

Multi-Conditioned Diffusion Synthesis of Sand Boils for Low-Resource Earthen-Levee Inspection

Padam Jung Thapa¹, Abdullah Bin Naeem², Ayon Dey², Anav Katwal², Md Tamjidul Hoque^{2,*}

¹University of Louisiana at Lafayette, Lafayette, LA, USA

²Department of Computer Science, Louisiana State University New Orleans, New Orleans, LA, USA

Emails: tpadamjung@gmail.com; {aneem, adey, akatwal, thoque}@lsuneworleans.edu

*Corresponding author: thoque@lsuneworleans.edu

Abstract—Sand boils on earthen levees are safety-critical defects, yet pixel-level detection from inspection imagery is limited by the scarcity of annotated examples. We address this low-resource setting with a diffusion-based synthesis pipeline. A Stable Diffusion XL backbone, fine-tuned with DreamBooth on a small curated reference set and conditioned by a multi-branch ControlNet stack, generates synthetic sand-boil imagery. A soft-mask inpainting protocol preserves the real defect region pixel-for-pixel while re-rendering the surrounding scene, avoiding the seams and colour casts of the seamless-cloning compositing used previously. A complementary mask-conditioned ControlNet generates a fresh boil into a chosen mask, so the conditioning mask is the segmentation label by construction, with the Convex Hull Annotator repurposed as a drift-checking quality gate—though certifying that label at scale is not yet solved by the available real-trained gate, so we release the soft-mask preset, not this one, as the default. Text conditioning is supplied by a taxonomy-driven *Prompt Atlas* that expands a single domain specification into a stratified, CLIP-validated prompt bank and ports to new defect classes without code changes. From the real training images the pipeline produces 1,020 synthetic candidates, of which 815 pass a CLIP admissibility filter to form the augmented dataset. We evaluate image quality with distribution and fidelity/diversity measures against the real reference set and a Poisson baseline, and audit the augmented set for out-of-distribution drift and memorisation. No single preset dominates: the presets trade off fidelity, diversity, and label reliability, so we release the label-reliable preset as the default and treat a curated mixture as the natural augmentation set. We scope our claims to image quality, label provenance, and diversity; downstream segmentation is left to future work. Code and an artefact manifest are released for reproducibility.

Index Terms—Sand boil detection, earthen levees, synthetic image generation, Stable Diffusion XL, DreamBooth, ControlNet, soft-mask inpainting, prompt curation, low-resource computer vision, civil infrastructure.

I. INTRODUCTION

Earthen levees protect large populations and property from flooding, yet their condition is hard to monitor at scale. The cost of inadequate inspection became clear

during the 2005 Hurricane Katrina failures, when independent investigators attributed the catastrophic flooding in New Orleans to a combination of design, construction, and inspection deficiencies [1]. Among the defect classes that inspectors are trained to recognise, sand boils are particularly consequential. A sand boil forms when water under hydraulic pressure carries fine sand upward through a permeable foundation layer, producing a small saturated dome at the surface (Figure 1). This is the early visible stage of internal erosion (also called piping), which, if left untreated, can progress to a sudden breach during a high-water event. Detecting and segmenting sand boils at the pixel level from routine field photographs therefore offers a direct lever for proactive flood-risk management.

Training a deep segmentation model for this task runs into a stubborn obstacle: annotated sand boil imagery is rare. The defect appears only during high water, is spread across long linear levee networks, and requires hydrologic expertise for accurate pixel-level annotation. The curated dataset used in this study (whose size is given in Section III-A) is, to our knowledge, among the larger curated sand-boil sets reported. Even at that scale it is too small to train a modern segmentation network without data augmentation or pretraining. Conventional spatial and photometric augmentation (rotation, flips, colour jitter, elastic distortion) inflates the sample count without broadening the underlying scene distribution. It cannot synthesise a sand boil under a hydraulic regime, soil type, or photographic style that the corpus does not already contain. Diffusion-based generative augmentation [2], [3] promises to fill exactly this gap.

Realising that promise on a niche scientific class raises three coupled problems, which this paper sets out to solve.

a) *Problem 1: Adapting a general-purpose diffusion model to a narrow defect class without retraining it.*: A stock Stable Diffusion XL pipeline [4] prompted with “a photo of a sand boil” returns images of volcanic vents, coastal surf, or children’s sandboxes; the pretrained base model has never seen civil-infrastructure defect imagery. Full fine-tuning is impractical at this reference-set scale. DreamBooth fine-tuning with low-rank adaptation [5], [6] confines the trainable parameters to a small additive adapter anchored by a rare trigger token. It fits comfort-

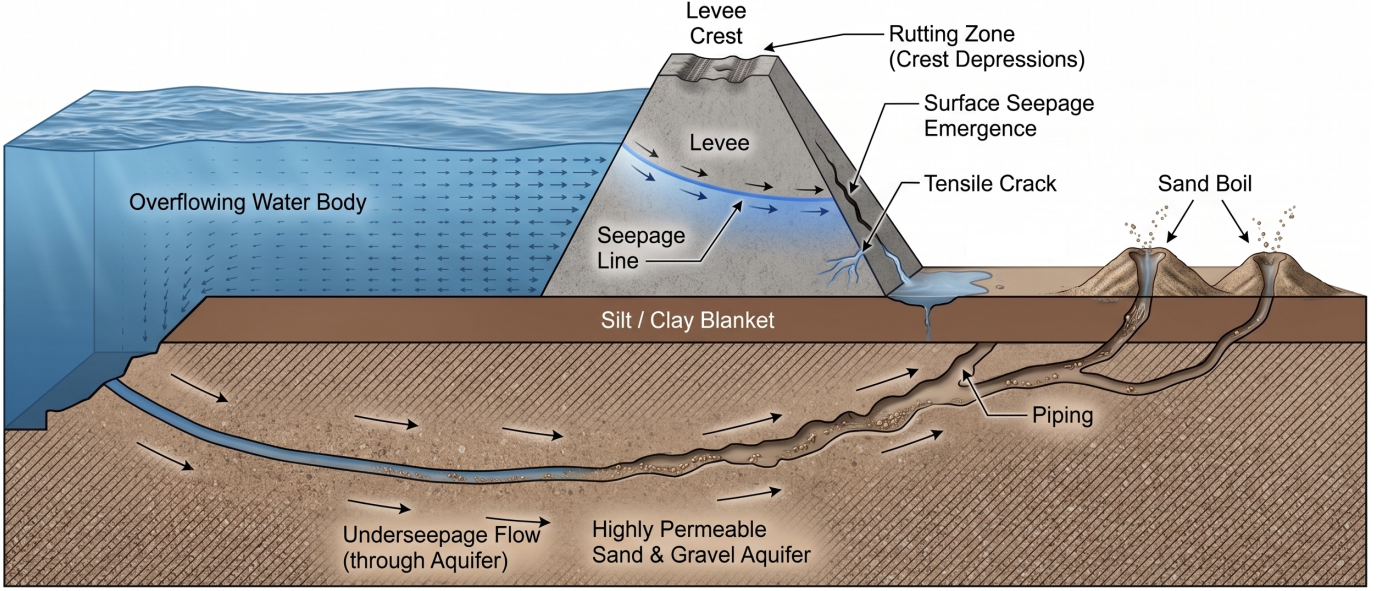


Fig. 1. Cross-sectional anatomy of a levee system and the surface defects it produces. High water in the river drives *under-seepage* through the permeable sand aquifer beneath the silt/clay blanket; where the flow concentrates, it erodes an upward channel (*piping*) that breaches the blanket and vents sand and water at the surface as a *sand boil*. The same hydraulic loading drives a phreatic *seepage line* through the embankment to a landside *seepage emergence*, while surface deterioration shows up as *cracks* and *rutting* on the crest. Sand boils, the early visible stage of internal erosion, are the focus of this paper, though the generation pipeline transfers to the other defect classes (Section VII).

ably on a single accelerator and preserves the frozen base for unrelated prompts. We adopt this recipe. Our main reason for building on SDXL rather than a higher-capacity rectified-flow [7] successor such as Stable Diffusion 3.5 [8] is the maturity of its publicly available ControlNet and IP-Adapter checkpoints, on which our conditioning stack depends; we also qualitatively observed the rectified-flow backbone to be more prone to memorise individual training frames at this adapter rank and data budget. The backbone choice is therefore deliberate, not a default; we examine it methodically in Section III-C.

b) Problem 2: Constraining the generated scene to a real field reference while leaving room for textural variety.: Text prompts alone cannot pin down the outline of a sand boil dome, its three-dimensional bulge against an otherwise flat ground surface, or the fine rim-texture transition where saturated sediment meets dry mud. To this end, we assemble a stack of four ControlNet branches [9] on top of the frozen DreamBooth-fine-tuned denoiser: Canny edges for object outline, monocular depth [10] for three-dimensional layout, surface normals for dome-rim relief, and holistically-nested edge detection (HED) soft edges [11] for rim texture. Each production preset activates a subset; the selected V4 preset combines HED soft edges, surface normals, and an IP-Adapter style anchor. To re-render the surrounding scene without disturbing the real defect pixels, we replace the image-to-image step with a soft-mask inpainting protocol. The inverted ground-truth mask is eroded and Gaussian-blurred into a soft gradient, and the denoiser predicts only inside that gradient at every step. The boundary ring is partially regenerated throughout the diffusion loop, so that by the time the

loop terminates, the transition has been negotiated against the new lighting and colour, and the hard-boundary seam of binary compositing is avoided. This addresses the two failure modes of the Poisson seamless-cloning step used in our earlier work [12], [13]: the hard mask boundary that betrayed compositing, and the colour bleed from the new background into the cloned region.

c) Problem 3: Producing textual conditioning that exhausts the scene distribution without manual prompt engineering.: A fine-tuned generator can only diversify within what the prompts ask for. Hand-written prompt lists scale poorly across defect classes, offer no guard against hallucinated or off-distribution prompts, and tie the augmentation pipeline to the domain expert who wrote them. We replace the hand-written list with a taxonomy-driven *Prompt Atlas*: a single JSON specification (concept, trigger token, axis labels, exemplar references, negative-concept list) is expanded by either template substitution or a small open-weight language model. It is then filtered by length, sparse text-similarity de-duplication, negative-concept screening, and image-text validation against the reference set with a CLIP similarity score [14]. The atlas is reproducible from the specification alone and is constructed to extend to new defect classes (seepage, sinkhole, rutting) without code changes; the empirical validation in this paper is restricted to the sand boil class (Section VI-E).

d) Contributions.: This paper contributes a generation pipeline whose components span three thrusts—defect-preserving generation, label provenance, and reproducible prompt curation—detailed below and assessed in one empirical study:

- **Defect-preserving soft-mask latent blending** (Section III-E): a per-step composition rule that pins the defect’s image-space appearance verbatim while giving the diffusion denoiser explicit freedom to negotiate a coherent transition with the re-rendered surrounding scene, removing the hard-seam and colour-bleed failure modes of binary inpainting and Poisson seamless cloning respectively.
- **Defect-size-adaptive soft-mask geometry** (Section III-E): a fully automated size-scaled rule that sets the soft-mask’s erosion radius and Gaussian kernel width from the defect’s bounding-box geometry, so the preservation margin and transition smoothness scale with the defect’s spatial extent rather than being hand-tuned per image.
- **Convex Hull Annotator** (Section III-A): a fully automated *shape-aware* probability-to-annotation pipeline (robust threshold $\rightarrow$ connected components $\rightarrow$ resolution-invariant area filter $\rightarrow$ shape-aware geometry adapter routed by form factor $\rightarrow$ dual raster/polygon output) that emits a single-policy training label *and* an editable polygon annotation. It routes each component between convex-hull, morphological-closing, and skeleton policies according to its geometry, so the same code path generalises across compact, elongated, and thin curvilinear defect classes (Figure 3). The same machinery serves a second role as a verification gate for the mask-conditioned path below.
- **Mask-conditioned generation with exact label provenance, as a zero-annotation diversity engine** (Section III-F): a binary-mask SDXL ControlNet, fine-tuned from a public soft-edge checkpoint, that generates *from* a chosen mask so the conditioning mask is the ground-truth label by construction. This removes the dependence on an external segmenter checkpoint that otherwise caps synthetic-label quality, and repositions CONVEX HULL ANNOTATOR as the drift-measuring quality-control gate for this path. Because the conditioning silhouette is decoupled from the real source pool, the same path renders single-boil scenes from a 39-mask bank under per-instance jitter. At zero annotation cost, this contributes exact labels and a variety of shapes, sizes, and placements that the source-anchored path cannot produce (Figure 7). We report its image quality and the limits of automatically certifying its labels, and make no downstream-segmentation claim for it here.
- **Reference-set similarity quality filtering** (Section III-I): an automated post-generation filter that drops out-of-distribution drift *and* memorisation at the same time, with an admissibility band auto-tuned from the reference set’s leave-one-out CLIP similarity statistics: no human review, no hand-picked thresholds.
- **Taxonomy-driven Prompt Atlas** (Section III-G): a five-stage prompt-curation pipeline that expands a single JSON domain specification into a stratified,

de-duplicated, image-text-validated prompt bank, replacing ad-hoc hand-written prompt lists and porting to a new defect class without code changes.

- **Empirical evaluation** of generative image quality: distribution distance (FID under two independent implementations, KID with standard deviation, CLIP image-text similarity, LPIPS perceptual diversity) re-based on the full reference set, *and* a fidelity-diversity decomposition via improved precision and recall [15] and density/coverage (Section V-F) that separates “looks like” from “covers” the references; a production-preset comparison (Section V-B); a quantitative comparison against a Poisson seamless-cloning baseline (Section V-E); empirical validation of the CLIP admissibility filter on the augmented set (815 of 1,020 synthetic candidates admitted), with a per-bucket diagnostic of where a global threshold is conservative (Section V-G); and a CLIP-space nearest-reference memorisation audit (Section V-H). Downstream segmentation use of the generated data is left to future work.

These components are not independent. The generation and labelling components interlock to give the augmented dataset a single-policy mask provenance and automatic quality envelopes on both the input and output sides (Section III-K), and together they are engineered to make the generated data usable as segmentation supervision—uniform label policy, on-distribution filtering—a property we design for but do not test downstream in this paper. The Prompt Atlas, together with the choice of SDXL over rectified-flow successors (Section III-C), is what ports across defect classes through a single JSON specification; transferring the full pipeline additionally requires a new DreamBooth adapter and structural-conditioning sources, so we describe it as designed to transfer rather than demonstrated on a second class.

e) Relation to prior work.: This pipeline builds on our earlier thesis [12], and we delineate the boundary explicitly. *New in this paper* are the defect-preserving soft-mask latent-blending compositor and its size-adaptive geometry (which replace the Poisson seamless-cloning step of the thesis), the mask-conditioned ControlNet with exact label provenance, the taxonomy-driven Prompt Atlas, the leave-one-out CLIP admissibility filter, the repositioning of CONVEX HULL ANNOTATOR as a drift-measuring verification gate, and an image-quality evaluation re-based on the full reference set under two independent FID implementations. *Inherited and refined* from the thesis are the core CONVEX HULL ANNOTATOR annotation algorithm and the Poisson-composition protocol; the latter appears here only as an evaluation baseline. This paper claims no contribution to downstream segmentation, which is left to future work.

The complete pipeline (DreamBooth checkpoint, Prompt Atlas record, multi-ControlNet inference configuration, soft-mask blending parameters, per-image seed cascade) is released alongside the paper, so the reported results can be reproduced.

f) Scope.: This paper focuses on the *generation* of synthetic sand boil imagery and on the methodology and metrics for evaluating that generation. Downstream use of the generated data (the leak-free cross-validated stacking ensemble and the confusion-score-mined hard-negative training phase) is left to future work.

The remainder of the paper is organised as follows. Section II positions the contribution against the literature on diffusion-based generative augmentation, DreamBooth and ControlNet, and prompt curation. Section III develops the pipeline. Section IV describes the experimental protocol. Section V reports quantitative and qualitative results. Section VI examines the design choices and the remaining failure modes. Section VII concludes.

II. RELATED WORK

The present work sits at the intersection of three research threads: automated detection of sand boils and related levee defects, diffusion-based image synthesis as a remedy for data scarcity, and the emerging practice of systematic prompt construction for generative-augmentation pipelines. We review each in turn and, in each, mark precisely where our pipeline departs from established practice. Segmentation architectures and ensemble learning, which consume the synthetic data produced here, lie outside this scope and are left to future work.

A. Sand-boil and levee-defect imagery

Automated inspection of earthen levees began with classical image processing (edge filters, thresholding, and handcrafted descriptors fed to shallow classifiers), which degrades badly under the lighting changes, mud, vegetation, and reflective standing water that dominate field photographs. Deep learning has since become the norm across infrastructure distress detection [16], and the sand-boil literature in particular has moved to convolutional encoder-decoders. Panta *et al.* [17] proposed SandBoilNet, a fully convolutional encoder-decoder whose partially fine-tuned ResNet-50v2 backbone carries PCA-based channel- and spatial attention with Inception blocks on its skip connections. Its reported intersection-over-union (IoU) of 0.5743 and balanced accuracy of 0.8552 on a held-out real test set are the reference point for the downstream segmentation in this series. Earlier, Kuchi *et al.* [18] were the first to couple a small synthetic sand-boil set with a CNN, but they classified whole images rather than segmenting pixels.

The same convolutional recipe recurs across neighbouring infrastructure defects: culvert-and-sewer segmentation with an enhanced feature-pyramid network [19], sinkhole detection with a depthwise-separable U-Net [20], and levee-seepage segmentation from the same group [21]. Synthetic-aperture-radar monitoring [22] probes subsurface water movement, but at a spatial resolution far too coarse for the small-scale cones that define a sand boil.

What unites these studies is that none treats diffusion synthesis as a primary source of training data. Our own

earlier prototype [12] took a first step, namely single-control diffusion augmentation on Stable Diffusion v1.5 with Poisson seamless cloning, but the present paper rebuilds it end to end: a stronger multi-conditioned generator, a soft-mask inpainting protocol that keeps the labelled defect essentially intact (up to a small, post-processed rim drift), and a reproducible prompt builder in place of a hand-written list.

B. Diffusion-based generative augmentation

Before diffusion, variational autoencoders [23] and generative adversarial networks [24] were the default synthesizers under data scarcity. Both struggle in our regime: on the small, visually homogeneous reference sets typical of levee inspection, StyleGAN variants [25], [26] are prone to mode collapse and unstable training, and tend to echo a handful of real images rather than broaden the distribution.

Diffusion models [27], [28] trade this instability for a many-step denoising process that is markedly easier to train. Latent diffusion moves the denoising into a compressed latent space [2], making high-resolution synthesis tractable on a single GPU, and Stable Diffusion XL (SDXL) [4] scales the backbone and native resolution further. Three extensions turn this backbone into a controllable, subject-specific generator. DreamBooth [5] binds a new subject to a rare trigger token from only a handful of reference images, so later samples render that subject under novel scenes and lighting. Low-Rank Adaptation [6] folds the fine-tuning update into a compact additive adapter (cutting the trainable parameters from billions to roughly ten million), so DreamBooth-style training fits on a single accelerator. ControlNet [9] attaches a parallel branch that injects a structural signal (a Canny edge map, a monocular depth or surface-normal estimate, or a holistic-edge map, abbreviated HED) at every skip connection of the frozen denoiser, enforcing geometry that text prompts cannot. Orthogonally, the IP-Adapter [29] feeds a reference-*image* embedding through a decoupled cross-attention path, supplying a style anchor that we exploit in the V3 and V4 presets (Section III-H).

That diffusion augmentation pays off is by now well documented: in image classification [3], [30], across remote-sensing tasks [31], and especially in data-scarce medical imaging [32], where it has become a routine response to small annotated cohorts. Generative augmentation of surface and industrial defects is itself an active area—GAN- and diffusion-based crack, weld, and anomaly synthesis, including mask-controlled defect generation [33], [34]—but pixel-level earthen-levee defects, and sand boils in particular, have seen almost none of it. We also note the newer rectified-flow diffusion models, which follow long compositional prompts more faithfully and render text better, but demand more memory and lack a mature ecosystem of structural-conditioning checkpoints. For the structural control our pipeline depends on, we therefore commit to SDXL, as argued in Section III-C.

a) Composition strategies.: When diffusion is used to augment defect imagery, the usual recipe is to synthesise a whole image from a prompt and then either trust the generator’s auto-produced mask or hand-label the result. A second line of work edits inside a fixed region: RePaint [35] re-noises and re-denoises masked pixels in image space, while Blended Diffusion [36] fuses a text-driven edit with the source under a region mask. Our soft-mask path takes the complementary region—preserving the masked defect verbatim and re-rendering only the surrounding scene—but its contribution over merely complementing an inpainting mask is the per-step latent composition schedule and the size-adaptive blend band that keeps the preserved defect aligned with the real source geometry. We position the method relative to RePaint and Blended Diffusion rather than benchmark against them: our objective is label-preserving scene re-rendering, not masked-region editing, so a like-for-like comparison is left to future work. The segmentation label itself is re-predicted post hoc rather than inherited (the re-prediction ceiling discussed in the next paragraph). This soft-mask path sidesteps both the labelling cost of fully synthetic generation and the hard seams and colour bleed of the Poisson seamless cloning [13] used in our earlier work [12].

b) The label-provenance gap.: A common thread runs through all of these recipes: the image is synthesised first and the *label* is recovered afterwards, whether by trusting an auto-produced mask, by hand-annotating it, or (as in our soft-mask path) by re-predicting the mask with a segmenter whose accuracy then caps the label quality. The alternative is to invert the dependency and let a chosen structural map *be* the label, so the generator is conditioned to fill it. ControlNet [9] makes this mechanically possible by accepting an arbitrary structural map as conditioning. Making the conditioning map the label by construction is the premise of semantic image synthesis and generate-with-labels work [37], [38] and, in the industrial setting, of mask-controlled anomaly and defect synthesis [33], [34]; it is comparatively under-explored for low-resource civil-infrastructure defects and, to our knowledge, for sand boils specifically. Our mask-conditioned path (MaskCN, Section III-F) takes this route in that setting and pairs it with the CHA drift gate: it fine-tunes a binary-mask SDXL ControlNet and generates a fresh defect into a chosen silhouette, so the mask is the label with no post-hoc prediction step. Because the conditioning silhouette need no longer come from a real photograph, the mask bank can be jittered per instance (flip, rotation, rescale, and repositioning) to obtain a variety the source-anchored inpainting path cannot produce. This positions MaskCN as a complementary diversity engine to the soft-mask path: the latter widens the scene distribution around real defect geometry; the former widens the shape, size, and placement distribution of the defect silhouette at zero annotation cost. We report its image quality and the limits of automatically verifying its labels with a real-trained segmenter, and defer any downstream-segmentation claim to future work.

C. Prompt curation

Prompt construction is a consequential part of any generative-augmentation pipeline. Most systems rely on a short, hand-written prompt list, which scales poorly to new domains and offers no protection against hallucinated or off-distribution prompts. Automated prompt construction has been explored mainly for classification (through chain-of-thought or template expansion), while the image-synthesis side remains comparatively thin.

We address this with the Prompt Atlas. A single JSON domain specification (concept, trigger token, axis labels, exemplar references, and negative concepts) is expanded by template substitution or a small open-weight language model, and the candidates are filtered by length, by sparse text-similarity de-duplication, by negative-concept matching, and by image-text similarity to the reference set [14], [39]. The atlas is reproducible from its specification alone and ports to a new defect class by editing that single file, with no changes to the builder.

III. METHODOLOGY

The generation pipeline has two coupled stages. First, a diffusion-based generator, fine-tuned on a small curated reference set, produces structurally faithful synthetic sand boil imagery while preserving the real defect region pixel-for-pixel. Second, a taxonomy-driven prompt builder supplies the generator with stratified, validated textual conditioning. Each synthetic record is paired with a binary mask from a uniform automated annotator, so every generated image is admissible as training supervision for downstream segmentation. Figure 2 gives a schematic overview, and the subsections that follow develop each component in turn.

A. Dataset and preprocessing

The real sand boil dataset is drawn from the U.S. Army Corps of Engineers (USACE) levee-inspection archive [40]. Field inspectors traverse the crest and adjacent areas of an earthen levee with mobile-phone or DSLR cameras and photograph any abnormalities they encounter. The archive contains over four thousand defect photographs. From these, a domain expert curated a subset of approximately three hundred sand boil images on the basis of visibility and unambiguous appearance. This subset was then filtered by automatic intersection-over-union agreement between the manual annotation and an initial Segment Anything [41] proposal. The final labelled set contains $N_{\text{train}}=199$ training images and $N_{\text{test}}=51$ held-out test images at variable native resolutions, each paired with a binary pixel-level mask. A high-quality *subset* of $N_{\text{ref}}=39$ images with the highest agreement score anchors the DreamBooth-LoRA appearance fine-tuning and supplies the conditioning-mask bank used by mask-conditioned generation (MaskCN, Section III-F). All counts in this paper refer back to these quantities; elsewhere the prose says “the training set”, “the held-out test set”, and “the curated subset” to avoid restating the absolute numbers.

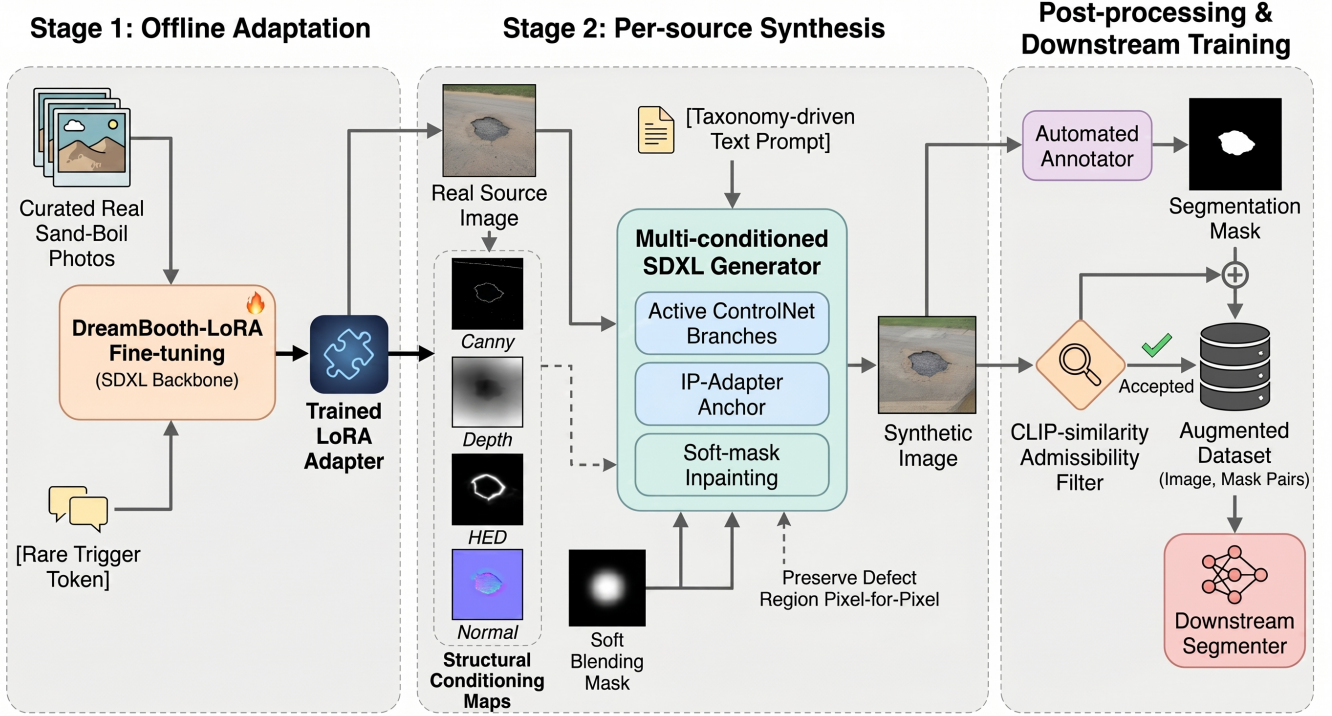


Fig. 2. Conceptual overview of the synthesis pipeline. **Stage 1 (offline adaptation)**: a DreamBooth-LoRA adapter is fit once to the curated reference set, anchored by the rare trigger token `sbx`, and loaded into the synthesis stack. **Stage 2 (per-source synthesis)**: for each real source image, the four-branch ControlNet stack (Canny, monocular depth, HED soft edges, and surface normal) and a soft blending mask are extracted; the multi-conditioned SDXL generator activates a preset-specific subset of the ControlNet branches, together with an IP-Adapter style anchor and soft-mask inpainting, and re-renders the surrounding scene while preserving the real defect region pixel-for-pixel (the selected V4 preset uses HED soft edges and surface normal). **Post-processing**: each synthetic image is paired with a mask by the CONVEX HULL ANNOTATOR and passed through the CLIP admissibility filter (Section III-I); accepted (image, mask) pairs enter the augmented dataset used to train the downstream segmenter.

a) *Annotation protocols.*: Two separate mask-production protocols are used: one for the real reference images and one for the synthetic outputs of the generator.

For the *real* images, every mask is hand-drawn at the pixel level and then cross-checked against a Segment Anything proposal of the same image [41]. The IoU between the two is recorded, and images with low agreement are returned for re-annotation rather than admitted to the training set. This yields masks that preserve concave rim features, and it adds an objective quality filter that the click-based protocol of our earlier work [12] lacked.

For the *synthetic* images, we introduce CONVEX HULL ANNOTATOR (CHA) [12], [42], a fully automated *shape-aware* probability-to-annotation pipeline. CHA takes a sigmoid probability map from a class-specific segmenter (SandBoilNet [17] for sand boils) and emits a paired binary mask plus a VGG-Image-Annotator (VIA)-format polygon annotation [43]. CHA is the second of two coupled mask-provenance mechanisms in our pipeline; the first is the soft-mask inpainting protocol of Section III-E. Because CHA is applied uniformly across all four production presets, the augmented dataset has a single, reproducible mask-production policy. The same shape-aware machinery serves a second role. In the mask-conditioned generation path of Section III-F the conditioning mask is already

the ground-truth label, so CHA acts instead as a quality-control gate that measures how far the rendered dome drifted from the requested silhouette. The shape-aware geometry adapter (stage 4 below) makes CHA class-agnostic in both roles. The pipeline therefore transfers from sand boils to seepage, rutting, sinkhole, or crack defects through a single domain-specification file, with no edits to the post-processing logic.

b) *Pipeline.*: Given a probability map $\hat{p} \in [0, 1]^{H \times W \times C}$ from a class-specific segmenter (collapsed across C output channels by per-pixel max), CHA proceeds in five stages (Figure 3):

- 1) **Robust threshold.** Binarise the probability map at an image-specific threshold τ . The default policy is Otsu’s method, which selects τ to maximise the between-class variance of $\hat{p}$; alternatives (Li, Yen, mean, fixed) are exposed in the domain specification. Otsu replaces the original CHA’s mean-of-probability threshold, which is fragile on near-empty and near-saturated probability maps.
- 2) **Connected components.** Label the binary mask with eight-connectivity.
- 3) **Resolution-invariant area filter.** Reject components whose area falls outside $[a_{\min}, a_{\max}]$ expressed as *fractions of image area*; defaults $a_{\min}=0.1\%$ and

$a_{\max}=85\%$ cover both threshold-noise suppression and segmenter-blow-up rejection. Expressing the bounds as fractions removes the original CHA’s 100-pixel-at-512² hard-coding, so the same configuration generalises across image resolutions.

- 4) **Shape-aware geometry adapter.** For each surviving component, the dimensionless form factor $F = \text{perimeter}^2 / (4\pi \cdot \text{area})$ is computed ($F=1$ is a perfect circle, larger values indicate thinness). The component is then cleaned with a shape-appropriate policy: **convex hull** when $F < 3$ (compact defects: sand boils, sinkholes), **morphological closing** when $3 \leq F < 10$ (elongated-but-thick defects: rutting grooves), or **skeleton + controlled dilation** when $F \geq 10$ (thin curvilinear defects: cracks). Auto-routing can be overridden by an explicit shape policy in the domain spec. This is the principal generalisation that makes CHA class-agnostic.
- 5) **Dual output with per-component quality scores.** Union the cleaned components into a 0/255 binary raster mask for downstream segmentation supervision, and emit each component as a VIA-format polygon [43] carrying a quality record ($s, \text{solidity}, F$) where $s = \text{mean}(\hat{p})$ inside the component. The polygons load directly into standard polygon-annotation tools (VIA, CVAT, LabelMe) for optional human review; the quality scores let downstream consumers filter low-confidence components without re-running the segmenter.

The dual raster-plus-polygon output is what makes CHA an *annotator* rather than a mask cleaner: the raster path feeds downstream training, and the polygon path feeds downstream human review. The shape-aware geometry adapter (stage 4) is what makes CHA class-agnostic: the same code path handles compact, elongated, and thin curvilinear defects through automatic form-factor routing, with no parameter change. This routing is validated on synthetic defects of all three shape classes in the reference implementation.

The same CHA pipeline is applied uniformly to every synthetic image, regardless of which production preset produced it (see Section III-H), so mask provenance is single-policy across the augmented dataset. As Figure 3 makes visually concrete, the cleanup is not cosmetic. The spurious small blobs that survive the naive $\hat{p} \geq 0.5$ threshold, if used as training labels directly, would teach the downstream segmenter to fire on threshold-uncertainty noise in regions of the synthetic image that contain no defect at all.

For downstream segmentation, images and masks are resampled to 512×512 pixels. Image resampling uses cubic interpolation; mask resampling uses nearest-neighbour interpolation followed by a binary threshold, so that labels remain strictly $\{0, 1\}$ after resize. After loading, channels are reordered from BGR to RGB and pixel values are scaled to $[0, 1]$. A suite of geometric and photometric augmentations is drawn from the Albumentations library [44],

applied jointly to image and mask for spatial transforms and to the image alone for intensity and noise transforms.

B. DreamBooth fine-tuning with low-rank adaptation

Stable Diffusion XL is a publicly released general-purpose text-to-image model trained on a large web-scale image corpus by Stability AI, and it has no prior exposure to civil-infrastructure defect imagery. A stock SDXL pipeline prompted with “a photo of a sand boil” tends to return unrelated imagery that the word evokes from web text (volcanic vents, coastal surf, children’s sandboxes), confirming that the base model does not encode the defect class. To adapt the model without retraining it, we use the DreamBooth-LoRA recipe of [5], [6]. The pretrained SDXL weights remain frozen throughout training: no value inside the released checkpoint is modified. Instead, a small additive adapter (about 10 million trainable parameters, against SDXL’s ~ 2.6 billion) is fit to the curated USACE reference set and paired with a rare trigger token **sbx**. The adapter is therefore the only component of the generator that has observed real sand boils. At inference time the frozen base supplies the general photographic world model (scenes, lighting, materials, geometry), and the adapter steers those priors toward the defect class whenever the prompt contains **sbx**. Unloading the adapter returns the pipeline to plain pretrained SDXL, and the off-domain behaviour reappears. We read this as evidence that the sand-boil-specific adaptation is carried by the adapter rather than the frozen base.

We fine-tune SDXL with a Low-Rank Adaptation [6] adapter inserted at the four attention projections (query, key, value, and output) of every transformer block in the denoiser. For each projection the original weight matrix W_0 is frozen and a trainable update $\Delta W = BA$ is added, with $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times d}$ and $r=16$ in the production SDXL configuration:

$$y = W_0 x + \frac{\alpha}{r} B A x, \quad (1)$$

where α is a fixed scaling constant and the gradient flows only through A and B (never through W_0). For an SDXL attention matrix of width $d \approx 1280$, each adapter contributes $\sim 2rd \approx 40,960$ parameters in place of the original $d^2 \approx 1.6 \times 10^6$, a per-matrix reduction of roughly $40\times$.

The trigger token is a rare three-letter string, **sbx**, selected because it does not appear meaningfully in the SDXL tokeniser. Reference images are paired with the instance prompt *a photo of sbx sandboil* during training. We train for 1,500 steps under a cosine learning-rate schedule (peak $1e-4$), batch size 1, and gradient accumulation 4, with class-image prior preservation (200 auto-generated *a photo of a sandboil* images), using mixed precision (fp16), the 8-bit Adam optimiser [45], and gradient checkpointing. (An initial rank-8, 800-step prototype was used during development; all results reported here use this production rank-16 adapter.) The trained adapter is serialised as a single safetensors file and loaded into the SDXL inference

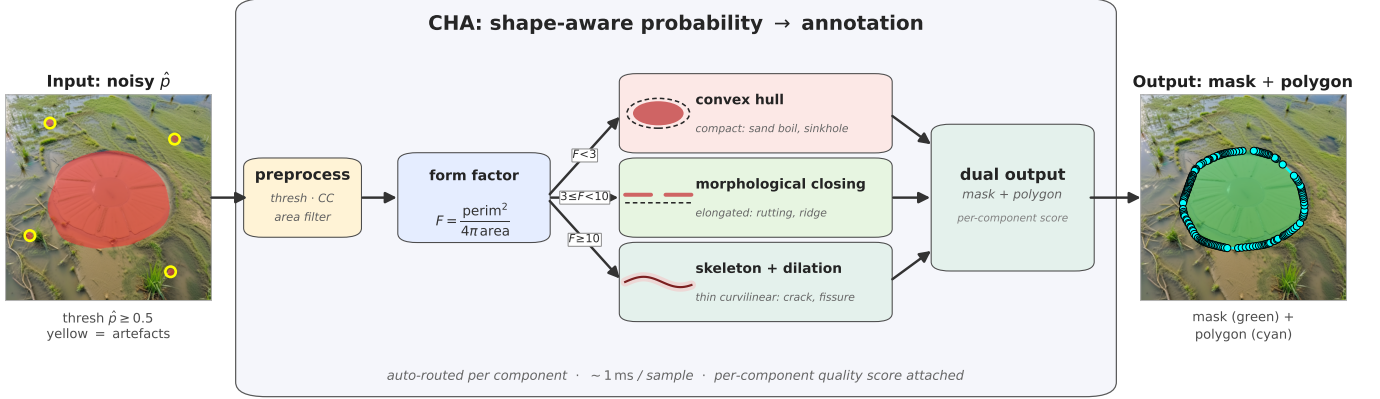


Fig. 3. CONVEX HULL ANNOTATOR (CHA), the shape-aware probability-to-annotation pipeline. After thresholding and component cleanup (connected components, area filter), each surviving component is routed by its form factor $F = \text{perim}^2 / (4\pi \text{area})$ to a convex hull ($F < 3$, compact), a morphological closing ($3 \leq F < 10$, elongated), or a skeleton with controlled dilation ($F \geq 10$, thin curvilinear), yielding a 0/255 binary raster mask and a VIA-format polygon with a per-component quality record.

pipeline. The full configuration is restated for reproducibility in Section IV-B.

C. Backbone choice: SDXL versus rectified-flow successors

Choosing Stable Diffusion XL over a more recent rectified-flow successor is a deliberate trade-off, because the two families occupy different operating points. The successors we considered are Stable Diffusion 3.5 [8] and comparable Multi-Modal Diffusion Transformer (MMDiT) backbones. Rectified-flow MMDiT backbones are widely reported to give stronger prompt adherence on long compositional captions, better in-image text rendering, and a more cinematic global appearance. They do so at noticeably higher VRAM cost per generated pixel and with a substantially less mature ecosystem of pretrained ControlNet, IP-Adapter, and structural-conditioning weights at the time of writing. SDXL, in contrast, retains the strongest publicly released ControlNet stack (Canny, depth, HED, normal, tile, soft-edge), the most mature LoRA tooling, and the lowest inference latency at 1024×1024 . Its cost is weaker prompt adherence on unusually long or compositional prompts and weaker text rendering. For the present application, none of MMDiT’s strengths are on the critical path: our prompts are short scene descriptions of a single defect on a levee, no in-image text is required, and the generation is structurally constrained by three ControlNet branches plus IP-Adapter rather than by the prompt alone. SDXL’s strengths, in turn, are exactly the ones we depend on, so the trade-off resolves cleanly in favour of SDXL.

Three further observations from our own pipeline experimentation reinforced this choice. We report them as qualitative findings rather than as headline numbers. First, the publicly available ControlNet and IP-Adapter checkpoints for SDXL were directly compatible with our soft-mask inpainting protocol of Section III-E, whereas equivalent production-quality conditioning checkpoints for

the rectified-flow backbone were not available in a form that interoperates with the same protocol. Second, at this application’s reference-set scale, a high-capacity MMDiT denoiser fine-tuned via DreamBooth-LoRA appeared, in qualitative inspection, more prone to memorising individual training frames: the generator reproduced the trained dome with little geometric variation across prompts and seeds. This overfitting failure mode was less pronounced on the lower-capacity SDXL U-Net at the same adapter rank and data budget. Third, on the H100 NVL hardware used here, the SDXL inpainting pipeline runs comfortably at 1024×1024 with up to three simultaneous ControlNets (or two ControlNets plus an IP-Adapter). In our informal profiling, the rectified-flow counterpart at the same conditioning depth left noticeably less headroom for the IP-Adapter style anchor that the present pipeline relies on for visual variety. We kept the comparison on a fixed structural-augmentation framework to keep it operationally simple and reproducible. We therefore report SDXL as the production backbone and treat the rectified-flow alternative as out of scope.

D. Multi-ControlNet conditioning

Text prompts alone cannot constrain the geometry of a generated image to that of a particular field reference. ControlNet [9] addresses this by training a parallel branch of the denoiser to consume a structural map and emit residual features that are additively combined with the main branch at every skip connection and at the mid block. The present pipeline makes four publicly released ControlNet variants available on top of the DreamBooth-fine-tuned SDXL backbone, so that each generated image can be constrained from up to four independent structural perspectives; each production preset activates a subset of these branches (Section III-H).

- 1) **Canny-edge conditioning.** The Canny detector [46] is run on the source image with low and high

thresholds of 30 and 90 (a dense setting selected to trace the low-contrast boil rim), producing a binary edge map that constrains the outline of the sand boil dome together with the dominant boundaries of the surrounding scene.

- 2) **Monocular-depth conditioning.** A pretrained vision transformer for dense depth prediction [10] produces a per-pixel depth estimate of the source image, which constrains the three-dimensional layout. A dome bulging out of an otherwise flat ground surface remains bulging in the synthetic counterpart.
- 3) **Soft-edge (HED) conditioning.** Holistically-Nested Edge Detection [11] produces a soft probability map rather than a binary boundary, capturing fine textural transitions (grain boundaries, wet versus dry patches, the rim where saturated sediment meets the surrounding mud) that the binary Canny map cannot represent.
- 4) **Surface-normal conditioning.** A surface-normal map estimated from the source image constrains the local surface orientation, reinforcing the dome-rim relief so that the bulge of the boil reads as a coherent three-dimensional form rather than a flat texture patch. It complements the depth branch, which fixes coarse layout, by supplying finer per-pixel orientation cues.

In the forward pass of every diffusion step, the residual at denoiser skip connection i becomes

$$\text{skip}[i] = h_i + \lambda_c r_i^c + \lambda_d r_i^d + \lambda_h r_i^h + \lambda_n r_i^n, \quad (2)$$

where h_i is the encoder feature, $r_i^c, r_i^d, r_i^h, r_i^n$ are the Canny, depth, soft-edge (HED), and surface-normal branch residuals at the corresponding resolution, and $(\lambda_c, \lambda_d, \lambda_h, \lambda_n) = (0.40, 0.25, 0.40, 0.40)$ are the conditioning scales used throughout. Individual presets activate a subset of the four branches; the unused branches are simply omitted from the sum. The same combination is applied at the mid block. The scale budget is itself a quality lever. Every additional control branch tightens the geometric constraint, but it also consumes a portion of the denoiser’s representational budget that would otherwise render high-frequency texture. Loading the stack indiscriminately therefore tends to produce visibly smoothed outputs even at 1024×1024 resolution. The Canny–depth–HED combination is the smallest configuration that satisfies our core geometric requirements (outline, three-dimensional layout, rim texture), while the surface-normal branch adds finer dome–rim relief when a preset calls for it. The ablation of Section IV-C traces the cost of denser combinations and the contribution of an IP-Adapter style anchor [29]. The anchor imports high-frequency texture from a real reference photograph and thereby partially restores fidelity in configurations where one of the structural controls has been dropped. Figure 4 illustrates the additive-fusion mechanism on three of these structural controls (Canny, depth, and HED soft edges) extracted from a real source; the surface-normal branch fuses identically.

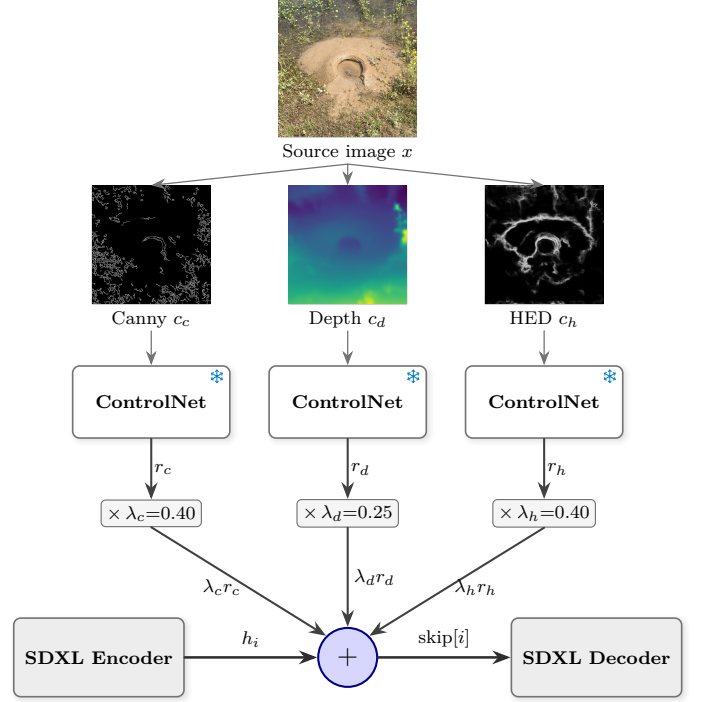


Fig. 4. Multi-ControlNet conditioning of SDXL. Canny, depth, and HED maps from the source x each drive a frozen, pretrained ControlNet; the residuals, scaled by their weights λ , are summed with the encoder feature h_i into $\text{skip}[i]$ (Eq. (2)) before the decoder. The SDXL backbone is frozen and carries the trained DreamBooth-LoRA adapter (Section III-B).

Figure 5 traces, for two real parents, the three control maps that drive each synthesis (Canny edges, surface normals, and HED soft edges) through to the generated output. Qualitatively, the branches capture complementary structure: the normal map encodes the dome–rim relief and HED the soft boil outline and sediment texture, while Canny contributes a dense map of high-frequency scene edges (vegetation, debris, and water lines) and is the weakest cue for the smooth, low-contrast boil itself. We do not quantify each branch’s contribution; their conjunction is what keeps the synthetic boil geometrically consistent with its parent while the DreamBooth trigger repaints faithful texture.

E. Defect-preserving soft-mask latent blending

Defect-preserving soft-mask latent blending is a per-step composition rule that pins the defect’s image-space appearance verbatim while giving the diffusion denoiser explicit freedom to negotiate a coherent transition with the re-rendered surrounding scene. This is the first of two coupled mask-provenance mechanisms in our pipeline; the second is the CONVEX HULL ANNOTATOR mask producer of Section III-A.

a) *The failure modes it replaces.*: The image-to-image and inpainting paths of Stable Diffusion accept a mask that controls which pixels the denoiser is allowed to modify. The naive choice for our setting (preserve the sand

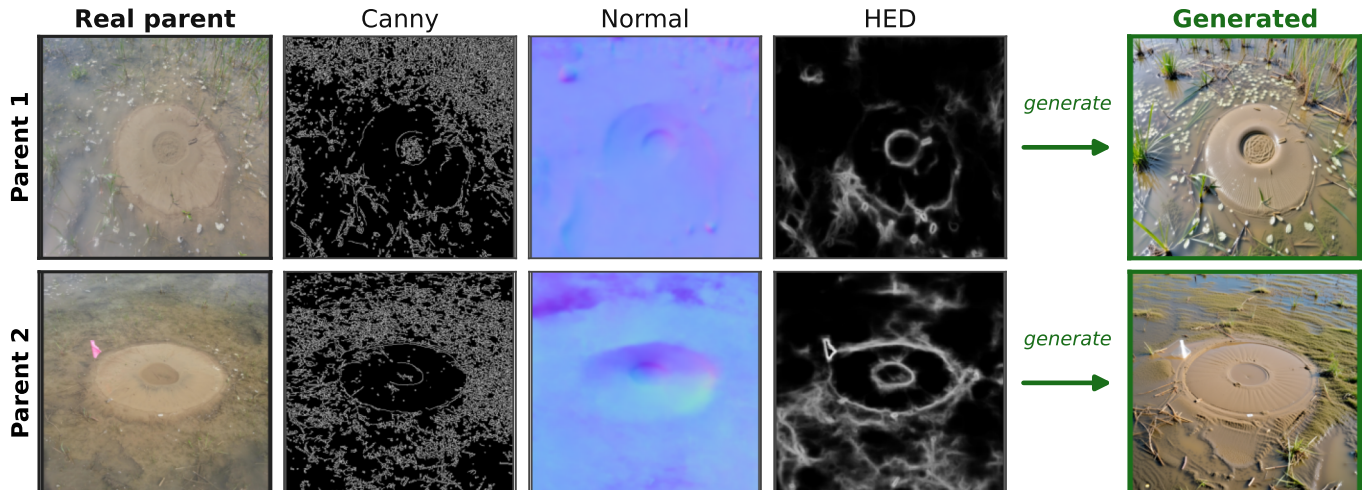


Fig. 5. ControlNet conditioning for two real sand boil parents. Each row shows the real parent and its three control maps (Canny edges, surface normal, and HED soft edges) that jointly condition the DreamBooth-LoRA SDXL generator (trigger “`sbx sandboil`”) to produce the synthetic output (green frame).

boil core, re-render the background) would pass the binary ground-truth mask directly. In practice this produces a hard, visible seam at the boundary, because the predicted and original latents change abruptly across a single pixel of the mask. Our previous work [12] used Poisson seamless cloning [13] to blend the regenerated background onto the original sand boil, but this introduced a separate failure mode: the cloned region tended to inherit the colour cast of the new background and appeared tinted. Soft-mask latent blending eliminates both failure modes by construction. The boundary is renegotiated continuously across the diffusion loop, rather than sharply at a single mask edge or post hoc in pixel space.

b) Construction of the soft mask.: The ground-truth sand boil mask is first inverted, so that 1 denotes the background region we wish to regenerate and 0 denotes the dome core we wish to preserve. The inverted mask is eroded with an elliptical structuring element (growing a small protected ring around the dome) and then convolved with a Gaussian kernel. The result is a smooth gradient mask $M \in [0, 1]^{H \times W}$ that is 0 in the core, 1 in the far background, and varies smoothly across a boundary ring.

c) Defect-size-adaptive geometry.: The erosion radius and the Gaussian kernel width control two complementary properties of the soft mask. The erosion controls the *preservation margin* (how much of the dome interior is locked verbatim), and the blur controls the *transition smoothness* (how rapidly the denoiser regains freedom as we cross the rim). A fixed (12, 41)-pixel pair, the choice in our earlier work [12], over-erodes small domes while barely affecting large ones. We instead derive both radii adaptively from the defect’s spatial extent. Let $d = \min(h_{\text{bbox}}, w_{\text{bbox}})$ be the smaller side of the dome’s tight

bounding box. We set

$$\begin{aligned} r_{\text{erode}} &= \max(4, \text{round}(0.04 d)), \\ r_{\text{blur}} &= \max(15, \text{odd}(\text{round}(0.12 d))), \end{aligned} \quad (3)$$

where $\text{odd}(\cdot)$ rounds up to the nearest odd integer (Gaussian kernel sizes must be odd) and the floors of 4 and 15 pixels prevent degenerate operations on extremely small domes. The coefficients 4% and 12% were selected so that the preservation margin and the transition ring scale linearly with the dome’s size: a 300-pixel-wide dome receives a 12-pixel erosion and a 37-pixel blur (close to the fixed (12, 41) values of our prior work); a 100-pixel dome receives a 4-pixel erosion and a 15-pixel blur; and a 600-pixel dome receives a 24-pixel erosion and a 73-pixel blur. The adaptive policy adds no manual tuning, requires no labelled data, and is computed in milliseconds from the bounding box of the binary mask that is already part of the inpainting input.

d) Per-step blending.: At each diffusion step t the inpainting pipeline blends the predicted denoised latent with the noised source latent under the soft mask:

$$z_t = M \cdot \hat{z}_t + (1 - M) \cdot z_t^{\text{src}}, \quad (4)$$

where $\hat{z}_t$ is the latent predicted by the denoiser at step t and z_t^{src} is the source image latent noised to the same step. Because M is smooth, the boundary ring is partially regenerated at every step, so the model can render a coherent transition between the preserved dome and the re-rendered background. By the time the diffusion loop terminates, the boundary has been negotiated against the new lighting and colour at every denoising step, so the hard-boundary seam of binary compositing is strongly suppressed. The boundary ring of M is therefore the exact set of pixels where the denoiser has freedom to negotiate

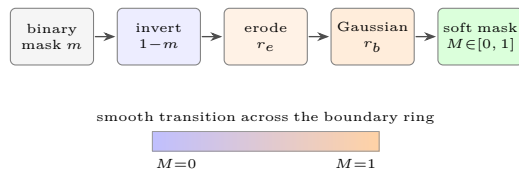


Fig. 6. Soft-mask construction. The binary defect mask m is inverted, eroded by r_e pixels, and Gaussian-blurred with kernel r_b into a smooth gradient mask M . Both radii are derived adaptively from the defect bbox minimum side d via (3) ($r_e=0.04d$, $r_b=0.12d$). Under per-step blending (4), $M=0$ preserves the dome verbatim, $M=1$ regenerates the far background, and the boundary ring is partially regenerated at every denoising step.

texture continuity. This property follows directly from (4), and it is what makes the soft mask preferable to either binary inpainting (hard seam) or Poisson seamless cloning (post-hoc colour bleed). Figure 6 traces the construction of M from the binary defect mask.

F. Mask-conditioned generation: the mask as ground truth

The soft-mask path of Section III-E preserves the dome in image space, but it does not inherit the label: because the boundary ring is regenerated, the ground-truth mask is *re-predicted* by CONVEX HULL ANNOTATOR (CHA) from the public SandBoilNet checkpoint (Section III-A). Label quality is therefore upper-bounded by that checkpoint’s accuracy on synthetic images. We remove this ceiling with a second, complementary provenance path in which the label is exact *by construction*.

a) Method.: We fine-tune a Stable Diffusion XL ControlNet whose conditioning input is the binary defect mask itself. Following standard practice for data-scarce ControlNet training, we do not train from scratch. We initialise from a publicly released soft-edge (HED) SDXL ControlNet [9], whose edge-to-structure prior is the closest match to a filled silhouette, and fine-tune on the 199 curated (image, mask) pairs with the mask replicated to three channels, at 768×768 in full precision with gradient checkpointing (the mixed-precision path in the public SDXL-ControlNet trainer is numerically unstable for this configuration). At inference the frozen SDXL backbone, the DreamBooth-LoRA adapter (Section III-B), and this mask-ControlNet are composed, and an image is generated *from* a chosen binary mask in pure text-to-image mode (there is no source photograph and therefore no denoising-strength, depth, HED, or Canny conditioning: the silhouette is the only structural signal). We use classifier-free guidance 8.5, a DPM++ 2M Karras sampler, and the same fp16-fix VAE as the rest of the pipeline; to suppress per-sample artefacts we draw a small candidate pool per prompt and retain the one with the highest no-reference CLIP-IQA quality [47]. Because the generator is conditioned on that mask, the mask *is* the ground-truth label: there is no post-hoc prediction step and no dependence on an external segmenter checkpoint.

b) MaskCN conditioning and per-instance diversity.: Mask-conditioned generation (*MaskCN*) decouples the

synthetic *label geometry* from the catalogue of real source domes, and this unlocks a second role beyond exact provenance: a controllable *diversity engine*. The conditioning mask need not come from a real photograph. Masks are drawn from a bank of 39 reference silhouettes. The ControlNet is applied at its default conditioning scale (1.0), yielding one boil per render at a requested size and position. To turn a fixed bank into a genuine diversity source, each mask is independently flipped, rotated, rescaled, and repositioned before conditioning (*per-instance jitter*), so repeated draws of the same source silhouette yield distinct boil shapes, sizes, and placements rather than identical copies. The rescale is sampled across a small-to-large size range that deliberately populates the small-object regime where pixel-level recall is weakest. Because every perturbation acts on the conditioning mask itself, the by-construction labelling guarantee is *exact*: one mask blob reliably yields one rendered boil. Figure 7 makes this concrete, pairing each conditioning mask with the render it produces and the mask-render overlay that shows label registration.

c) CHA as a verification gate.: This is where CHA’s role shifts rather than disappears. For the diffusion-augmentation presets (V1–V4), CHA remains the single-policy label producer (Section III-A); for the mask-conditioned path it is repositioned as a *quality-control gate*. CHA runs on the rendered image, and the intersection-over-union between the conditioning mask and CHA’s prediction measures how faithfully the generator placed the dome where the mask asked. A sample whose IoU falls below a threshold is flagged: the rendered dome drifted from the requested silhouette, so the by-construction label no longer matches the pixels. The same shape-aware routing (convex hull / closing / skeleton) that makes CHA a class-agnostic annotator thus also makes it a class-agnostic verifier, and CHA’s standalone library [42] serves both roles unchanged. The gate’s reliability is itself contingent on the SandBoilNet checkpoint transferring to the synthetic renders; we quantify how tight that ceiling is in Section V-C.

d) Trade-off.: Mask-conditioned generation gives exact, checkpoint-independent provenance and lets a user draw a mask to obtain a paired (image, mask) record, at the cost of training one ControlNet per defect class and curating a bank of conditioning masks. The soft-mask path, in turn, needs no extra training and inherits the real dome geometry pixel-for-pixel, at the cost of the CHA re-prediction ceiling. We release both paths; the soft-mask path with CHA labelling produced the augmented dataset evaluated in this paper, and the mask-conditioned ControlNet is released as the checkpoint-independent, diversity-widening alternative with CHA as its verifier. We are careful to scope MaskCN’s claim. It is a strong *image* and *label-provenance* engine, but we do not assert a downstream segmentation benefit here; the gate that would certify its labels at scale is the present bottleneck (Section V-C).

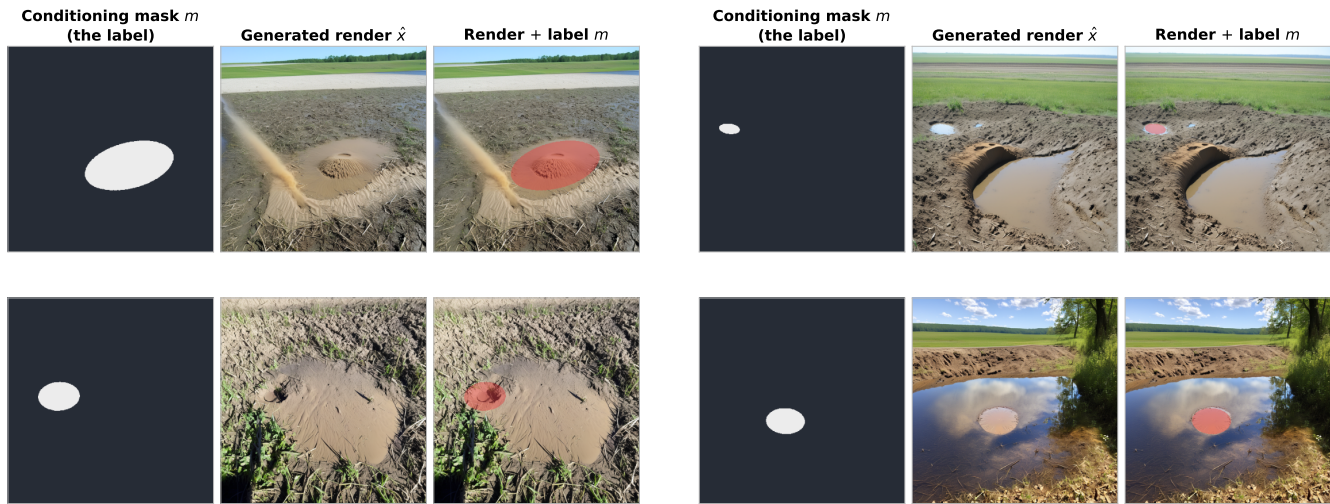


Fig. 7. Mask-conditioned generation (MaskCN) as a *label-by-construction* engine: the conditioning mask m is *exactly* the segmentation label of the image $\hat{x}$ it produces, so no manual annotation is performed. Each row is a triplet [conditioning mask m (the label) $\rightarrow$ generated render $\hat{x} \rightarrow$ render with m overlaid]. Masks are drawn from the 39-silhouette bank under per-instance jitter, spanning the requested-mask size range from small incipient domes to large mature cones; one mask blob yields one rendered boil. The right column overlays the conditioning mask on the render to show label registration; the overlays are shown verbatim, including imperfect cases, and the quantitative label-drift of these renders under the SandBoilNet-based gate is analysed separately in Section V-C. The distinct sampler (DPM++ 2M Karras) and mask-only conditioning (no denoising strength, depth, HED, or Canny) distinguish this path from the image-to-image presets of Section III-H.

G. Prompt Atlas

The Prompt Atlas turns a domain into a stratified, validated bank of text prompts. The builder’s input is a JSON domain specification with the following fields: the **concept** (here, *sandboil*); the **trigger_token** `sbx` introduced during DreamBooth training; a one-sentence **geometry** description used as system context when querying a language model; a **taxonomy** dictionary mapping axis names to value lists (we use eight axes: geographic setting, soil type, hydraulic regime, vegetation, season, weather, lighting, and camera perspective); a list of **negative_concepts** that prompts must not contain (e.g., *volcano*, *lava*, *cartoon*); a set of **exemplar_refs** pointing to real images used for image–text validation; a **prompt_template** with placeholders for axis values; and an explicit list of **cross_products** of axes whose combinations are to be expanded. The same builder accepts any specification of this shape, so the atlas extends to a new defect class by writing a new specification file without code changes.

The builder runs five stages.

a) Stage 1: candidate generation.: In template mode, the builder enumerates the requested axis cross-products and fills the template with one value from each axis, drawing the remaining axes by seeded random sampling. In language-model mode, a small open-weight model (Qwen2.5-3B [48]) is queried once per axis cell. The query uses a system prompt incorporating the geometry, the scientific context, and the negative-concept list, and the model returns three single-sentence variations.

b) Stage 2: length filter.: Prompts shorter than 12 words or longer than 60 words are discarded.

c) Stage 3: TF-IDF de-duplication.: Each prompt is vectorised under a unigram–bigram TF-IDF model and

the pairwise cosine similarity is computed. Whenever a prompt is at or above 0.85 cosine similarity with one already retained, it is dropped.

d) Stage 4: negative-concept screening.: A regular expression synthesised from the **negative_concepts** list is matched against each prompt; matches are removed.

e) Stage 5: image–text validation.: Each surviving prompt is encoded with a CLIP ViT-B/32 text encoder [14]; the average image embedding of the exemplar references is computed with the corresponding CLIP image encoder. Cosine similarity yields a ranking, and prompts below the tenth-percentile similarity are dropped.

A final stratified sampling over the cross-product cells yields the requested number of prompts. The output is a JSON Lines record per prompt with a stable identifier, the prompt text, the axis bucket, the CLIP image–text similarity score, and the source flag (template or language model). The sand boil atlas is built incrementally, and the production atlas used in this paper is a strict superset of earlier revisions.

f) Production atlas statistics.: The v1 sand boil atlas contains $N = 91$ prompts produced from 102 template candidates (target 100). The per-stage yield is $102 \rightarrow 102 \rightarrow 102 \rightarrow 102 \rightarrow 91$: the length, TF-IDF, and negative filters dropped no candidates, and the CLIP image–text validator removed the bottom 11. The CLIP similarity to the exemplar embedding is $\mu = 0.295 \pm 0.015$ (range [0.269, 0.331]); the tenth-percentile threshold is 0.276. Prompts average 43.3 words (min 31, max 55, within the 12–60 filter). Lexical diversity is low at the unigram level ($\text{distinct-1} = 0.041$) and higher at the bigram level ($\text{distinct-2} = 0.095$), reflecting a templated style with substantial axis-driven variation. Of the 44 taxonomy values declared across the eight axes, 40 (91%) appear in

at least one rendered prompt. The four missing values all belong to the **season** axis. As a known limitation, the v1 prompt template references seven of the eight axes and omits the season slot; season values are therefore sampled and recorded in the **fill** field but do not reach the rendered prompt text.

g) *Example prompts.*: Two representative atlas entries make the rendered style concrete:

“photo of sbx sandboil at the landside toe of an earthen levee, saturated foundation sand, steady low-velocity discharge, wetland reeds and rushes, drizzle with wet ground reflections, raking side light highlighting the dome relief, wide context shot showing the levee crest in the background, color-graded raw camera output.”

“photo of sbx sandboil along a riverbank during receding flood, saturated foundation sand, intermittent flow with sediment fan deposits, sparse short grass, foggy early morning, raking side light highlighting the dome relief, oblique angle inspector survey, professional documentary photography.”

Figure 8 summarises the builder and these statistics.

H. Production presets

The five-row ablation of Section IV-C traces individual design axes (each control branch, the LoRA, the inpainting path, the soft-edge branch) under fixed seeds and sources. A smaller, four-cell production-preset comparison is also useful as the practical distillation of those axes. Each preset bundles a complete recipe (control set, IP-Adapter use, denoising path) and is the unit a downstream operator actually picks. All four share the SDXL backbone, the DreamBooth-LoRA adapter, 45 diffusion steps at 1024×1024 , and the conditioning-scale schedule $(\lambda_c, \lambda_d, \lambda_h, \lambda_n) = (0.40, 0.25, 0.40, 0.40)$ on whichever branches each preset activates. IP-Adapter is loaded with scale 0.55, anchoring style to a real reference photograph rotated through the source pool. Classifier-free guidance is 8.5 for the image-to-image presets and is per-image jittered for the soft-mask inpainting preset ($\text{CFG} \in [7.5, 9.5]$); strength is jittered for the inpainting preset in $[0.85, 0.99]$.

The four presets span a single design axis that trades geometric constraint for style supervision. Two presets load three structural controls each (Canny plus two of Depth/HED/Normal) but no IP-Adapter, putting the burden of texture entirely on the denoiser’s residual capacity after the controls have been satisfied. One preset drops Canny and Depth, retains HED and Normal, and adds an IP-Adapter style anchor; the anchor injects high-frequency texture (granular sand, wet-dry transitions, photographic noise) that a structural control cannot represent. The fourth preset carries the heaviest conditioning stack (HED, Normal, IP-Adapter, plus the soft-mask inpainting protocol of Section III-E). This preset preserves the dome geometry through the diffusion loop, eliminates the Poisson seam-and-tint failure modes of the prior pipeline [12], [13], and exposes the IP-Adapter to a coherent unmasked context for style transfer. The inpainting preset is selected when geometric fidelity to a specific source-mask dome is paramount; the style-anchored image-to-image preset

is selected when style-rich sand boil texture without inpainting is preferred. Without intervention, the inpainting preset’s heavy conditioning would make every output a close variant of the same recipe. To prevent that, the deployed pipeline injects seeded per-image jitter on the HED and Normal scales, the IP-Adapter scale, the soft-mask erode kernel and Gaussian blur kernel, the sand boil attention boost, the CFG, and the strength. The same seed-derived RNG drives every jittered axis, so a given (source, seed, prompt) triple is fully reproducible.

I. Reference-set similarity quality filtering

DreamBooth fine-tuning on a small reference set can fail in two distinguishable directions. It can drift *off* the target domain when the prompt pushes the denoiser into a region of the prior the adapter did not anchor, yielding out-of-distribution outputs. It can also drift *onto* a memorised training frame when the prompt is close to one the adapter saw during training, yielding near-duplicates of the reference set. Both failures degrade downstream training: OOD samples teach the segmenter spurious correlations, and memorised samples inflate the effective duplication of a handful of real frames. The filter introduced here guards against both modes with no human review.

a) *The filter.*: Let $\Phi : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^D$ denote the CLIP [14] ViT-B/32 image encoder normalised to unit length. Compute the centroid embedding of the reference set, $\bar{\phi}^{\text{ref}} = \text{normalize}\left(\frac{1}{N_{\text{ref}}} \sum_{x \in \mathcal{R}} \Phi(x)\right)$, once at pipeline-initialisation time. For each synthetic image $\hat{x}$, the quality score is the cosine similarity to the centroid:

$$s(\hat{x}) = \Phi(\hat{x})^\top \bar{\phi}^{\text{ref}}. \quad (5)$$

Samples whose score falls outside the admissibility band $[s_{\text{lo}}, s_{\text{hi}}]$ are excluded from the augmented dataset.

b) *Auto-tuned thresholds.*: Rather than hand-pick the band, we compute the per-image leave-one-out CLIP similarity on the reference set itself: for each $x_k \in \mathcal{R}$, $s_k = \Phi(x_k)^\top \text{normalize}\left(\sum_{j \neq k} \Phi(x_j) / (N_{\text{ref}} - 1)\right)$. The empirical distribution $\{s_k\}$ defines what “in-distribution” looks like at the reference-set scale. We set

$$s_{\text{lo}} = \text{median}(\{s_k\}) - 2.5 \sigma(\{s_k\}), \quad s_{\text{hi}} = \max(\{s_k\}) + \epsilon, \quad (6)$$

with σ the median absolute deviation and ϵ a small positive offset (we use 0.01). The lower threshold rejects synthetic samples that are systematically less reference-like than the most peripheral real sample. The upper threshold rejects synthetic samples that are *more* reference-like than any real sample, which under the small-reference-set regime is the signature of memorisation rather than legitimate within-distribution generation.

c) *Properties.*: The filter is automated (no labelled data, no manual scoring), reproducible (the band is a deterministic function of the reference set and a fixed encoder), and content-aware (CLIP’s vision tower compresses both texture and semantic context, so a sand-boil-styled volcano scene and a near-duplicate sand boil

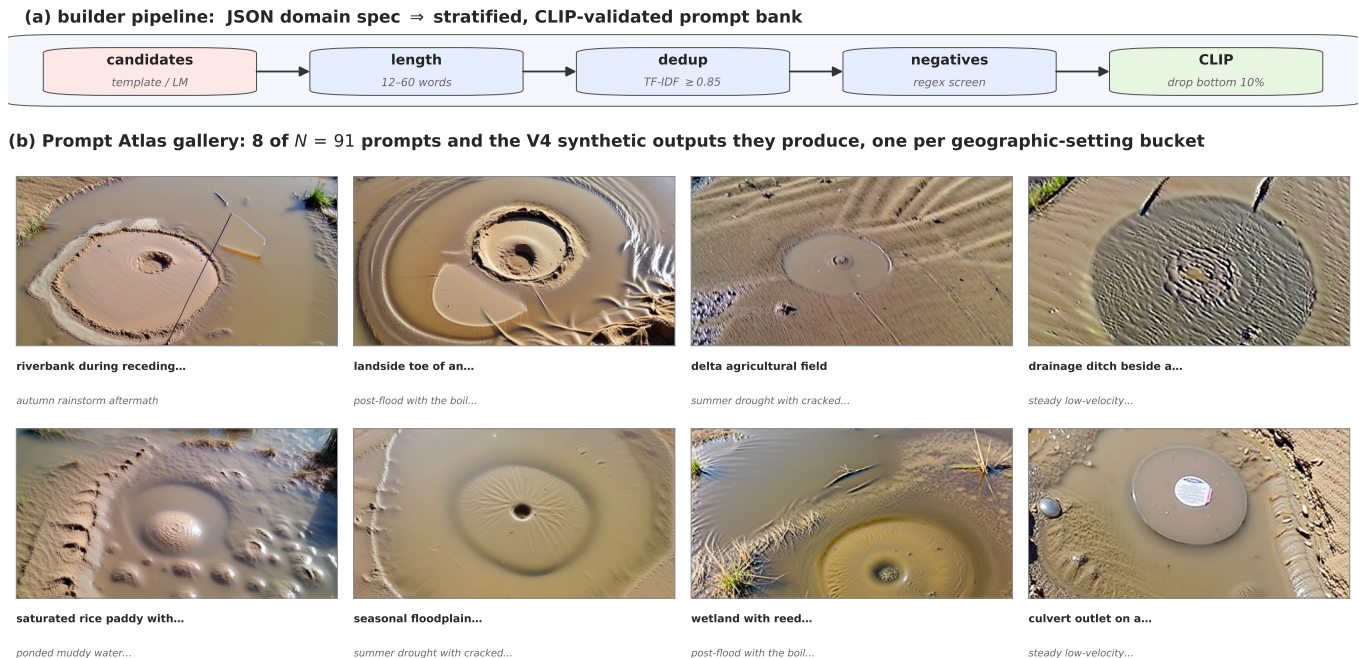


Fig. 8. The sand boil v1 Prompt Atlas, presented as a visual catalogue. (a) *Builder pipeline*. A JSON domain specification is expanded into candidate prompts (by template substitution or a small open-weight language model) and then passed through a length filter, a sparse TF-IDF de-duplication, a negative-concept regex screen, and a CLIP image-text validation against the exemplar embedding. The builder is reproducible from the specification alone and ports to a new defect class by writing a new specification file with no code change. (b) *Atlas gallery*. Eight of the $N=91$ production prompts paired with the V4 synthetic outputs they generate (one image per prompt, all from the same generator and identical scheduler/CFG), one per geographic-setting bucket in the taxonomy. Each caption gives the prompt’s primary axis value (bold) and one secondary axis value (italic) so the reader can map an output back to the stratification cell that produced it. The visual variety is the operational meaning of the atlas: a fixed defect geometry rendered under controlled scene variation. The per-prompt CLIP score distribution ($\mu=0.295 \pm 0.015$, $p_{10}=0.276$, yield $102 \rightarrow 91$) is reported in the text.

are rejected for different but principled reasons). The computation is approximately 4 ms per sample on a single GPU. The same filter transfers to any new defect class through its own reference centroid.

d) *Per-bucket diagnostic.*: A single global floor implicitly assumes every scene type sits the same distance from the reference centroid. To test that assumption we stratify the synthetic set by its Prompt-Atlas geographic-setting bucket and recompute, within each bucket, a bucket-local floor $s_{lo}^{(b)} = \text{median}(\{s_b\}) - 2.5\sigma(\{s_b\})$ from that bucket’s own scores. The global floor ($s_{lo} = 0.839$, set by the tight reference leave-one-out distribution) proves *stricter* than all bucket-local floors (which span 0.806–0.834 across buckets): it rejects 205 of the 1,020 synthetic images, against 121 when each bucket is judged by its own floor. The rejections are also far from uniform; they concentrate in scene types that sit systematically lower in centroid similarity (most of all the riverbank and delta-field buckets), exactly the cases a single threshold cannot tell apart from genuine out-of-distribution drift. We therefore report the per-bucket floors as a diagnostic of *where* the global filter is conservative, and keep the global floor in production because it errs toward dropping borderline samples.

J. Dataset assembly with source tagging

Each synthetic image is annotated with a pointer to the real source image from which it was derived, together with the prompt, seed, and generator configuration that produced it. The synthetic ground-truth mask is produced uniformly across all production presets by running the synthetic image through the CONVEX HULL ANNOTATOR pipeline of Section III-A: the public SandBoilNet checkpoint emits a per-pixel probability map, the threshold-label-hull cleanup of [12], [42] produces a binary mask, and that mask becomes the label for the synthetic record. Source tagging is preserved as an artefact-level discipline: every synthetic record carries the `source_image_id` of the real parent from which it was derived, even though the leak-free fold-filtering machinery that consumes those tags is left to future work.

K. Integration: what the contributions together enable

Three mutually reinforcing claims hold once the pipeline is assembled end-to-end.

a) *Single-policy mask provenance.*: Two coupled mechanisms guarantee that every synthetic image in the augmented set has a mask whose origin we can name. Defect-preserving soft-mask latent blending (Section III-E) ensures the defect *image pixels* are inherited

verbatim from the source. CONVEX HULL ANNOTATOR (Section III-A) produces every *mask label* from a single, reproducible post-processing pipeline applied uniformly across all production presets. The two mechanisms address different sides of the same problem (image-side preservation and label-side production) and together yield an augmented dataset with one mask-production policy and no per-preset special-casing. The mask-conditioned ControlNet of Section III-F offers a third, stricter provenance option in which the label is exact by construction and CHA becomes a verifier; we release it as a checkpoint-independent alternative to the soft-mask+CHA path used for the dataset here.

b) Automatic quality envelopes at both ends of the training pipeline.: Defect-size-adaptive soft-mask geometry (Section III-E) is an *input-side* quality envelope: it ensures the preservation margin and transition smoothness are well-scaled to each individual defect rather than tuned to an average dome. Reference-set similarity quality filtering (Section III-I) is the *output-side* envelope: it removes samples that fall outside the in-distribution band defined by the reference set’s leave-one-out similarity statistics. Both envelopes are automated, both are content-aware, and both add zero human-annotation effort to the per-defect-class extension cost.

c) Portability without code change.: The Prompt Atlas (Section III-G) reduces the per-class textual specification cost to writing a JSON file. CHA depends only on a class-specific segmenter checkpoint, the adaptive soft-mask geometry depends only on the binary mask geometry, and the quality filter depends only on a reference-set centroid in CLIP space. None of these depends on the sand boil class itself. Re-instantiating the pipeline on a seepage, sinkhole, or rutting domain therefore requires one new domain specification, one new DreamBooth adapter on a small new reference set, and one new segmenter checkpoint, and nothing else changes.

These integration properties make the augmented dataset *admissible as supervision* for a downstream cross-validated stacking ensemble, and they are the methodological return on the components developed above.

IV. EXPERIMENTS

This section describes the experimental protocol. It proceeds in three phases: a recap of the DreamBooth-adapter configuration, a generator-pipeline ablation, and a production-preset image-quality comparison against the real reference set.

A. Hardware and software

All experiments run on a single NVIDIA H100 NVL GPU (95 GB HBM). Diffusion-based image generation uses an fp16-mixed-precision PyTorch stack with the standard Diffusers inference pipelines. Two decoder settings are critical for image quality. First, the stock SDXL variational autoencoder produces blotch and colour artefacts when run in fp16 (a well-documented

SDXL issue); we therefore load the fp16-corrected VAE (`sdxl-vae-fp16-fix` [49]), which decodes without the fp16 colour artefacts. Second, because the 95 GB HBM removes any memory pressure, we *disable* VAE tiling and slicing: those split the decode into tiles and leave visible grid seams (a “torn” look) at 1024^2 . The Prompt Atlas builder runs on CPU only and builds the sand-boil prompt domain in under twenty seconds. The downstream segmentation work uses a PyTorch stack (`SEGMENTATION_MODELS.PYTORCH`).

B. DreamBooth adapter configuration

The adapter used by the production generator is a DreamBooth-LoRA module trained for 1,500 steps at rank 16 and learning rate $1e-4$ under a cosine schedule. Training uses class-image prior preservation (200 auto-generated images for the prompt *a photo of a sandboil*) on the curated reference set, with the trigger phrase **sbx sandboil**. The DreamBooth-LoRA recipe of Section III-B otherwise applies unchanged. Section III-C justifies restricting the configuration to the SDXL backbone.

C. Generator-pipeline ablation

To isolate the contribution of each generator component, we run a five-row ablation, with each row evaluated on the same five source images, prompts, and seeds. Table I summarises the ablation grid.

The last two rows answer two targeted questions. Row D tests whether adding a HED soft-edge ControlNet on top of Canny and depth improves visual quality. Row E tests whether the soft-mask inpainting path yields better seam behaviour than the plain image-to-image path of row D. The new rows share identical settings: 45 diffusion steps, classifier-free guidance scale 8.5, denoising strength 0.85, the DPM-Solver++ (order 3) scheduler [50], [51], 1024×1024 output resolution, and a full negative prompt that suppresses *volcano*, *lava*, *cartoon*, and other unrelated concepts. We chose the denoising strength of 0.85 after a coarse sweep on five sources. Lower values left the output too similar to the input, which reduced background diversity; higher values began to perturb the dome geometry on the non-inpaint rows. The ControlNet conditioning scales active in this ablation are $(\lambda_c, \lambda_d, \lambda_h) = (0.40, 0.25, 0.40)$ for Canny, depth, and HED soft-edge respectively. These are the first three entries of the full four-way schedule $(\lambda_c, \lambda_d, \lambda_h, \lambda_n) = (0.40, 0.25, 0.40, 0.40)$; the surface-normal branch λ_n is inactive here. The soft mask in row E uses the fixed 12-pixel erosion followed by a 41-pixel Gaussian blur for this ablation; the deployed V4 preset instead uses the adaptive radii of Section III-E. The five source images span varied environments (rice paddy, levee toe, riverbank, foggy field, ponded depression). The five prompts are drawn from the Prompt Atlas (Section III-G) in a stratified manner, so that each source maps to a different geographic-setting cell.

Table II consolidates the generation hyperparameters for reproducibility. The fixed settings are shared across

TABLE I

GENERATOR-PIPELINE ABLATION MATRIX. EACH ROW PRODUCES FIVE IMAGES ON THE SAME SOURCES, PROMPTS, AND SEEDS. CN COLUMNS MARK THE CONTROLNETS ACTIVE IN THIS ABLATION (CANNY, DEPTH, HED SOFT-EDGE; THE SURFACE-NORMAL BRANCH IS EXERCISED ONLY IN THE PRODUCTION PRESETS); LoRA MARKS WHETHER THE DREAMBOOTH ADAPTER IS LOADED; PATH IS EITHER IMG2IMG (WHOLE IMAGE) OR SOFT-MASK INPAINTING (PRESERVES THE DOME CORE).

ID	LoRA	Canny CN	Depth CN	HED soft-edge CN	Path	Notes
A	–	–	–	–	img2img	SDXL baseline (no conditioning)
B	–	✓	✓	–	img2img	Canny + Depth
C	✓	✓	✓	–	img2img	+ LoRA
D	✓	✓	✓	✓	img2img	+ HED soft-edge
E	✓	✓	✓	✓	soft-mask inpaint	+ soft-mask inpaint

TABLE II

GENERATION HYPERPARAMETERS FOR THE SAND-BOIL SYNTHESIZER. FIXED SETTINGS ARE HELD ACROSS PRESETS. THE DENOISING STRENGTH AND CONTROLNET CONDITIONING SCALE ARE THE TWO PARAMETERS WITH THE LARGEST EFFECT ON THE DIVERSITY/FIDELITY TRADE-OFF; THE SELECTED V4 PRESET JITTERS THE STRENGTH IN [0.85, 0.99].

Parameter	Value
Backbone	SDXL base 1.0
Domain adapter	DreamBooth-LoRA (rank 16, 1,500 steps)
Adapter weight / trigger	1.0 / sbx sandboil
ControlNet stack	Canny + Depth + Normal + HED soft-edge
preset V1	Canny + Depth + HED soft-edge
preset V2	Canny + Normal + HED soft-edge
Scheduler	DPM-Solver++ (order 3, Karras)
Inference steps	45
Guidance scale (CFG)	8.5
Resolution	1024 × 1024
Random seed	deterministic per image
Prompts	stratified Prompt Atlas
img2img denoising strength	0.85 (V4: [0.85, 0.99])
ControlNet conditioning scale	0.40

presets; the two parameters with the largest effect on the output are the image-to-image denoising strength and the ControlNet conditioning scale, which trade background diversity against fidelity to the parent’s boil. The diversity-oriented production setting (strength = 0.85, scale 0.40) maximises background variety, and the selected V4 preset jitters the strength in [0.85, 0.99] around it (Section III-H). Because the synthetic ground-truth mask is produced by the CONVEX HULL ANNOTATOR from the rendered image itself (Section III-A), the label tracks the boil as actually rendered; out-of-distribution and memorised samples are then removed by the leave-one-out CLIP admissibility filter (Section III-I).

D. Production-preset comparison

The deployed pipeline exposes four production presets, summarised in Table III. For all four presets, the synthetic ground-truth mask is produced uniformly post hoc by the CONVEX HULL ANNOTATOR pipeline of Section III-A.

Each preset is evaluated on a fixed batch of 182 renders (the per-preset evaluation volume reported in Table VII). We evaluate each preset’s samples under the protocol of Section IV-F: Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) against the real reference set on Inception-v3 features, CLIP score against each sample’s own Prompt Atlas prompt, and Learned Perceptual Image

Patch Similarity (LPIPS) diversity within the per-preset set. The seed cascade, source rotation, and Prompt Atlas stratification match those of Section IV-E.

a) *Mask-conditioned generation (MaskCN).*: Alongside the four img2img presets we evaluate the mask-conditioned ControlNet of Section III-F under the same image-quality protocol. MaskCN is a text-to-image path with mask-only structural conditioning; it uses no source photograph and therefore has no denoising-strength, depth, HED, or Canny inputs. It shares the SDXL backbone, the DreamBooth-LoRA adapter, classifier-free guidance 8.5, and the fp16-corrected VAE with the other presets, but swaps the scheduler for DPM++ 2M Karras. Masks are drawn from a 39-silhouette bank, yielding one boil per render; per-instance jitter (flip, rotation, rescale, reposition) spans boil shape, size, and placement across draws. The diversity singles folded into the augmentation mixture use the default conditioning scale (1.0), whereas the MaskCN image-quality numbers in Table IV are measured at the best evaluation configuration selected by the sweep in Section V-C (conditioning scale 1.8, 45 steps, DPM++ 2M Karras, with best-of-three retention). This diversity contribution is also evaluated qualitatively in Section V-D; the label-verification gate is examined separately in Section V-C.

E. Augmented dataset generation

The augmented training set is produced by the V4 production preset (Table III) at 1024 × 1024 resolution, with approximately five synthetic variations per real training image. The production run yields 1,020 synthetic candidates in total; the leave-one-out CLIP admissibility filter (Section III-I) then admits 815 of them into the augmented dataset and rejects 205 as out-of-distribution. Seeds are assigned deterministically, and prompts are drawn round-robin from the Prompt Atlas so that the synthetic set is uniformly stratified across the taxonomy axes. Section IV-G describes the deterministic seed cascade and the full parameter manifest.

F. Synthetic-image quality evaluation

The image-quality evaluation protocol assesses the augmented dataset against the full real evaluation set ($N_{\text{eval}} = 199$ images) on two families of measures. We use the full

TABLE III

PRODUCTION PRESETS EXPOSED BY THE DEPLOYED PIPELINE. ALL SHARE THE SDXL BACKBONE, THE DREAMBOOTH-LoRA ADAPTER, 45 DIFFUSION STEPS AT 1024×1024 , AND CONDITIONING-SCALE SCHEDULE $(\lambda_c, \lambda_d, \lambda_h, \lambda_n) = (0.40, 0.25, 0.40, 0.40)$ ON WHICHEVER BRANCHES EACH PRESET ACTIVATES. “IPA” IS IP-ADAPTER [29] AT SCALE 0.55 WITH A REAL REFERENCE ROTATED THROUGH THE SOURCE POOL. CFG IS 8.5 FOR V1–V3; V4 JITTERS CFG IN [7.5, 9.5] AND STRENGTH IN [0.85, 0.99].

ID	Control set	IPA	Path	Notes
V1	Canny + Depth + HED soft-edge	–	img2img	Three structural controls (Depth), no style anchor
V2	Canny + Normal + HED soft-edge	–	img2img	Three structural controls (Normal), no style anchor
V3	HED soft-edge + Normal	✓	img2img	Style-anchored image-to-image
V4	HED soft-edge + Normal	✓	soft-mask inpaint	Style-anchored soft-mask inpainting (selected)

199-image set rather than the smaller curated 39-image reference because the latter leaves the 2048-dimensional Inception covariance rank-deficient and inflates FID, a known small-sample bias [52]. These 199 images are the full real training pool, which also serves as the source set for the image-to-image presets; FID against it is therefore an in-distribution reference, and we rely on KID and the precision/recall decomposition to separate fidelity from diversity. Because the soft-mask path preserves real dome pixels, its synthetic images share pixel content with this reference set, which biases its distribution distances downward and makes them not directly comparable across families to MaskCN (which uses no source photograph) or the Poisson baseline; we therefore read these absolute distances within rather than across the generation families, and rest the selection argument on the fidelity/diversity decomposition and on label provenance. A disjoint held-out real set is reserved for downstream evaluation and is not used for these distribution metrics. The first is distribution distance: FID [53] and KID [54] (mean and standard deviation) on Inception-v3 [55] features. We report FID with two implementations (standard torchmetrics and clean-fid [56], which removes resize and quantisation aliasing) and treat the unbiased KID as the primary distribution metric at this sample size. We further report CLIP score [14], [39] on the OpenAI CLIP ViT-B/32 model and LPIPS diversity [57] (AlexNet variant, at most 16 randomly sampled images for tractability). The second is a fidelity–diversity decomposition that a single FID number cannot provide: improved precision and recall [15] and density and coverage [58], computed on the same 2048-dimensional Inception features using k -nearest-neighbour manifold radii ($k=5$). Precision and density measure how much synthetic mass falls inside the real manifold (fidelity); recall and coverage measure how much of the real manifold the synthetic set spans (diversity). We compute the distribution measures for the full augmented set and for each production preset, and the precision/recall/density/coverage decomposition for each production preset, MaskCN, and the Poisson baseline. The generator ablation (rows A–E) is instead assessed qualitatively (Section V-A). Together these separate “looks like the references” from “covers the references”, which is the crux of the preset trade-off in Section V. Each configuration is scored on a single deterministic-seed batch, so KID is the only measure we report with a spread;

differences of a few FID points or a few thousandths of CLIP or LPIPS should be read as indicative rather than statistically separated.

G. Reproducibility

Every reported number is reproducible. Each numerical entry in Section V is the deterministic product of three identifiers: the trained-adapter checkpoint, the Prompt Atlas record, and the per-image seed. Each synthetic record carries a JSON sidecar with the source image identifier, prompt identifier, seed, ControlNet conditioning scales, and the production preset. The released code emits a manifest of every adapter, prompt-bank revision, and seed cascade used to produce a given augmented set. From the released code and this manifest, a reader can re-derive every reported number without manual reconfiguration.

V. RESULTS AND ANALYSIS

This section reports the generator-pipeline ablation; the production-preset image-quality comparison; the mask-conditioned ControlNet (MaskCN) results; a curated two-engine diversity gallery and source-paired qualitative outputs; a fidelity–diversity decomposition and a Poisson seamless-cloning baseline; the empirical behaviour of the CLIP admissibility filter and the memorisation audit; and the wall-clock cost on a single NVIDIA H100 NVL GPU.

A. Generator-pipeline ablation

To isolate how each conditioning stage shapes output quality, Figure 9 shows the qualitative outputs of the five ablation columns on the same five source images. We focus on two observations, concerning columns D and E.

Column D (img2img with soft edges). Adding the soft-edge branch on top of Canny and depth recovers fine-scale rim structure that the binary Canny map cannot resolve, in particular the wet–dry transition and the granular sand texture. On sources whose rim is the most pronounced, the additional branch eliminates the painterly artefact present in column C. The cost is one extra ControlNet forward pass per diffusion step, which raises VRAM by approximately 1.5 GB and per-step latency by roughly 12%.

Column E (soft-mask inpainting). The soft-mask path retains the source sand-boil region pixel-for-pixel and re-renders only the surrounding scene. Column D denoises

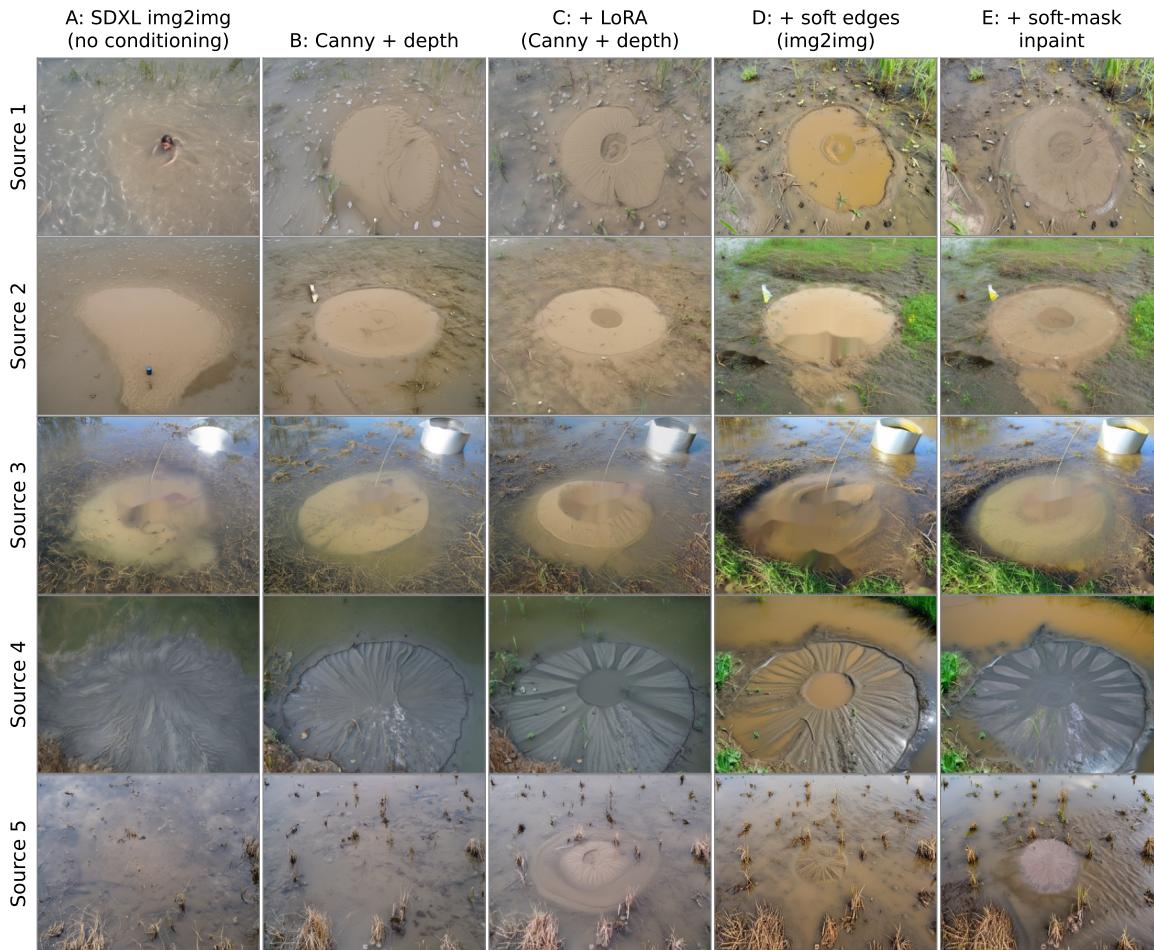


Fig. 9. Qualitative outputs of the five generator-pipeline ablation columns on the same five sources, prompts, and seeds. Columns: **A** SDXL img2img (no conditioning); **B** Canny + depth; **C** Canny + depth with the DreamBooth-LoRA adapter; **D** **C** plus a soft-edge (HED) ControlNet; **E** the soft-mask inpainting path of Section III-E. From A to E, the dome becomes progressively better localised and the rim texture sharper; column E produces the seam-free, dome-preserving outputs selected for the augmented dataset.

every pixel and may slightly perturb the dome geometry. In contrast, column E preserves the dome exactly and produces a seam-free, colour-consistent boundary, eliminating the two failure modes of the Poisson seamless-cloning step used in our previous work [12]. The dome is preserved pixel-for-pixel in image space, but the synthetic ground-truth label is not inherited from the source mask. As discussed in Section III-A, the soft-edge blending region permits a small amount of rim drift, so the label is re-predicted by the CONVEX HULL ANNOTATOR pipeline on the rendered synthetic image. This keeps mask provenance uniform across all production presets.

B. Synthetic-image quality across production presets

The image-quality measures, applied to the four image-to-image production presets of Section III-H and to the mask-conditioned ControlNet of Section III-F, isolate the trade-off between geometric over-conditioning and style-anchored texture. We read this table as a *menu of complementary engines*, not a ranking: no single preset dominates, and minimum distribution distance is deliberately

not the selection criterion. For data destined to augment a segmenter, label reliability and scene diversity matter more than proximity to the small reference set—indeed the lowest-FID options (V1 here, and the degenerate Poisson baseline of Section V-E) hug the reference distribution rather than expand it. Accordingly, the released augmented set is generated by the single *label-reliable* preset (V4, Section IV-E). A curated *mixture* across these complementary engines—V1 for manifold coverage, V3/MaskCN for diversity, V4 for real-structure labels—is the natural augmentation strategy, to be evaluated in future work. Each preset is evaluated at 1024×1024 against the full real reference set ($N_{\text{eval}} = 199$); results are reported in Table IV. We compute FID with two independent implementations (standard torchmetrics and clean-fid [56]), which agree throughout, and treat the unbiased KID as the primary distribution-match metric at this sample size.

The synthetic ground-truth mask is produced uniformly across all four image-to-image production presets V1–V4 by the CONVEX HULL ANNOTATOR pipeline of Sec-

TABLE IV

SYNTHETIC-IMAGE QUALITY FOR THE FOUR IMAGE-TO-IMAGE PRODUCTION PRESETS AND THE MASK-CONDITIONED CONTROLNET (MASKCN), EVALUATED AT 1024×1024 AGAINST THE FULL REAL REFERENCE SET ($N_{\text{eval}}=199$). FID IS REPORTED UNDER TWO IMPLEMENTATIONS (TORCHMETRICS AT FEATURE DIMENSION 2048, AND CLEAN-FID [56]), WHICH AGREE WITHIN THREE POINTS. LOWER IS BETTER FOR FID/KID, HIGHER FOR CLIP IMAGE-TEXT SIMILARITY AND LPIPS WITHIN-SET DIVERSITY; THE BEST VALUE PER COLUMN IS IN BOLD. V4 (†) IS THE PRESET SELECTED FOR THE AUGMENTED DATASET (SECTION IV-E).

Preset	FID ↓	clean-fid ↓	KID ↓	CLIP ↑	LPIPS ↑
V1	169.1	169.4	0.048 ± 0.008	0.302	0.595
V2	198.7	197.0	0.073 ± 0.007	0.295	0.588
V3	193.9	192.8	0.074 ± 0.007	0.311	0.644
V4†	201.8	201.2	0.091 ± 0.010	0.302	0.633
MaskCN	189.4	190.1	0.048 ± 0.006	0.317	0.646

tion III-A. This pipeline runs the public SandBoilNet checkpoint over each synthetic image and convex-hull cleans the prediction before it is admitted to the training set. The policy holds even for V4. Soft-mask inpainting preserves the dome geometry of the source mask more faithfully than the image-to-image presets V1–V3, yet we deliberately do not inherit the source mask as the label. As a result, every synthetic sample has the same mask provenance, and the dataset carries a single, reproducible label-production policy. The mask-conditioned ControlNet is the one exception to this policy (it is designed to make the conditioning mask the label by construction), and we examine how well that design holds in Section V-C.

C. Mask-conditioned ControlNet and the limits of CHA as a gate

The mask-conditioned ControlNet (MaskCN, Section III-F) attains the strongest KID, CLIP, and LPIPS scores of any preset in Table IV: it ties V1 for the lowest KID (0.048), attains the highest CLIP image-text similarity (0.317) and the highest LPIPS within-set diversity (0.646), and its FID (189.4, clean-fid 190.1) is second only to V1. But, as we show next, its labels cannot yet be certified by the available gate, and it is therefore *not* adopted as the production preset. By generating from a chosen binary mask rather than from a source photograph, it promises an exact label by construction: the conditioning mask *is* the ground-truth label.

We test that promise with the CONVEX HULL ANNOTATOR (CHA) repurposed as a quality-control gate: for each render we run the public SandBoilNet checkpoint and measure the IoU between the conditioning mask and the predicted dome (the “drift”). The result is negative but informative. Across the retained set the drift IoU has median 0.15 and mean 0.17, and no sample clears the IoU ≥ 0.5 acceptance gate. As Figure 10 shows, the renders are realistic and placed at the requested mask, so the cause is not principally generator error but a *transfer* failure of the gate. SandBoilNet, trained on real photographs, alternately under- and over-segments the synthetic renders, so

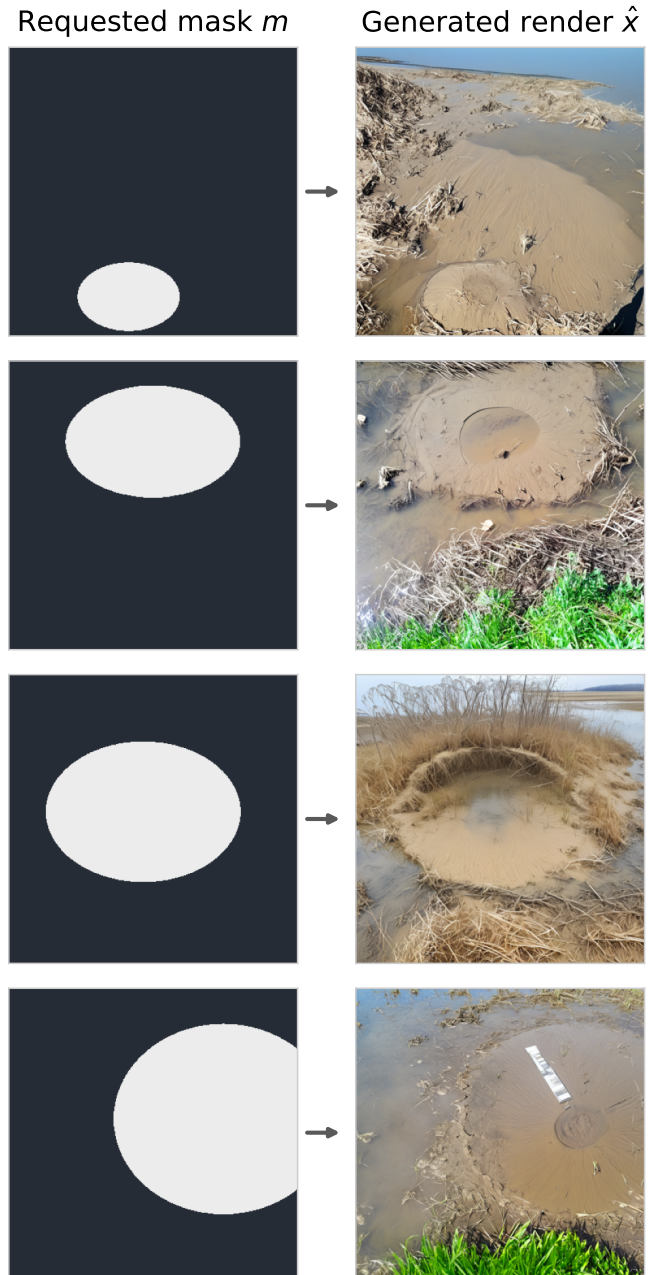


Fig. 10. Mask-conditioned ControlNet: controllable generation. Each row pairs a requested binary mask m (left) with the realistic sand boil the model generates at that location (right), across four examples spanning the mask-size range (not cherry-picked). Because the image is generated *from* the mask, the mask is the ground-truth label by construction. Whether the automated CONVEX HULL ANNOTATOR gate can *certify* that label on these synthetic renders (where it under- and over-segments, yielding a median mask-render IoU of only 0.15, i.e. large drift) is taken up in the surrounding text.

the IoU it reports is an unreliable proxy for true mask adherence on this new generation path.

To separate the two effects we swept, on a held-out subset, the generation controls that govern adherence: the ControlNet conditioning scale ($\{1.0, 1.8\}$), the number of inference steps ($\{30, 45\}$), and the sampler (DPM++ 2M Karras vs. the stock scheduler). We scored each configuration by drift IoU and by FID against the 199-image reference. Raising the conditioning scale improved the drift IoU only marginally ($0.137 \rightarrow 0.143$), and the sampler and step count made no measurable difference. This confirms that the residual drift the gate reports is dominated by the segmenter’s transfer behaviour rather than by a generation setting we can tune away. At the best configuration (conditioning scale 1.8, 45 steps, DPM++ 2M Karras) we additionally applied a best-of-three retention step, keeping the candidate with the highest CLIP-IQA no-reference quality [47] per prompt; this lifted the mean drift IoU to 0.17 and produced the image-quality numbers reported above.

The practical consequence is that MaskCN is an excellent *image* generator whose *label* reliability cannot yet be certified by the available gate. We therefore retain V4 (soft-mask inpainting), whose dome is a real preserved structure, as the production preset for the augmented dataset, and report MaskCN as a strong complementary path whose label-verification gate is the current bottleneck rather than its image quality.

Two properties nonetheless make MaskCN a first-class *diversity* engine alongside the img2img presets, independent of the unresolved gate. First, its labels are exact *by construction* (Section III-F); the drift IoU above measures only how well the available SandBoilNet-based gate can *verify* that label on this new render distribution, not a defect in the label itself. Second, by decoupling the synthetic label geometry from the catalogue of real source domes, MaskCN can place a boil of a chosen shape, size, and position into a scene, independent of where a real dome happened to sit in any single source image. Figure 7 makes the label-by-construction guarantee concrete; the resulting cross-engine diversity is shown in Section V-D. We make no claim here that adding MaskCN samples improves downstream segmentation IoU; that evaluation is ongoing and out of scope for this paper. We claim only that MaskCN contributes exact zero-annotation labels and a per-instance variety of boil shape, size, and placement that the img2img production preset cannot produce by design.

D. Qualitative diversity across both generation engines

Figure 11 presents the two generation engines side by side under a single curation policy. Row (A) shows img2img preset outputs: each synthetic image retains the dome geometry of its real source under multi-ControlNet structural conditioning and acquires a varied surrounding scene from the Prompt Atlas, spanning soil type, lighting, vegetation, and viewpoint. Row (B) shows the mask-conditioned engine, whose *singles* render one boil into a

chosen silhouette at an arbitrary size and placement, with the conditioning mask serving as the exact label. The two engines are therefore complementary diversity sources: img2img widens the *scene* distribution around real defect geometry, while MaskCN supplies exact zero-annotation labels and places boils at sizes and positions not tied to any single real source. The gallery is curated for scene variety, not ranked by any quality metric, and accordingly implies no claim about downstream segmentation accuracy.

Within the MaskCN row, per-instance mask jitter further broadens the engine’s output: perturbing each mask’s boil shape, size, and position turns a fixed set of conditioning masks into a wide spread of boil sizes and off-centre placements rather than collapsing to a few repeated layouts.

Beyond boil morphology, the mask-conditioned generator also varies the surrounding *scene* (season, weather, and ground cover) through the Prompt Atlas. Such background variety is, in principle, what keeps a downstream segmenter from keying on a single “muddy levee” context, though we make no segmentation claim here. We bound this claim conservatively, however: the variation stays within the natural-levee / floodplain domain, because the DreamBooth-LoRA and mask-ControlNet priors are strong enough that prompts for urban, night, or snow scenes collapse back to a levee setting—a limit we flag rather than overstate, since the generator does not produce that wider scene range.

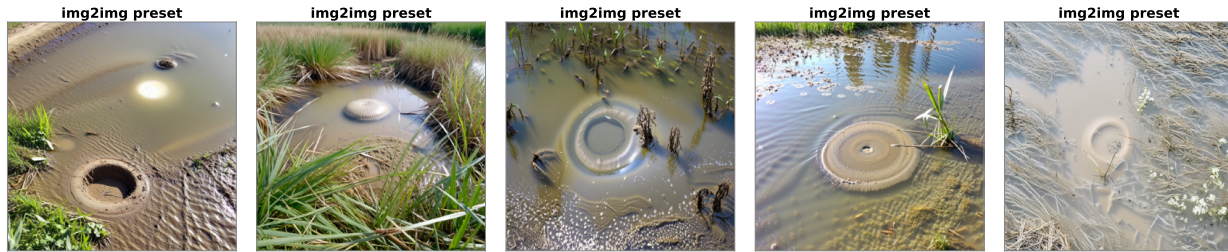
Figure 12 complements the diversity gallery by placing each img2img synthetic image directly beside its real source, so that source-to-synthetic dome preservation is legible row by row. The synthetic images preserve the dome geometry of the source through the soft-mask inpainting path (up to a small rim drift in the soft-blend region) and acquire varied surrounding scenes from the Prompt Atlas, including different soil types, lighting, vegetation, and viewpoint. The synthetic ground-truth masks track the rendered dome.

E. Comparison against a Poisson seamless-cloning baseline

A natural non-generative reference point is the Poisson seamless-cloning composition strategy [13] adopted in our earlier work [12]: cut the masked sand-boil region from one real reference image and paste it into a different real reference photograph under Poisson colour blending, with no diffusion model in the loop. We reproduce that baseline (five Poisson composites per source over the curated foregrounds, 1024×1024) and evaluate it against the same full 199-image reference under the same FID/KID/LPIPS measure (Table V).

a) *Why FID alone favours the baseline, and why the baseline is nonetheless not the better source of training data.*: The Poisson baseline still scores lower on FID (135.0 against 201.8 for V4) and KID (0.021 ± 0.006 against 0.091 for V4), though the margin is far narrower against the full reference than against the curated 39-image set, where the

(A) img2img presets — multi-ControlNet, strength 0.85



(B) MaskCN singles — mask-conditioned ControlNet, per-instance mask jitter, conditioning scale 1.0



Fig. 11. Curated best-diverse gallery of synthetic sand boils, arranged as two labelled rows. (A) **Img2img presets** (multi-ControlNet on the DreamBooth-LoRA SDXL backbone, denoising strength 0.85; samples drawn from the V2-curated and V3 production presets of Section III-H). (B) **MaskCN singles** (mask-conditioned ControlNet, DPM++ 2M Karras, per-instance mask jitter, conditioning scale 1.0). Tiles are hand-curated for variety, not ranked by any metric; the panel illustrates image and scene variety only and makes no segmentation-quality claim. The img2img path widens the *scene* around a fixed real dome, while MaskCN places a boil of a chosen shape, size, and position via its conditioning mask, both within the *natural-levee / floodplain domain*: the strong DreamBooth-LoRA and mask-ControlNet priors mean prompts for urban, night, or snow collapse back to a levee scene.

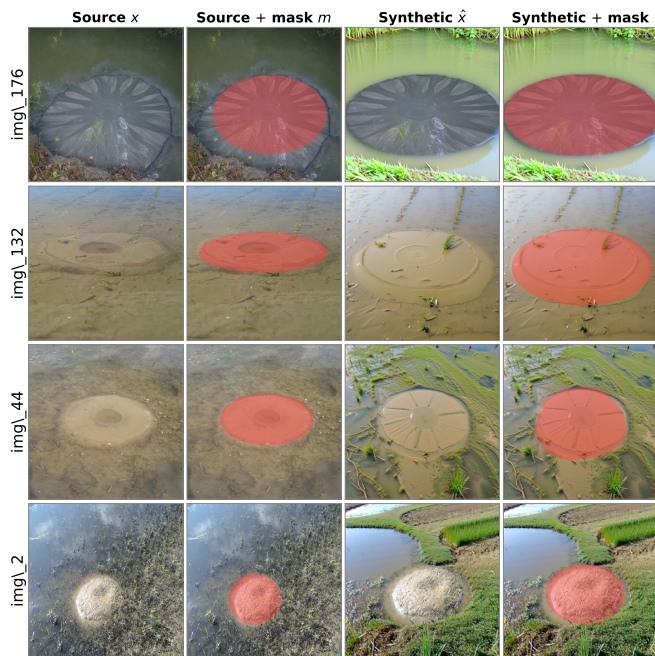


Fig. 12. Source-paired synthetic outputs (one source per row). Columns: real source x ; x with ground-truth mask m overlaid; synthetic $\hat{x}$ produced under the soft-mask inpainting preset; $\hat{x}$ with its synthetic mask overlaid. The dome geometry is preserved between x and $\hat{x}$ by the soft-mask path (Section III-E); the surrounding scene varies under Prompt Atlas conditioning. Synthetic masks shown are human-verified corrections; CONVEX HULL ANNOTATOR is the default label for uncorrected samples.

same baseline scored an artificially low FID of 101. This is the expected behaviour, not evidence in favour of the baseline. FID and KID measure the distance between feature distributions. The Poisson baseline’s outputs are built by pasting *real* sand-boil foregrounds onto *real* reference photographs, so its output distribution is a near-subset of the real reference distribution and the distance to the reference is artificially small. Three properties separate the two methods qualitatively despite the FID gap. First, the Poisson baseline cannot vary the defect geometry beyond the 39 real-foreground templates: every synthetic sand boil is one of the 39 real foregrounds, at most shifted in the frame, so the effective foreground catalogue is fixed and small. Second, the Poisson colour-blend step has known seam and colour-cast failure modes [12], [13], visible on most composites in side-by-side qualitative inspection. Third, LPIPS within-set diversity is close between the two methods ($V4 = 0.633$ vs. baseline = 0.620), so the Poisson baseline’s headline FID advantage does not translate into scene-level variation. We conclude that FID against a small reference set is a necessary but not sufficient quality measure for an augmented training set. The soft-mask inpainting preset is preferred because it produces (i) seam-free boundaries, (ii) controlled re-rendering of the surrounding scene under Prompt Atlas conditioning, and (iii) a synthetic dataset that is not trivially derivable from the reference set itself.

TABLE V

POISSON SEAMLESS-CLONING BASELINE VERSUS THE PROPOSED SOFT-MASK INPAINTING PRESET (V4), BOTH EVALUATED AGAINST THE FULL 199-IMAGE REAL REFERENCE AND THE SAME NUMBER OF SYNTHESISED OUTPUTS PER SOURCE. LOWER IS BETTER FOR FID/KID, HIGHER FOR LPIPS DIVERSITY. THE BASELINE REUSES REAL FOREGROUNDS (NO GENERATIVE MODEL), SO ITS FID AGAINST THE REFERENCE IS FAVOURABLE WHILE ITS SCENE DIVERSITY IS BOUNDED BY THE REFERENCE SET ITSELF; CLEAN-FID AGREES WITH THE FID SHOWN TO WITHIN THREE POINTS. BEST VALUE PER COLUMN IS IN BOLD; THE PROPOSED METHOD’S ROW LABEL IS BOLD.

Method	FID ↓	KID ↓	LPIPS ↑
Poisson seamless-cloning baseline	135.0	0.021 ± 0.006	0.620
V4 (proposed, soft-mask inpaint)	201.8	0.091 ± 0.010	0.633

TABLE VI

FIDELITY-DIVERSITY DECOMPOSITION AGAINST THE FULL 199-IMAGE REFERENCE SET (THE SAME REFERENCE AS TABLE IV). PRECISION AND DENSITY MEASURE SYNTHETIC MASS INSIDE THE REAL MANIFOLD (FIDELITY); RECALL AND COVERAGE MEASURE HOW MUCH OF THE REAL MANIFOLD THE SYNTHETIC SET SPANS (DIVERSITY). ALL ON 2048-D INCEPTION FEATURES WITH $k=5$ NEAREST-NEIGHBOUR MANIFOLD RADII. V4 (†) IS THE SELECTED PRESET; THE BEST VALUE IN EACH COLUMN AMONG THE GENERATIVE PRESETS IS IN BOLD.

Preset	Precision ↑	Recall ↑	Density ↑	Coverage ↑
V1	0.656	0.497	0.577	0.673
V2	0.478	0.307	0.309	0.432
V3	0.517	0.171	0.454	0.482
V4†	0.315	0.226	0.149	0.322
MaskCN	0.829	0.236	1.249	0.663
Poisson baseline	0.915	0.397	1.266	0.804

F. Decomposing fidelity from diversity

FID and KID collapse fidelity and diversity into one scalar, which is why V1 and V4 are hard to separate on those measures alone. The Improved Precision and Recall [15] and Density and Coverage [58] decomposition, computed on the same Inception features, pulls the two apart (Table VI).

Three findings follow. First, among the image-to-image presets the decomposition tracks the FID ordering: V1 sits closest to and best spans the real manifold (precision 0.656, the highest recall 0.497), while V4 has the lowest precision (0.315) and density—the signature of a preset that ventures furthest from the tight real manifold, which is consistent with its high LPIPS within-set diversity in Table IV (0.633, the second-highest among the image-to-image presets, after V3). V4 is thus *not* selected for manifold fidelity, on which V1 dominates it, but for its single-policy soft-mask label provenance and its scene-level diversity. Second, MaskCN has the highest precision (0.829) and density (1.249) of any generative preset: it sits very close to the real manifold, consistent with its low FID/KID, at modest recall (0.236): a high-fidelity, moderate-diversity profile. Third, the Poisson baseline posts the highest precision (0.915) and density (1.266), high recall (0.397), and the highest coverage (0.804) in the table: because it pastes real defect foregrounds onto

real backgrounds, its samples lie *inside* the real manifold by construction. Its strong recall and coverage reflect this reuse of real content rather than genuinely new scenes—the manifold coverage a degenerate copy-paste route attains without adding novelty. This is the classic degenerate-augmentation signature, and the same effect that flatters its FID in Section V-E. The decomposition therefore confirms that distribution-matching scores must be read alongside the provenance of the samples being scored, and that V4’s lower fidelity numbers reflect real scene synthesis rather than a defect.

G. Empirical demonstration of the CLIP quality filter

Section III-I introduces a post-generation CLIP admissibility filter whose lower and upper band thresholds are computed from leave-one-out reference-set statistics. We apply the filter to the full 1,020-image V4 augmented set without modification.

- Reference-set leave-one-out similarity to the cohort centroid: median 0.889, MAD 0.020.
- Admissibility band: $[s_{lo}, s_{hi}] = [0.839, 0.963]$.
- Augmented-set scores: mean 0.868, range $[0.627, 0.946]$.
- Filter verdict: 815 accepted, 205 below band (out-of-distribution drift), 0 above band (memorisation signature).

The filter rejects a non-trivial fraction of generations without any human review or hand-picked threshold. The rejected examples fall into one of two failure modes whose causes are diagnostically different: OOD drift and memorisation; here every rejection is OOD drift.

H. Memorisation audit

A recurring concern with DreamBooth-fine-tuned diffusion models on small reference sets is overfitting to individual reference frames. We evaluate this directly against the 39-image curated set the adapter was *trained* on (the relevant reference for memorisation, as distinct from the 199-image evaluation set used for FID/KID). For every synthetic image, we compute the cosine similarity of its CLIP-ViT-B/32 embedding to its *nearest* reference image. The nearest-reference similarity distribution over the 1,020-image V4 augmented set (Figure 13) has mean 0.860 ± 0.029 , with 99th-percentile 0.913 and a maximum of 0.925. The maximum is comfortably below an identity-level value (cosine ≈ 1.0), so no synthetic sample is a CLIP-space duplicate of any reference frame, and even the most reference-like 1% of synthetic samples (99th-percentile similarity 0.913) stay well below that identity level.

I. Interpretability of the conditioned generator

Two views make the pipeline legible: *how faithful* the generated distribution is to the real one, and *where* the class concept is grounded. (The conditioning signals themselves, and what each ControlNet branch contributes, are shown in Figure 5 of Section III-D.)

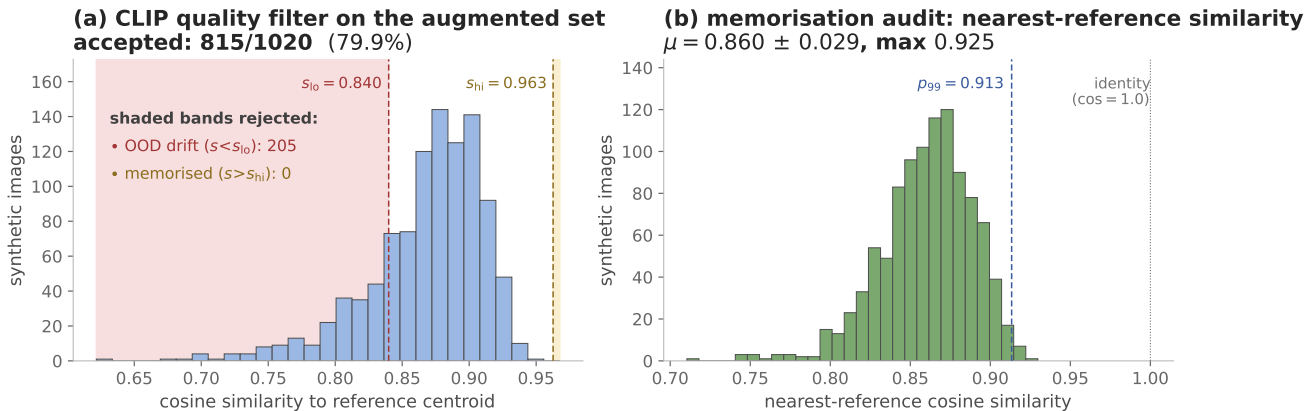


Fig. 13. Empirical audits over the V4 augmented set ($N_{\text{synth}}=1,020$, $N_{\text{ref}}=39$). **(a)** Histogram of each synthetic image’s cosine similarity to the reference centroid. The shaded bands are the leave-one-out admissibility band of Section III-I: pink rejects out-of-distribution drift, yellow rejects memorisation-signature outputs. Both thresholds are deterministic functions of the reference set with no hand-picked values. **(b)** Histogram of nearest-reference cosine similarity for every synthetic image; the 99th percentile sits well below the identity-level cosine of 1.0, so even the most reference-like samples are far from identity at the small-reference-set scale.

Synthetic outputs overlap the real texture distribution. Figure 14 embeds real and synthetic images in the 26-dimensional handcrafted texture/colour space (the descriptor underlying our curation filter) via PCA followed by t-distributed stochastic neighbour embedding (t-SNE). On this 2-D t-SNE projection the synthetic samples interleave with the real points rather than occupying a disjoint region; we note, however, that t-SNE geometry is only indicative and the quantitative recall and coverage (Table VI) qualify this. The overlap is consistent with the synthetic samples sharing much of the real texture/colour distribution; a 2-D t-SNE overlap is suggestive rather than conclusive and does not by itself establish statistical equivalence. We further note that the selected V4 preset trades manifold fidelity for diversity: its precision and density are the lowest among the presets (Table VI). The minority of synthetics that sit farthest from their nearest real neighbours form a *separable* cluster; these are precisely the texture outliers the curation step removes, which is why curation sharpens the pool without discarding faithful samples.

J. Computational cost

Table VII gives the wall-clock cost of the major phases of the generation pipeline on a single NVIDIA H100 NVL.

VI. DISCUSSION

The results in Section V show that the DreamBooth-fine-tuned SDXL generator, combined with multi-ControlNet conditioning and soft-mask inpainting, produces synthetic sand boil imagery with controlled stratification, defect-region preservation, and uniform mask provenance. We restrict the claims of this paper to the generation pipeline and its image-level quality measures. Whether training on *real plus synthetic* versus *real only* helps a downstream segmenter is left to future work. This section explains the main design choices and reviews the failure modes that remain.

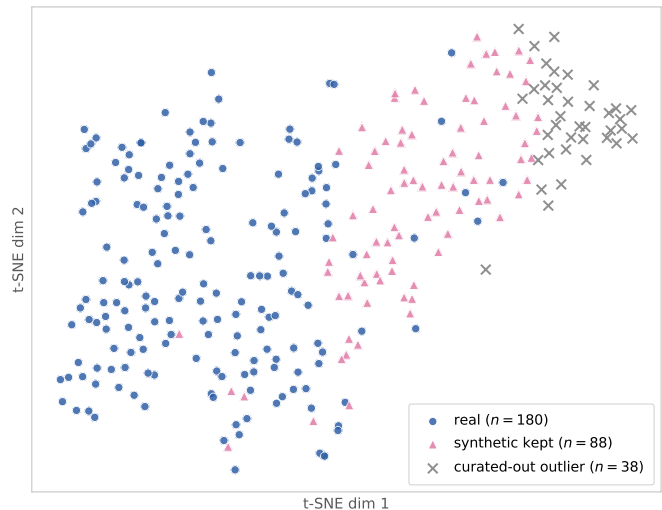


Fig. 14. Texture-feature faithfulness of the synthetic pool. Real and synthetic images are embedded with PCA followed by t-SNE on the 26-dimensional handcrafted texture/colour descriptor. Synthetic samples overlap the real distribution, while the minority farthest from their nearest real neighbours form a separable cluster (the texture outliers the curation filter removes).

A. What each component contributes

The generator-pipeline ablation of Section V-A makes a clear case for keeping all three ControlNets. Canny constrains the broad outline of the sand boil and the major scene boundaries. Depth constrains the three-dimensional layout. This matters because a sand boil is a small bulge above an otherwise flat surface, and without depth conditioning the generator occasionally flattens it. The soft-edge branch captures the fine textural transition at the boil’s rim, which is exactly where Canny and depth are weakest. Removing any single control visibly degrades boundaries on at least one of the five test sources.

The soft-mask inpainting path (row E) is preferred

TABLE VII
WALL-CLOCK COST ON A SINGLE NVIDIA H100 NVL GPU,
MEASURED FROM SLURM JOB-ACCOUNTING RECORDS OF THE
PRODUCTION RUN.

Phase	H100
Prompt Atlas build (CPU)	20 s
Augmented dataset generation (1,020 images @ 1024 × 1024)	2 h 13 min
Per-image generation throughput	~7.8 s/img
Synthetic-image quality (FID/KID/LPIPS, 300 samples)	<2 min
Per-preset evaluation generation (182 images each)	~24 min

to plain image-to-image (row D) for augmented-dataset generation, but the comparison is asymmetric. For visual variety alone, row D is adequate. For dataset generation, row E is strongly preferred because its synthetic ground-truth mask aligns pixel-for-pixel with the real mask, up to the small rim drift that the CONVEX HULL ANNOTATOR (CHA) step is designed to handle. Under row D the dome shape can drift slightly during denoising, and the auto-produced ground-truth mask must be relied upon entirely. Under row E the mask comes from a single pass over a known-anchored dome region, giving more reliable boundary labels in our qualitative inspection.

a) Why V4 is selected despite lower-FID alternatives.: Table IV shows V1 with the lowest FID among the image-to-image presets (169.1) and the mask-conditioned ControlNet (MaskCN) strongest on CLIP image-text similarity and within-set diversity, tied with V1 for the lowest KID, while V4 carries a numerically higher FID (201.8). We nonetheless select V4 for the augmented dataset, because the metric leaders each fail a requirement that FID does not see. MaskCN’s labels cannot be certified: its CHA quality gate reports a median drift IoU of only 0.15 (Section V-C), so the mask it is conditioned on is not a trustworthy ground-truth label for the dome it actually renders. V1’s labels are sounder but still re-predicted, and three further non-FID reasons favour V4. First, V1 generates whole-image img2img outputs in which the dome can drift slightly during denoising, so the synthetic ground-truth mask must be re-predicted by CHA over a dome that has shifted relative to the source. Under V4’s soft-mask inpainting path the dome pixels are preserved verbatim, so the mask-production policy is anchored to a known-stable dome region and yields cleaner boundary labels in qualitative inspection. Second, the precision/recall decomposition (Table VI) shows that V1 leads V4 on manifold fidelity (precision 0.656 vs 0.315) and on recall (0.497 vs 0.226). V4’s advantage over V1 is therefore not manifold coverage but its single-policy soft-mask label provenance, together with a modestly higher LPIPS within-set diversity (0.633 versus V1’s 0.595; a small-sample estimate we treat as suggestive rather than decisive), at equal CLIP image-text similarity (0.302). V4 trades manifold fidelity for a trustworthy label and a wider spread of synthetic scenes—the properties that matter when the set is used as seg-

menter supervision rather than as a sample from the real distribution. Third, distribution-matching scores reward *being close to the reference distribution*, which a degenerate route can achieve. The Poisson seamless-cloning baseline of Section V-E pastes real foregrounds onto real backgrounds and shows exactly this: FID $\approx$ 135, high precision and density, and high recall and coverage by construction, yet no genuinely new content. The selection criterion is therefore not minimum FID but the joint requirement of high within-set diversity (LPIPS spread), adequate fidelity (precision in the generative-preset range), high CLIP image-text similarity, and single-policy mask provenance.

We release V4 as the single *label-reliable* production preset: it is the only one whose synthetic dome is a preserved real structure rather than a freely re-rendered one. The presets are nonetheless complementary: V1 attains the broadest real-manifold coverage (Table VI), while V3 and the mask-conditioned ControlNet attain the highest within-set diversity. A curated *mixture* of presets, admitted by the CLIP quality filter and then sub-sampled for feature-space spread, is therefore a natural way to widen augmentation diversity without sacrificing on-distribution quality; quantifying its effect on downstream segmentation accuracy is left to future work.

B. Mask-conditioned generation as a diversity and zero-annotation engine

The mask-conditioned ControlNet (MaskCN) earns a first-class place in the pipeline for two reasons that are distinct from (and unaffected by) the unresolved label-verification gate of Section V-C. We state both carefully, because MaskCN’s downstream segmentation benefit is not asserted in this paper.

First, MaskCN is a *zero-annotation* labelling engine. By generating a fresh boil *into* a chosen binary mask, it makes the conditioning mask exactly the segmentation label of the image it produces (Figure 7), with no manual annotation and no dependence on a segmenter checkpoint. This removes the re-prediction ceiling that bounds the soft-mask+CHA labels: the soft-mask path’s label quality is capped by how well SandBoilNet transfers to synthetic renders, whereas MaskCN’s label is correct by construction *if* the generator fills the conditioning silhouette. What remains open is the *verification* of that adherence at scale: the currently available, real-trained gate reports a median drift IoU of 0.15 on these renders, which we read as largely a transfer artefact of the gate rather than evidence of a wrong label, though with the present gate we cannot fully separate gate transfer failure from generator non-adherence (Section V-C).

Second, MaskCN is a label-provenance diversity engine that the source-anchored img2img path structurally cannot match. Because its label geometry is decoupled from the catalogue of real source domes, MaskCN draws masks from a bank and jitters each instance with flips, rotations, rescaling, and repositioning (Figure 11, block B). The boil shape, size, and placement therefore vary independently of

any one real source dome. MaskCN thus widens a different axis of diversity from the img2img presets: where img2img varies the *scene* around a fixed real defect geometry, MaskCN varies the *label geometry itself* at zero annotation cost. A natural next step is to fold curated MaskCN singles into the augmentation mixture as a zero-annotation label source. Whether their by-construction labels hold up under an independent adherence check and improve downstream segmentation IoU is left to future work.

C. Backbone choice as a load-bearing decision

The choice of SDXL over a more recent rectified-flow alternative is, by design, a load-bearing decision rather than a default. Section III-C details the trade-off. In summary, SDXL pairs with the most mature publicly available ControlNet and IP-Adapter checkpoints, fits comfortably on a single H100 NVL at 1024×1024 with up to three simultaneous controls (or two controls plus an IP-Adapter), and, in our own qualitative experiments, was less prone than higher-capacity rectified-flow alternatives to memorising individual training frames at the reference-set scale used here. The same pipeline could in principle be re-instantiated on a rectified-flow backbone once equivalent structural-conditioning weights mature, with no change to the prompt curator.

D. Why prompt curation matters for niche datasets

When real data is scarce, the value of each synthetic sample depends disproportionately on its prompt. A synthetic image with an off-topic prompt, such as a sand boil rendered in a cartoonish style or in an unrelated geographic setting, drags the augmented distribution away from the real one. Under scarcity, this contributes far more to model degradation than the same defect would in a million-image dataset. Hand-written prompt lists offer no guard against this failure mode and tie the augmentation pipeline to the domain expert who wrote them. In response, the Prompt Atlas of Section III-G replaces the hand-written list with a stratified, validated, image-text-anchored bank that is reproducible from a single domain specification and portable across defect classes. The CLIP-based image-text validation stage does real work here: prompts whose CLIP similarity to the reference images falls below the cohort minimum are dropped, and the resulting bank provides explicit, reproducible coverage of the taxonomy axes that a hand-written list of comparable size does not guarantee.

E. Generalisation to other defect classes

Although the experiments focus on sand boils, the pipeline is domain-agnostic by construction. Extending to a new defect (rutting, seepage, sinkhole, and the like) requires only three changes: writing a new domain specification for the Prompt Atlas with the appropriate concept, taxonomy axes, and exemplar images; fine-tuning a new DreamBooth adapter on the new reference set;

and supplying the structural-conditioning sources for the multi-ControlNet stack. The Prompt Atlas builder accepts any domain specification without code changes, and the DreamBooth-LoRA recipe reuses the same training script.

One practical caveat is that the taxonomy axes that matter for one defect class do not transfer to another. A rutting domain specification needs road-surface, wheel-path-pattern, and depth-severity axes, and would not benefit from a hydraulic-regime axis. Each new domain therefore requires a brief modelling exercise to choose its axes. This effort is small relative to the cost of curating real data or hand-writing prompts for every scenario.

F. Failure modes and limitations

Several limitations remain, which we set out below.

Style of the reference set. The DreamBooth adapter is fit to the curated USACE reference set. The visual style of the synthetic images is consequently anchored to that of USACE photography: camera type, colour profile, and ground-level perspective. A deployed application that receives drone (UAV) or smartphone imagery from a different source may not generalise as well as it does to USACE-style imagery. A straightforward remedy is to augment the adapter’s reference set with images from additional sources.

Reflective surfaces. The downstream segmenter still struggles with images in which the sand boil is partially obscured by reflective standing water. The reflective patch breaks the texture cues that the network relies on. We leave to future work a separate reflection-masked training condition, or a two-pass approach that first predicts reflection regions and then segments within the non-reflective area.

Downstream evaluation scope. This paper reports generative quality only (FID, KID, CLIP image-text similarity, and LPIPS diversity) together with the empirical behaviour of the CLIP admissibility and memorisation audits. A fuller multi-architecture downstream-segmentation evaluation, with cross-validation and a leak-free fold filter, is left to future work, which would consume the augmented dataset produced here.

Cross-domain validation. The pipeline is constructed to be class-agnostic: a single JSON specification ports the Prompt Atlas, and CHA’s shape-aware geometry adapter routes by form factor between compact, elongated, and thin curvilinear cleanup. The empirical validation in this paper, however, is restricted to the sand boil class. The class-agnostic claim is exercised by self-tests on synthetic shape inputs in the released CHA implementation, and the released code ships domain-specification files for sand-boil, sinkhole, rutting, and crack. An end-to-end empirical demonstration on a non-sand-boil class is left to a follow-up release once the corresponding reference sets and DreamBooth adapters are curated.

Prompt Atlas template gap. The v1 prompt template references seven of the eight taxonomy axes; the **season** axis is sampled into the per-prompt fill record

but is not rendered into the prompt text itself. The downstream effect is captured by the 40/44 rendered-coverage figure in Section III-G. Adding the missing template slot would be a one-line specification edit, but we did not re-run the production atlas for this paper, in order to preserve a one-to-one comparison with the previously generated augmented set.

Synthetic-mask quality ceiling. The synthetic ground-truth mask is produced by CHA from the publicly released SandBoilNet checkpoint. The quality of the resulting label is therefore upper-bounded by the prediction quality of that checkpoint on the synthetic image. For most samples the convex-hull post-processing absorbs minor prediction noise. On edge cases where the checkpoint disagrees with the rendered dome by a large margin, however, the synthetic label is unreliable and the sample must be filtered out. The deployed pipeline exposes a human-in-the-loop correction interface that allows an inspector to overwrite the synthetic label per image; the operational use of that interface is left to future work.

CLIP circularity in prompt validation. The Prompt Atlas filters candidate prompts by their CLIP image-text similarity to the reference set, and the V3/V4 presets are in turn conditioned with a CLIP-aligned style. Validating prompts with the same family of encoder that guides generation is mildly circular and could bias the atlas toward CLIP-favoured phrasings. The effect is limited here because field inspection photography is visually narrow and well-lit, but in a more heterogeneous domain the validation stage should use an ensemble of image-text encoders distinct from the generator’s conditioning encoder.

Un-ablated generation constants. The soft-mask erosion and blur coefficients, the per-branch ControlNet conditioning scales, and the classifier-free guidance value were selected on the five-source generator ablation rather than swept independently, and the SDXL backbone choice rests on a qualitative memorisation observation (Section III-C) rather than a controlled backbone benchmark. We report them as fixed, reproducible settings; a full sensitivity analysis and a rectified-flow backbone comparison are left to future work and do not affect the conclusions drawn from the fixed configuration evaluated here.

G. Threats to validity

Two threats to the conclusions drawn here deserve flagging. First, the generative-quality measures (FID, KID, and LPIPS) computed against a reference set of the size used here have known sample-size bias. We report KID with its standard deviation alongside FID to mitigate the small-sample bias of FID, but a larger reference set would be needed to make fine-grained between-preset comparisons fully reliable. Second, the augmented dataset is generated by a single DreamBooth adapter trained on a single curated reference set. A residual concern is that any generation-side stylistic bias of that adapter could narrow the diversity of the augmented set rather than

reflect a general property of diffusion-based augmentation. To address this directly we trained an ensemble of three DreamBooth-LoRA adapters on disjoint 13-image partitions of the reference set and evaluated each on a common prompt/seed set. The three adapters agree tightly: FID 235.3 ± 1.9 (range 4.5), KID 0.053 ± 0.005 , and LPIPS diversity 0.588 ± 0.007 across the three independent partitions. Because adapters fit to disjoint image subsets produce near-identical quality and diversity, the augmentation behaviour appears to be driven by the recipe rather than by one particular adapter’s stylistic bias. This rules out single-adapter bias, though not a dataset- or source-level bias within this single-source reference set. The ensemble-training and evaluation scripts are released so the control can be reproduced and extended to disjoint reference sets from other camera sources.

H. Ethical considerations

The intended use is to support, not to replace, trained inspectors. Sand boils are safety-critical defects: a false negative during a high-water event can delay intervention long enough to permit a breach. Any operational deployment that consumes a segmenter trained on this augmented data should expose model uncertainty rather than collapse it to a binary decision, and should retain a human-in-the-loop correction path for high-stakes inspections. We also note that the synthetic images produced by this pipeline are stylistic inheritors of the USACE reference photography. They are not photorealistic in a forensic sense and should not be presented as field documentation of actual sand boils.

VII. CONCLUSION AND FUTURE WORK

We presented a generation pipeline for synthetic sand boil imagery. The pipeline couples a DreamBooth-fine-tuned Stable Diffusion XL backbone, four independent ControlNet branches (Canny, monocular depth, surface normal, and HED soft-edge) from which each preset activates a subset, a soft-mask inpainting protocol that preserves the real defect region pixel-for-pixel, and a taxonomy-driven Prompt Atlas. The Prompt Atlas turns a single JSON domain specification into a stratified, de-duplicated, image-text-validated prompt bank in under a minute. We chose SDXL over more recent rectified-flow alternatives deliberately: at the reference-set scale and under the multi-ControlNet inpainting regime this application imposes, SDXL offered the best combination of structural-conditioning maturity, VRAM headroom, and overfit resistance.

We evaluated the pipeline on standard generative-image-quality measures (FID under two independent implementations, KID, CLIP image-text similarity, and LPIPS within-set diversity), together with a precision/recall and density/coverage decomposition that separates fidelity from diversity. We reported the production preset selected for the augmented dataset alongside three image-to-image variants, a mask-conditioned ControlNet,

and a Poisson seamless-cloning baseline (the compositing strategy used in our earlier pipeline). All were evaluated against the full real reference set, with KID reported alongside its standard deviation. We find that no single preset dominates: the configurations are complementary, trading manifold coverage, within-set diversity, and label reliability. We therefore release the label-reliable preset as the default and identify a curated mixture across them as the natural augmentation set. We demonstrated the leave-one-out CLIP admissibility filter empirically on the augmented set, and used a CLIP-space nearest-reference audit to estimate the memorisation rate at the small-reference-set scale. Downstream use of the augmented dataset for segmentation is left to future work.

The pipeline is designed to transfer to a new defect class through a single domain specification. The released code emits a complete artefact manifest (DreamBooth checkpoint, Prompt Atlas record, multi-ControlNet inference configuration, and per-image seed cascade), making the reported results reproducible.

Several extensions remain open.

Closed-loop diffusion as a segmentation teacher. The present augmentation is open-loop: the synthetic dataset is generated once, and then training proceeds. A closed-loop variant would use a downstream segmenter’s prediction uncertainty to steer subsequent rounds of synthesis toward the specific image conditions on which the current model fails. This is structurally close to active learning, with diffusion-based augmentation taking the role usually played by oracle labelling.

Scaling the mask-conditioned path. The mask-conditioned ControlNet of Section III-F is released alongside the soft-mask pipeline. It generates from a chosen mask so the label is exact by construction, with CHA as its drift verifier. It already serves as a zero-annotation diversity engine: single boils are drawn from a curated mask bank under per-instance jitter. These contribute exact labels and a variety of boil shapes, sizes, and placements that the source-anchored path cannot produce. Two extensions remain open. First, certifying its by-construction labels at scale needs a gate that transfers to synthetic renders, since the present CHA gate (built on the real-trained SandBoilNet checkpoint) under- and over-segments them (Section V-C). Second, the ControlNet could be scaled from the 199-pair fine-tune to a larger multi-source corpus, and a generative prior could be learned over plausible defect masks so that the curated mask bank is extended by a learned mask distribution. This would let the pipeline synthesise paired (image, mask) records end-to-end with no real source image at all. Whether MaskCN samples improve downstream segmentation accuracy is a question for future work, not this paper.

Larger curated benchmark. A publicly available levee-defect benchmark with paired real and synthetic data, source tags, and a held-out test set would let the community compare methods directly. Building such a benchmark on top of the dataset and reproducibility framework released here, together with a deployed

inspection-workflow prototype, is left to future work.

Domain transfer to other defect classes. The Prompt Atlas builder ports to a new domain through its JSON specification alone; re-instantiating the full multi-ControlNet inpainting pipeline on a seepage, sinkhole, or rutting domain additionally requires a new DreamBooth adapter and structural-conditioning sources. Such a transfer would provide external validation of the generation pipeline that the present paper’s single-class focus does not.

CODE AND DATA AVAILABILITY

The complete generation pipeline (DreamBooth-LoRA training script, multi-ControlNet inference configuration, soft-mask inpainting implementation, CONVEX HULL ANNOTATOR reference implementation, Prompt Atlas builder, and the per-image seed cascade) is released alongside this paper. The released artefact includes the trained LoRA safetensors checkpoint (~240 MB), the Prompt Atlas JSON Lines record for the sand boil domain (v1), and the parameter manifest needed to reproduce every numerical entry in Section V from a single command. The CONVEX HULL ANNOTATOR pipeline is additionally distributed as a standalone, class-agnostic library [42] with its own domain specifications for sand boil, sinkhole, rutting, and crack defect classes, so the post-processing component is reusable outside the present pipeline.

On acceptance, the full artefact (training and inference code, the trained LoRA checkpoint, the Prompt Atlas record, and the reproducibility manifest) will be published in a public Git repository and archived under a versioned Zenodo DOI for long-term citation.

The curated reference set is derived from the U.S. Army Corps of Engineers (USACE) levee-inspection archive [40]. Its redistribution is subject to the USACE archive’s terms of use, so the underlying photographs are not redistributed; the pixel-level binary masks produced by the authors will be released under the Creative Commons Attribution 4.0 (CC BY 4.0) license alongside the code.

CONFLICTS OF INTEREST

The authors declare no competing financial or non-financial interests. This work was carried out in an academic research setting and is derived from the corresponding author’s Master’s thesis [12].

ACKNOWLEDGMENTS

The authors thank the U.S. Army Corps of Engineers for making the levee-inspection imagery archive [40] available for academic research, and the maintainers of the open-source diffusion-model ecosystem (Stability AI, the diffusers library, the ControlNet authors, and the IP-Adapter authors) whose publicly released checkpoints made this pipeline tractable on a single accelerator. This work was conducted on an institutional GPU cluster at Louisiana State University New Orleans.

REFERENCES

- [1] R. Seed, R. Bea, A. Athanasopoulos-Zekkos *et al.*, “Investigation of the performance of the new orleans flood protection systems in hurricane katrina on august 29, 2005,” Independent Levee Investigation Team, University of California Berkeley, Tech. Rep., 2006.
- [2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [3] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, “Effective data augmentation with diffusion models,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [4] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “SDXL: Improving latent diffusion models for high-resolution image synthesis,” *arXiv preprint arXiv:2307.01952*, 2023.
- [5] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [7] X. Liu, C. Gong, and Q. Liu, “Flow straight and fast: Learning to generate and transfer data with rectified flow,” in *International Conference on Learning Representations (ICLR)*, 2023.
- [8] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach, “Scaling rectified flow transformers for high-resolution image synthesis,” in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- [9] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [10] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [11] S. Xie and Z. Tu, “Holistically-nested edge detection,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [12] P. J. Thapa, “Toward robust semantic segmentation in levee infrastructure monitoring: Enhancing accuracy with high-fidelity synthetic data and ensemble learning,” Master’s Thesis, University of New Orleans, New Orleans, LA, USA, 2025, university of New Orleans Theses and Dissertations No. 3231. <https://scholarworks.uno.edu/td/3231/>.
- [13] P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing,” *ACM Transactions on Graphics (SIGGRAPH)*, vol. 22, no. 3, pp. 313–318, 2003.
- [14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning (ICML)*, 2021.
- [15] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, “Improved precision and recall metric for assessing generative models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [16] S. Guan, H. Liu, H. R. Pourreza, and H. Mahyar, “Deep learning approaches in pavement distress identification: A review,” *arXiv preprint arXiv:2308.00828*, 2023.
- [17] M. Panta, M. T. Hoque, K. N. Niles, J. Tom, M. Abdelguerfi, and M. Flanagan, “Deep learning approach for accurate segmentation of sand boils in levee systems,” *IEEE Access*, vol. 11, pp. 126 263–126 282, 2023.
- [18] A. Kuchi, M. T. Hoque, M. Abdelguerfi, and M. C. Flanagan, “Machine learning applications in detecting sand boils from images,” *Array*, vol. 3, p. 100012, 2019.
- [19] R. Alshawi, M. M. Ferdous, M. Abdelguerfi, K. Niles, K. Pathak, and S. Sloan, “Imbalance-aware culvert-sewer defect segmentation using an enhanced feature pyramid network,” *arXiv preprint arXiv:2408.10181*, 2024.
- [20] R. Alshawi, M. T. Hoque, and M. C. Flanagan, “A depth-wise separable U-Net architecture with multiscale filters to detect sinkholes,” *Remote Sensing*, vol. 15, no. 5, p. 1384, 2023.
- [21] M. Panta, P. J. Thapa, M. T. Hoque, K. N. Niles, S. Sloan, M. Flanagan, K. Pathak, and M. Abdelguerfi, “Application of deep learning for segmenting seepages in levee systems,” *Remote Sensing*, vol. 16, no. 13, p. 2441, 2024.
- [22] J. V. Aanstoos, K. Hasan, C. G. O’Hara, S. Prasad, L. Dabiru, M. Mahrooghy, B. Gokaraju, and R. Nobrega, “Earthen levee monitoring with synthetic aperture radar,” in *2011 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2011.
- [23] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *International Conference on Learning Representations (ICLR)*, 2014.
- [24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
- [25] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [26] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training generative adversarial networks with limited data,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [27] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [28] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *International Conference on Learning Representations (ICLR)*, 2021.
- [29] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, “IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models,” *arXiv preprint arXiv:2308.06721*, 2023.
- [30] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet, “Synthetic data from diffusion models improves ImageNet classification,” *Transactions on Machine Learning Research*, 2023.
- [31] Y. Liu, J. Yue, S. Xia, P. Ghamisi, W. Xie, and L. Fang, “Diffusion models meet remote sensing: Principles, methods, and perspectives,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–22, 2024, arXiv:2404.08926.
- [32] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hachililoglu, and D. Merhof, “Diffusion models in medical imaging: A comprehensive survey,” *Medical Image Analysis*, vol. 88, p. 102846, 2023.
- [33] Y. Duan, Y. Hong, L. Niu, and L. Zhang, “Few-shot defect image generation via defect-aware feature manipulation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [34] T. Hu, J. Zhang, R. Yi, Y. Du, X. Chen, L. Liu, Y. Wang, and C. Wang, “AnomalyDiffusion: Few-shot anomaly image generation with diffusion model,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [35] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, “RePaint: Inpainting using denoising diffusion probabilistic models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [36] O. Avrahami, D. Lischinski, and O. Fried, “Blended diffusion for text-driven editing of natural images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [37] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [38] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler, “DatasetGAN: Efficient labeled data factory with minimal human effort,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [39] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, “CLIPScore: A reference-free evaluation metric for image captioning,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

- [40] U. A. C. of Engineers, “Levee owner’s manual for non-federal flood control works,” <https://www.mvr.usace.army.mil/Portals/48/docs/EC/LSP/LeveeOwnersManual.pdf>, accessed 2025-02-17.
- [41] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [42] P. J. Thapa, “Convex_Hull_Annotator: Automated convex-hull mask annotation from a SandBoilNet probability map,” https://github.com/padam56/Convex_Hull_Annotator, 2024, companion tool to [12]; applies thresholding, connected-component analysis, and per-component convex hulls to a CNN-predicted sandboil probability map.
- [43] A. Dutta and A. Zisserman, “The VIA annotation software for images, audio and video,” in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM ’19. ACM, 2019, pp. 2276–2279.
- [44] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and flexible image augmentations,” *Information*, vol. 11, no. 2, p. 125, 2020.
- [45] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer, “8-bit optimizers via block-wise quantization,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [46] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.
- [47] J. Wang, K. C. Chan, and C. C. Loy, “Exploring CLIP for assessing the look and feel of images,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [48] A. Yang *et al.*, “Qwen2.5 technical report,” *arXiv preprint arXiv:2412.15115*, 2024.
- [49] O. Boer Bohan, “SDXL-VAE-FP16-Fix: a numerically stable fp16 variational autoencoder for SDXL,” <https://huggingface.co/madebyollin/sd-xl-vae-fp16-fix>, 2023, hugging Face model repository.
- [50] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, “DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [51] —, “DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models,” *arXiv preprint arXiv:2211.01095*, 2022.
- [52] M. J. Chong and D. Forsyth, “Effectively unbiased FID and inception score and where to find them,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [54] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [55] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [56] G. Parmar, R. Zhang, and J.-Y. Zhu, “On aliased resizing and surprising subtleties in GAN evaluation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [58] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, “Reliable fidelity and diversity metrics for generative models,” in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.