CoSTA-Diff: Medical Concept-Conditioned Diffusion for Targeted Skin Lesion Augmentation

Anonymous authors
Paper under double-blind review

Abstract

Diffusion-based augmentation can expand medical training sets, but class labels alone do not specify which within-class patterns should be generated. We present CoSTA-Diff, a latent diffusion framework that uses shared medical concept coordinates to connect generation conditions with augmentation targets. A frozen MONET model provides image-specific evidence for 22 dermoscopic concepts and fixed text directions. An additive conditioner combines these quantities with the diagnosis to construct class and concept tokens for cross-attention, without requiring manual per-image concept annotations. The same evidence coordinates define within-class neighborhoods, allowing sparsely represented training profiles to serve as synthesis anchors. New images are generated from fresh noise under the anchors' observed labels and concept evidence, supported by a compact wavelet latent autoencoder. In a retrospective evaluation on a retained ISIC 2018 split, adding 1,800 synthetic images yields mean ResNet-50 balanced accuracy of 71.48% and macro-F1 of 73.73% across three classifier seeds, exceeding real-only training by 2.70 and 2.33 percentage points and a class-conditioned latent diffusion baseline by 3.04 and 3.65 points. Design studies and cross-backbone comparisons characterize the conditioning choices and architecture dependence of these gains. CoSTA-Diff provides a shared representation for concept-conditioned synthesis and targeted data expansion.

1 Introduction

Diffusion augmentation can enlarge a medical training set without resolving uneven coverage within each diagnosis. Lesions from the same class may exhibit different dermoscopic patterns, yet a class-balanced synthesis budget does not specify which patterns to revisit. Effective augmentation needs both a representation of what to generate and a rule for allocating generation to observed training profiles.

We ask: *can the same medical concept coordinates connect generation conditions with augmentation targets?* We present **CoSTA-Diff**, which uses frozen MONET to obtain evidence for 22 dermoscopic concepts and their fixed text directions. An additive conditioner converts this information and the diagnosis into cross-attention tokens. Within-class neighborhoods in the same evidence coordinates identify sparsely represented profiles for fresh-noise synthesis under their observed conditions. A wavelet latent autoencoder (WSL-AE) supports synthesis through asymmetric subband pathways and detail reconstruction supervision.

Across three classifier seeds, CoSTA-Diff improves ResNet-50 balanced accuracy and macro-F1 over real-only training by 2.70 and 2.33 points. Our contributions are:

1. **An explicit class-and-concept diffusion interface.** We combine fixed medical text directions, image-specific evidence, and concept identities into 22 additive concept tokens alongside a diagnosis token, without manual per-image concept labels or a direct global-image branch.
2. **Generation and allocation in a shared representation.** Same-class neighborhood support selects observed concept profiles as synthesis anchors; their evidence supplies both the generation conditions and the basis for allocating a class-level budget.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

3. **Evaluation of utility and design dependence.** Matched-budget comparisons quantify downstream classification gains, while conditioning and latent design studies examine the implementation choices. Cross-backbone, anchor-selection, and budget studies characterize how augmentation utility varies across evaluation settings.

2 Related Work

Medical and concept-conditioned synthesis. LesionGen (Fayyad et al., 2026) uses concept-rich captions; DiDGen (Shentu et al., 2026) combines semantic and spatial information; and Kim et al. (2025) combine mask-focused features with lesion-to-normal interpolation. Concept bottlenecks (Koh et al., 2020; Kulkarni et al., 2025) motivate named intermediate attributes. Salient concept-aware augmentation (Zhao et al., 2025) uses a discriminative embedding, whereas CPDE (Sun et al., 2026) uses MONET in a masked editing loss. CoSTA-Diff instead supplies named concept evidence directly through cross-attention tokens, without masks or generated captions.

Targeted data expansion. Target selection has used class–context associations (Yang et al., 2026), learning difficulty (Nguyen et al., 2026), metadata-defined subgroups (Drexlin et al., 2026), or feature prototypes (Zhu et al., 2024); Diff-II (Wang & Chen, 2025) combines concept learning with inversion and interpolation. CoSTA-Diff measures neighborhood support in the same concept coordinates used for conditioning (Table 11).

Latent representations. Latent diffusion (Rombach et al., 2022) and frequency-aware autoencoders (Lin et al., 2023; Medi et al., 2026) motivate detail-preserving spatial representations. Our WSL-AE supports the concept interface through asymmetric subband pathways and explicit detail supervision.

3 Method

3.1 Overview

Given a dermoscopic training set $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$, CoSTA-Diff constructs a labeled synthetic set using concept conditioning, a wavelet latent representation, and concept-space anchor selection (Figure 1). MONET and WSL-AE are frozen before diffusion training; the resulting generator produces images offline for a separately trained classifier.

3.2 Medical Concept Conditioning

Our additive class-and-concept interface uses one token per dermoscopic concept and a separate diagnosis token, with no direct global-image token branch. Frozen MONET (Kim et al., 2024) supplies unit-normalized image/text embeddings for a fixed vocabulary of 22 concepts; no per-image manual concept annotation is used.

Concept evidence and text directions. For concept m, each prompt pair $p \in \mathcal{P}_m$ contains a concept description and a generic lesion reference. Let g_i denote the unit-normalized MONET image embedding of x_i, and let h_{mp}^{+} and h_{mp}^{ref} be their unit-normalized text embeddings. We compute

$$c_{i,m} = \frac{\lambda}{|\mathcal{P}_m|} \sum_{p \in \mathcal{P}_m} g_i^{\mathsf{T}} (h_{mp}^{+} - h_{mp}^{\mathrm{ref}}),$$

$$\bar{d}_m = \frac{1}{|\mathcal{P}_m|} \sum_{p \in \mathcal{P}_m} (h_{mp}^{+} - h_{mp}^{\mathrm{ref}}), \qquad d_m = \frac{\bar{d}_m}{\|\bar{d}_m\|_2}. \tag{1}$$

Here λ is the fixed MONET similarity scale, $c_i \in \mathbb{R}^{22}$ is continuous image–concept evidence, and $d_m \in \mathbb{R}^{768}$ is a fixed text direction. The two quantities play complementary roles: d_m supplies a shared semantic direction for concept m, whereas $c_{i,m}$ records how strongly image x_i agrees with that direction. Keeping the direction, score, and concept identity in separate terms allows the conditioner to distinguish what a concept represents from how much evidence an image provides for it, while the class token retains the diagnostic context.

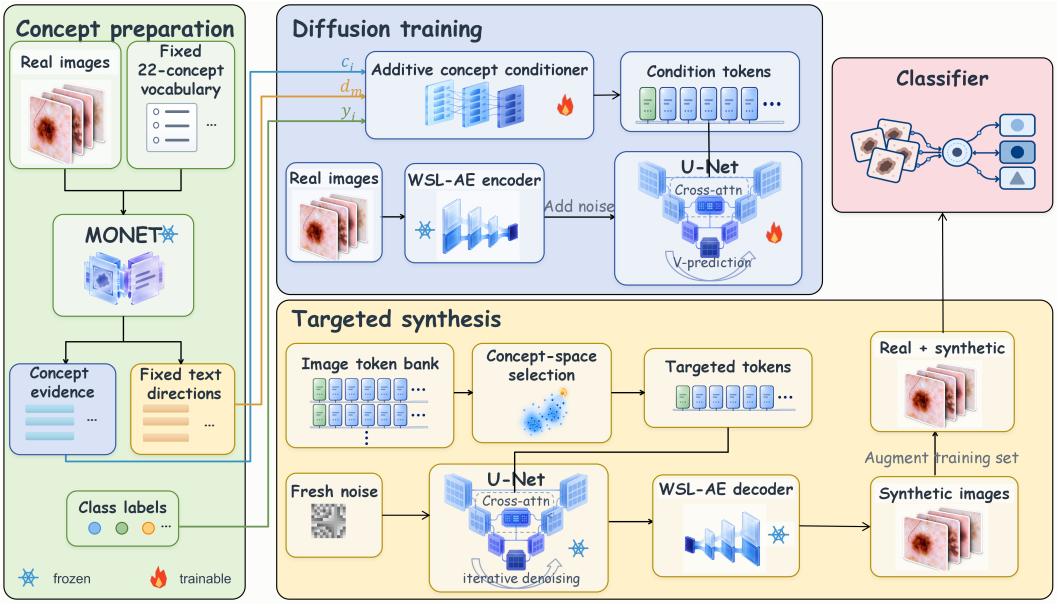

Figure 1: CoSTA-Diff: shared concept evidence supports conditioning and anchor selection; fresh-noise synthesis augments classifier training. MONET and WSL-AE are frozen during diffusion training. The depicted token bank is indexed by concept evidence; selection distances use evidence rather than token embeddings.

Additive class-and-concept tokens. The conditioner maps $(y_i, c_i; \mathbf{d})$ to $\mathcal{C}_i \in \mathbb{R}^{23\times 256}$, as detailed in Figure 2. We write $\bar{c}_{i,m}$ for evidence standardized using fixed training-cache statistics (Appendix A.12).

With learned class/concept type embeddings $e^{\text{cls}}, e^{\text{con}}$ and concept identity embeddings e^{id}_m, the tokens are

$$\begin{aligned} t_i^{\text{cls}} &= \text{LN}_{\text{cls}}(E_y(y_i) + e^{\text{cls}}), \\ t_{i,m}^{\text{con}} &= \text{LN}_{\text{con}}\left(W_d d_m + S(\bar{c}_{i,m}) + e^{\text{id}}_m + e^{\text{con}}\right), \\ \mathcal{C}_i &= [t_i^{\text{cls}}, t_{i,1}^{\text{con}}, \ldots, t_{i,22}^{\text{con}}] \in \mathbb{R}^{23\times 256}. \end{aligned} \tag{2}$$

Here S embeds scalar evidence and W_d projects the text direction; both maps are shared across concepts (Appendix A.10).

The sequence supplies U-Net cross-attention keys and values (Rombach et al., 2022); timestep conditioning remains separate.

Conditional diffusion in the normalized latent space. WSL-AE maps images to posterior-mode latents in $\mathbb{R}^{8\times 64\times 64}$, standardized by training-set channel statistics (Appendix A.12). With the forward process of Ho et al. (2020) and velocity target v_t (Salimans & Ho, 2022), the U-shaped denoiser (Ronneberger et al., 2015; Rombach et al., 2022) v_θ and conditioner ψ minimize

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{i,t,\epsilon}[\text{MSE}\left(v_\theta(\widetilde{z}_t, t, \mathcal{C}_\psi(y_i, c_i; \mathbf{d})), v_t\right)], \tag{3}$$

with uniform timestep sampling (training settings in Appendix A.1).

DDIM (Song et al., 2021) starts from fresh Gaussian noise; the final latent is denormalized and decoded. Anchors provide conditions only.

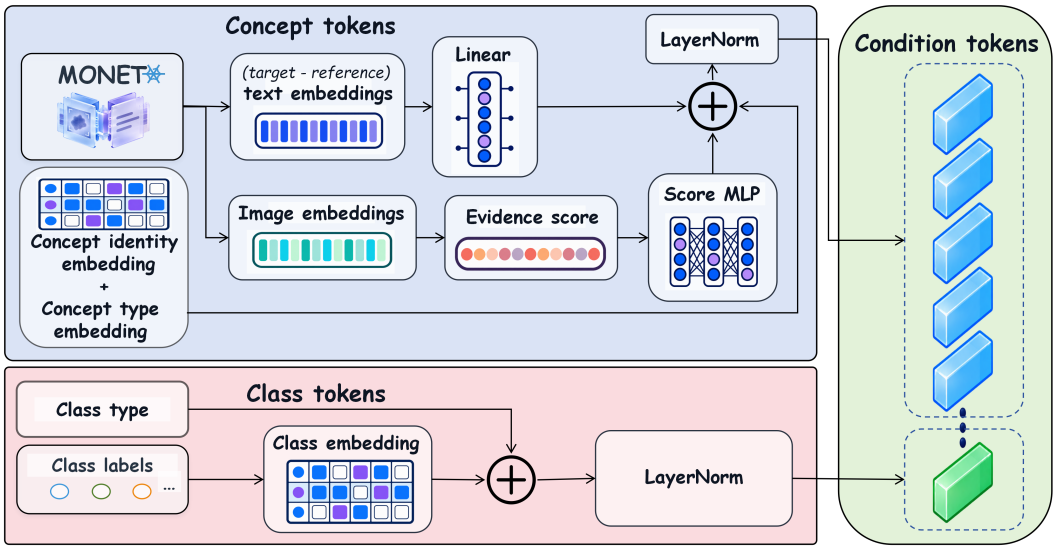

Figure 2: Token construction by addition and LayerNorm, yielding 22 concept tokens and one class token. Text-direction normalization, image–text evidence scoring, and score standardization are abbreviated (Eqs. 1, 12); the context has shape 23×256.

3.3 Wavelet Latent Autoencoder

A fixed orthonormal Haar transform $\mathcal{W}$ decomposes $x \in \mathbb{R}^{3\times256\times256}$ into approximation x_{LL} and detail bands $x_{LH}, x_{HL}, x_{HH} \in \mathbb{R}^{3\times128\times128}$. Two encoders process these bands (Figure 3):

$$\begin{aligned} f_L &= E_L(x_{LL}), & f_H &= E_H([x_{LH}, x_{HL}, x_{HH}]), \\ (\mu_L, \ell_L) &= P_L(F([f_L, f_H])), & (\mu_H, \ell_H) &= P_H(f_H). \end{aligned} \tag{4}$$

Here $[\cdot,\cdot]$ concatenates channels. Only the low posterior fuses both branch features. The diagonal-Gaussian posterior has mean $\mu = [\mu_L, \mu_H]$ and log variance $\ell = [\ell_L, \ell_H]$; its mode $z = \mu = [z_L, z_H]$ has six low-pathway and two detail-pathway channels. The low decoder receives the complete latent, while its features condition detail decoding:

$$\begin{aligned} u_L &= D_L^{\text{feat}}([z_L, z_H]), & \widehat{x}_{LL} &= O_L(u_L), \\ \widehat{x}_{\text{high}} &= D_H([U_H(z_H), u_L]), & \widehat{x} &= \mathcal{W}^{-1}(\widehat{x}_{LL}, \widehat{x}_{\text{high}}). \end{aligned} \tag{5}$$

Here U_H projects and upsamples z_H, and $\widehat{x}_{\text{high}}$ concatenates the three predicted detail bands. Inverse Haar reconstructs RGB without encoder-to-decoder skips. Training combines image MAE, mean detail-band MAE, and a diagonal-Gaussian KL penalty:

$$\mathcal{L}_{\text{AE}} = \mathbb{E}_{x\sim\mathcal{D}_{\text{tr}}}\left[\mathcal{L}_{\text{image}} + 0.1\mathcal{L}_{\text{high}} + 10^{-6}\mathcal{L}_{\text{KL}}(x)\right]. \tag{6}$$

The autoencoder is frozen before diffusion training and uses no adversarial loss, perceptual network, or lesion mask.

3.4 Concept-Space Targeted Synthesis

We target weakly supported within-class profiles. Let s_i be concept evidence standardized using selection-specific training statistics (Appendix A.12); sparsity denotes low neighborhood support, not a small number of nonzero coordinates.

For a class c with $N_c \geq 2$ training images, let $\mathcal{N}_{k_c}(i)$ be the $k_c = \min(5, N_c - 1)$ nearest same-class neighbors of image i, excluding itself. Define

$$a_i = \frac{1}{k_c}\sum_{j\in\mathcal{N}_{k_c}(i)} \|s_i - s_j\|_2, \qquad r_i = \text{Percentile}(\{a_j : y_j = y_i\}, a_i). \tag{7}$$

We use Euclidean distance. Percentiles lie in $[0, 1]$ with tied groups assigned midpoint ranks.

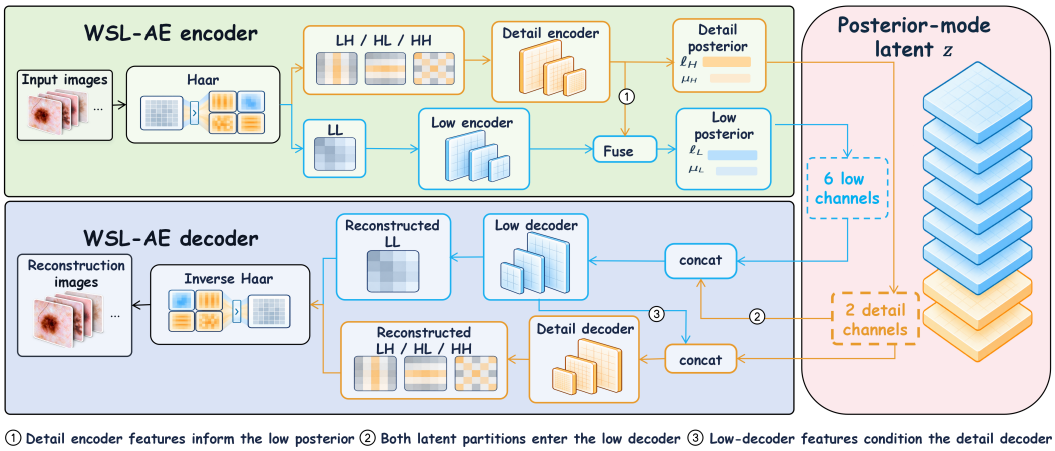

Figure 3: WSL-AE: detail features inform the low posterior, both latent partitions enter the low decoder, and low-decoder features condition detail decoding. The $6 + 2$ latent supports inverse-Haar reconstruction.

Anchor selection and allocation. For each class with budget $B_c > 0$, we select the highest-sparsity $\min(B_c, \lceil \rho N_c \rceil)$ profiles to form $\mathcal{A}_c$. We allocate n_i jobs per anchor as evenly as possible, using $\rho = 0.4$ and 300 images per non-NV class (details in Appendix A.12). The selection score sets sampling priority: a high a_i identifies a profile with weak local support among same-class evidence vectors. Once selected, the anchor retains its observed label and evidence, so allocation changes how often each conditional distribution is sampled. Fresh noise then produces new draws under that condition.

The induced augmentation distribution. Let $\mathcal{I}_c = \{i : y_i = c\}$, $\kappa_i = (y_i, c_i; \mathbf{d})$, and $p_G(\widehat{x} \mid \kappa_i)$ denote the image distribution induced by the frozen conditional generator and fresh Gaussian noise. For a class with positive budget, the augmentation distribution is

$$q_\pi(\widehat{x} \mid y = c) = \sum_{i \in \mathcal{I}_c} \pi_i \, p_G(\widehat{x} \mid \kappa_i), \qquad \pi_i \geq 0, \quad \sum_{i \in \mathcal{I}_c} \pi_i = 1. \tag{8}$$

Here $\pi_i = n_i / B_c$ for selected anchors and zero otherwise: conditioning defines the components, while targeting changes their mixture weights.

Each generated image inherits its anchor label. The classifier trains on $\mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{syn}}$; the sampling equation is given in Appendix A.12.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate seven-class skin-lesion classification on ISIC 2018 Task 3, using the HAM10000 training collection (Tschandl et al., 2018). Stratified image-level sampling with split seed 42 reserves 2,003 of the 10,015 training images for validation. Combined with the original 193 validation images, this yields 8,012/2,196 training/validation images; the original 1,512-image test set is unchanged. Each augmentation method adds 1,800 images to the training set, with 300 per class except the majority class, melanocytic nevus (NV). Class counts and the full protocol appear in Appendix A.1.

Metrics, baselines and implementation. We report balanced accuracy (BA, mean per-class recall), macro-F1, macro one-versus-rest AUROC, and accuracy in percent, with mean $\pm$ sample SD over classifier seeds 42, 1042, and 2026 unless stated otherwise. Baselines are Real-only, class-conditioned Vanilla LDM, DA-Fusion (Trabucco et al., 2024), DistDiff (Zhu et al., 2024), and Diff-II (Wang & Chen, 2025); they share the downstream split and synthetic budget while retaining method-specific

generative pretraining and adaptation. ResNet-50 uses ImageNet initialization, 256-pixel inputs, batch size 64, and 40 training epochs with SGD (learning rate 0.002, momentum 0.9, weight decay 10^{-4}), cosine decay to 10^{-6}, and label smoothing 0.05. Maximum validation BA over epochs 10–40 selects the checkpoint, with ties resolved by the earliest epoch. CoSTA-Diff trains for 100 diffusion epochs using AdamW (learning rate 2×10^{-4}, batch size 16) with frozen MONET and WSL-AE; validation diffusion loss selects the generator. DDIM sampling with 100 steps, $\eta = 0$, and no classifier guidance produces a fixed synthetic pool reused across classifier seeds. Experiments use an RTX 5060 Ti (16 GB). Appendix A.1 provides optimizer, schedule, and baseline-specific details.

4.2 Downstream Utility and Transfer

The primary test asks whether CoSTA-Diff helps under a shared downstream budget. Table 1 compares the synthesis pipelines through ResNet-50 classification, with Real-only as the reference.

Table 1: ResNet-50 comparison: mean ± sample SD for primary seeds 42, 1042, and 2026. Higher is better; bold marks the highest mean. The additional seed 2048 is reported in the supplement.

Method	BA	Macro-F1	AUROC	Accuracy
Real-only	68.78 ± 0.90	71.40 ± 0.86	96.07 ± 0.08	82.87 ± 0.07
Vanilla LDM	68.44 ± 1.06	70.08 ± 0.68	95.83 ± 0.23	82.23 ± 0.49
DA-Fusion	67.84 ± 1.40	70.34 ± 1.02	96.02 ± 0.64	82.12 ± 0.70
DistDiff	70.22 ± 0.88	71.43 ± 0.51	95.96 ± 0.21	82.67 ± 0.30
Diff-II	68.67 ± 1.73	70.94 ± 1.48	95.98 ± 0.45	82.89 ± 1.16
CoSTA-Diff (Ours)	**71.48 ± 1.71**	**73.73 ± 1.64**	**96.12 ± 0.26**	**83.64 ± 0.46**

CoSTA-Diff improves BA and macro-F1 over Real-only by 2.70 and 2.33 points, and over Vanilla LDM by 3.04 and 3.65 points. Vanilla LDM uses an ordinary AE and class conditioning with the same 100 diffusion-training epochs, but does not improve either metric over Real-only. Thus a larger training set alone does not explain the gain; the synthesis pipeline matters. These results compare complete augmentation pipelines; run-level results are in Appendix A.2.

Table 2 reports three downstream architectures using the same synthetic pool. BA and macro-F1 gains differ in magnitude across ResNet-18, ResNet-50, and DenseNet-121; the effect size depends on the downstream classifier.

Table 2: Paired cross-backbone results over seeds 42, 1042, and 2026 (mean ± sample SD). Δ is CoSTA-Diff minus Real-only in percentage points. All classifiers reuse one synthetic pool; full metrics are in Table 10.

	BA			Macro-F1		
Backbone	Real-only	Ours	Δ	Real-only	Ours	Δ
ResNet-18	68.81 ± 1.66	70.94 ± 1.04	+2.13	70.82 ± 1.46	71.72 ± 1.20	+0.90
ResNet-50	68.78 ± 0.90	71.48 ± 1.71	+2.70	71.40 ± 0.86	73.73 ± 1.64	+2.33
DenseNet-121	69.91 ± 0.99	71.40 ± 1.06	+1.49	72.64 ± 1.11	73.49 ± 1.09	+0.85

4.3 Ablation and Design Studies

We examine how the conditioning interface, latent structure, and anchor selection affect reconstruction and downstream classification.

Conditioning interface. Table 3 compares five completed interfaces. Additive class-and-concept conditioning achieves the highest mean BA and macro-F1. Compared with FiLM concept-only conditioning, it gains 1.87 and 1.54 points, respectively, suggesting that the evidence-fusion interface matters even when neither model uses a direct global-image branch. Class conditioning is retained throughout; "concept-only" means that the direct global-image branch is absent. The final interface is therefore used in all subsequent experiments. Full metrics appear in Table 13.

Table 3: Conditioning design study, mean ± sample SD across three classifier seeds (42, 1042, 2026). Class conditioning is retained in every interface.

Conditioning interface	BA	Macro-F1
FiLM with constant concept slots	68.83 ± 1.02	70.91 ± 0.53
FiLM with image-specific concept evidence	69.68 ± 0.64	71.69 ± 0.49
Legacy additive global-and-concept	69.63 ± 0.98	71.96 ± 0.81
FiLM concept-only	69.61 ± 0.54	72.19 ± 0.69
Additive class-and-concept (Ours)	**71.48 ± 1.71**	**73.73 ± 1.64**

Latent structure and reconstruction. Table 4(a) compares eight configurations from the earlier 193-image structural study. At $6 + 2$ channels, decoder-side exchange improves PSNR over independent branches (44.781 versus 43.643 dB), whereas encoder-only exchange does not (43.527 dB). Combining both exchanges gives the highest PSNR and lowest lesion/boundary errors among the tested configurations. The $6 + 2$ allocation also outperforms $4 + 4$ and $7 + 1$ within the full-exchange family, supporting an asymmetric channel allocation rather than an equal split.

Panel (b) separately evaluates saved checkpoints on 2,196 images: WSL-AE improves PSNR by 1.857 dB while lowering LPIPS and detail error. These observations motivate the detail-preserving design used in the final generator. Full metrics and the 575-image common-unseen subset are in Tables 12 and 15.

Table 4: Latent design evidence. (a) Earlier structural study on 193 images; exchange denotes cross-frequency connections in the encoder and/or decoder. (b) Separate saved-checkpoint comparison on 2,196 images. These panels use different checkpoints and training histories. MAE columns are scaled by 10^3; bold marks the best value within each panel.

(a) Cross-frequency exchange and channel allocation

Exchange	Low+high	PSNR ↑	Lesion MAE ↓	Boundary MAE ↓
Ordinary AE	8	44.567	5.887	4.470
None	$6 + 2$	43.643	6.563	4.974
Encoder only	$6 + 2$	43.527	6.672	5.036
Decoder only	$6 + 2$	44.781	5.700	4.413
Decoder only	$7 + 1$	44.266	6.090	4.611
Encoder + decoder	$4 + 4$	44.128	6.389	4.605
Encoder + decoder	$7 + 1$	44.995	5.509	4.355
Encoder + decoder	$6 + 2$	**45.057**	**5.478**	**4.290**

(b) Saved-checkpoint reconstruction

Model	PSNR ↑	LPIPS ↓	Detail MAE ↓
Ordinary AE	42.406	0.09140	4.679
WSL-AE	**44.263**	**0.05703**	**4.156**

Concept-space selection. Uniform uses all eligible anchors with the same generator, sampler, and class budgets. Concept targeting improves its primary mean BA and macro-F1 by 1.85 and 1.82 points, although per-seed differences have both signs. The anchor-count-matched random control tests whether sparsity ranking adds value beyond restricting the anchor set. Its performance is close to concept targeting, indicating that the observed benefit comes from the combined conditioning-and-targeting design rather than ranking alone (Table 5). The two comparisons use the same classifier seeds but different anchor controls and are reported separately.

4.4 Sensitivity to Augmentation Budget

We draw class-balanced nested subsets of 600 and 1,200 images from the fixed 1,800-image pool using a fixed hash order. Table 6 includes the same-seed Real-only baseline. BA increases across the tested budgets, while macro-F1 gains are concentrated at 1,800 images. For this seed-42 study,

Table 5: Anchor selection controls (mean ± sample SD). Each panel is a paired comparison over classifier seeds 42, 1042, and 2026. The synthetic budget is 1,800 throughout.

Selection rule	BA	Macro-F1
Seeds 42, 1042, 2026: all versus selected anchors		
Uniform	69.63 ± 1.36	71.91 ± 1.25
Concept targeting	71.48 ± 1.71	73.73 ± 1.64
Seeds 42, 1042, 2026: matched anchor counts		
Random selection	70.09 ± 1.48	72.57 ± 1.38
Concept targeting	69.93 ± 1.91	72.67 ± 1.24

the clearest budget-related gains are in BA and macro-F1; AUROC and accuracy are not monotonic (Table 16).

Table 6: Synthetic-budget sensitivity, ResNet-50 seed 42. All classifiers use 40 epochs and validation-selected checkpoints. Nonzero budgets use nested subsets of one pool.

Synthetic images	0 (Real-only)	600	1,200	1,800
BA	68.00	69.57	70.31	71.99
Macro-F1	70.75	71.13	71.36	74.10

4.5 Qualitative Synthesis

Figure 4 pairs real anchors with CoSTA-Diff generations across all six targeted classes, illustrating class coverage and within-class variation.

Single-concept perturbations produce mostly subtle texture, colour, or boundary changes, while noise changes are larger (Figure 5). The examples show the range of responses observed under the fixed generation procedure.

Across the controlled studies, three patterns emerge. The largest downstream gains appear when class information and image-specific concept evidence are fused in the additive conditioner, while the cross-backbone results show that the benefit persists across three classifiers. Concept targeting improves over using all eligible anchors, whereas the matched random control is close, suggesting that the main contribution is the shared conditioning-and-allocation design rather than a ranking rule in isolation. Finally, budget sensitivity shows that the benefit grows with the number of generated images, with the clearest macro-F1 improvement at 1,800 samples. These results motivate CoSTA-Diff as a coordinated synthesis-and-allocation framework.

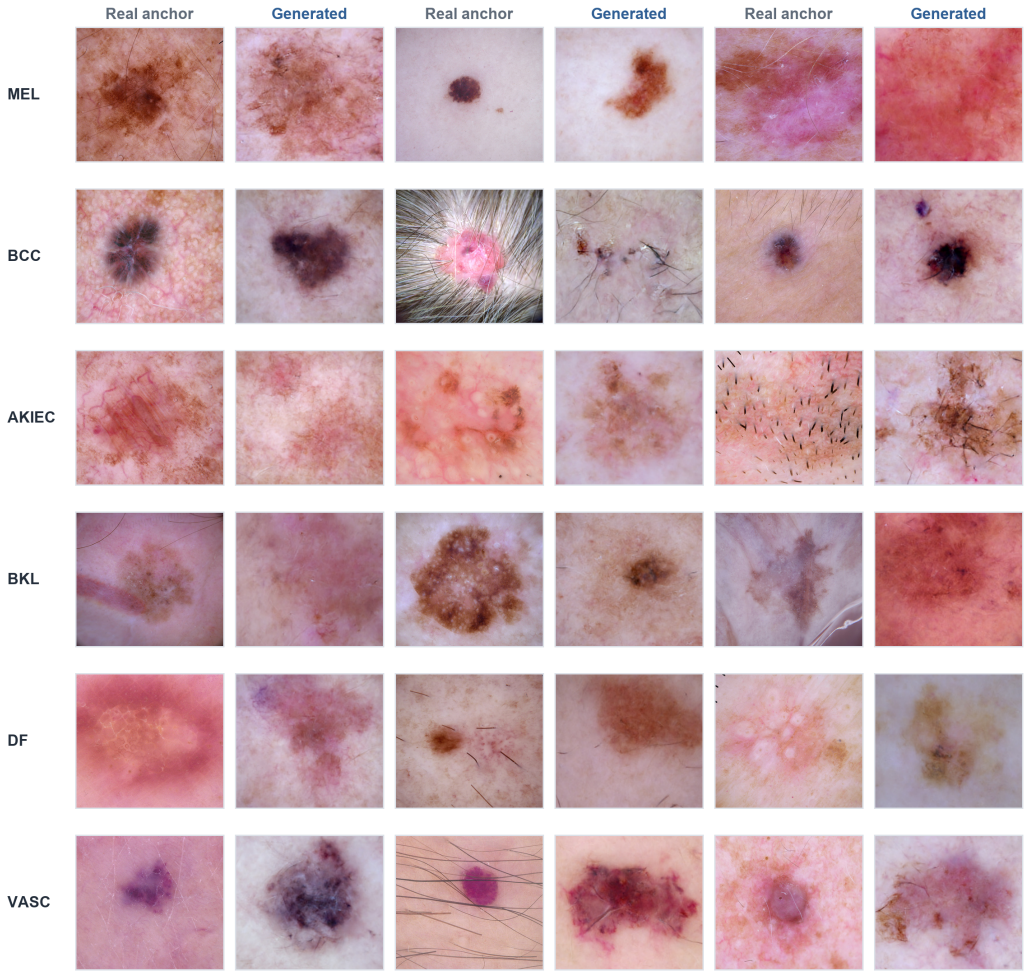

Figure 4: CoSTA-Diff generations paired with real anchors across six lesion classes (one class per row). Each anchor supplies its label and recorded concept evidence; generation starts from fresh noise. Three pairs per class are selected at fixed sparsity quantiles to illustrate class coverage and within-class variation.

5 Conclusion

CoSTA-Diff connects what to generate with where to expand the training set through shared medical concept evidence. Frozen MONET features supply image-specific scores and text directions to an additive class-and-concept conditioner, while neighborhoods in the same evidence space identify sparsely represented within-class profiles. Fresh-noise synthesis under the selected conditions, supported by a compact wavelet latent autoencoder, turns these profiles into additional training examples without manual per-image concept annotations. On the evaluated ISIC 2018 split, a fixed budget of 1,800 synthetic images improves mean ResNet-50 BA and macro-F1 over real-only training by 2.70 and 2.33 percentage points, respectively. Design studies and cross-backbone comparisons characterize the roles of the conditioning interface, latent structure, anchor allocation, and synthesis budget, supporting a shared-representation approach to targeted augmentation.

The current evaluation covers one dataset split, and classifier repetitions reuse a fixed generator and synthetic pool. MONET-derived scores provide image–concept similarity evidence. Future work will evaluate generalization across institutions and datasets, repeat generator training and synthesis to measure full-pipeline variability, and incorporate expert-validated concept annotations to study clinically meaningful augmentation targets.

AI USE STATEMENT

The authors used AI tools to assist with and polish the paper's writing. AI-assisted text and suggestions were checked against the implementation, experiment records, and cited sources. The authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 specify the model, targeting rule, and evaluation setup. Appendices A.1, A.2, A.10, A.12, and A.13 provide detailed protocols, run-level results, implementation details, and the concept vocabulary. Code, experiment configurations, and reproduction instructions will be publicly released upon publication.

REFERENCES

David Jacob Drexlin, Jonas Dippel, Julius Hense, Niklas Prenißl, Grégoire Montavon, Frederick Klauschen, and Klaus-Robert Müller. MeDi: Metadata-guided diffusion models for mitigating biases in tumor classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15973 of *Lecture Notes in Computer Science*, pp. 379–388, 2026. doi: 10.1007/978-3-032-05185-1_37. URL https://link.springer.com/chapter/10.1007/978-3-032-05185-1_37. First online 20 September 2025.

Jamil Fayyad, Nourhan Bayasi, Ziyang Yu, and Homayoun Najjaran. LesionGen: A concept-guided diffusion model for dermatology image synthesis. In *Skin Image Analysis, and Computer-Aided Pelvic Imaging for Female Health*, volume 16149 of *Lecture Notes in Computer Science*, pp. 3–12, 2026. doi: 10.1007/978-3-032-05825-6_1. URL https://link.springer.com/chapter/10.1007/978-3-032-05825-6_1. DGM4MICCAI 2025 workshop; first online September 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Chanwoo Kim, Soham U. Gadgil, Alex J. DeGrave, Jesutofunmi A. Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine*, 30:1154–1165, 2024. doi: 10.1038/s41591-024-02887-x. URL https://www.nature.com/articles/s41591-024-02887-x.

Mujung Kim, Jisang Yoo, Soonchul Kwon, Byung Jun Kim, Changsik John Pak, Chong Hyun Won, Suk-Ho Moon, Woo Jin Song, Han Gyu Cha, and Kyung Hee Park. Diffusion-based skin disease data augmentation with fine-grained detail preservation and interpolation for data diversity. *PLOS ONE*, 20(10):e0331404, 2025. doi: 10.1371/journal.pone.0331404. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0331404.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2007.04612.

Akshay Kulkarni, Ge Yan, Chung-En Sun, Tuomas Oikarinen, and Tsui-Wei Weng. Interpretable generative models through post-hoc concept bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.19377.

Xinmiao Lin, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Yu Kong. Catch missing details: Image reconstruction with frequency augmented variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52729.2023.00173. URL https://arxiv.org/abs/2305.02541.

Tejaswini Medi, Hsien-Yi Wang, Arianna Rampini, and Margret Keuper. Decoupling high and low frequencies for faithful image generation with fine details. arXiv:2509.05441v4, 2026. URL

https://arxiv.org/abs/2509.05441v4. First posted in 2025 as FAVAE-Effective Frequency Aware Latent Tokenizer; version 4 revised 20 August 2026.

Dang Nguyen, Jiping Li, Jinghao Zheng, and Baharan Mirzasoleiman. Do we need all the synthetic data? targeted image augmentation via diffusion models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/98bf3b8505c611ac21055dd9d355c66e-Abstract-Conference.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2112.10752.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. URL https://arxiv.org/abs/1505.04597.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00512.

Junjie Shentu, Matthew Watson, and Noura Al Moubayed. DiDGen: Diffusion-based dual-task synthesis for dermoscopic data generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15970 of *Lecture Notes in Computer Science*, pp. 74–84, 2026. doi: 10.1007/978-3-032-05141-7_8. URL https://link.springer.com/chapter/10.1007/978-3-032-05141-7_8. First online September 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

Tao Sun, Yun Jiang, Yarong Jin, Huanting Guo, and Zequn Zhang. Context-preserving dermoscopic editing: Mask-guided lesion-aware diffusion for attribute modification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5364–5373, 2026. URL https://openaccess.thecvf.com/content/WACV2026/papers/Sun_Context-Preserving_Dermoscopic_Editing_Mask-Guided_Lesion-Aware_Diffusion_for_Attribute_Modification_WACV_2026_paper.pdf.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2302.07944.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5: 180161, 2018. doi: 10.1038/sdata.2018.161. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6091241/.

Yanghao Wang and Long Chen. Inversion circle interpolation: Diffusion-based image augmentation for data-scarce classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25560–25569, 2025. URL https://arxiv.org/abs/2408.16266.

William Yang, Xindi Wu, Zhiwei Deng, Esin Tureci, and Olga Russakovsky. Beyond objects: Contextual synthetic data generation for fine-grained classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22498–22508, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_Beyond_Objects_Contextual_Synthetic_Data_Generation_for_Fine-Grained_Classification_CVPR_2026_paper.html.

Tianchen Zhao, Xuanbai Chen, Zhihua Li, Jun Fang, Dongsheng An, Xiang Xu, Zhuowen Tu, and Yifan Xing. Salient concept-aware generative data augmentation.

In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8fbfec781df35a01bb3ceb7ffdb6b521-Abstract-Conference.html.

Haowei Zhu, Ling Yang, Jun-Hai Yong, Hongzhi Yin, Jiawei Jiang, Meng Xiao, Wentao Zhang, and Bin Wang. Distribution-aware data expansion with diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ba2e53000899e45e6018f639cb7469fa-Abstract-Conference.html.

A Supplementary Experimental Material

A.1 Data and Evaluation Protocol

The source ISIC 2018 training collection contains 10,015 images. Stratified image-level sampling with split seed 42 reserves 2,003 for validation; adding the original 193 validation images gives 2,196 validation images and leaves 8,012 for training. The original 1,512-image test set is unchanged. Classes are melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis/intraepithelial carcinoma (AKIEC), benign keratosis-like lesions (BKL), dermatofibroma (DF), and vascular lesions (VASC). Training class counts in this order are 890, 5,364, 411, 262, 879, 92, and 114. Split seed 42 is distinct from the classifier-training seeds.

The primary classifier analysis uses seeds 42, 1042, and 2026. Seed 2048 was added later as an additional robustness run. We retain the primary three-seed and complete four-seed summaries below, together with every per-seed score. The cross-backbone tables report matched Real-only and CoSTA-Diff runs over seeds 42, 1042, and 2026; the conditioning table reports the completed three-seed design study using seeds 42, 1042, and 2026.

Classifier training and checkpoint selection. ResNet-50 classifiers use ImageNet initialization, 256-pixel inputs, batch size 64, SGD learning rate 0.002, momentum 0.9, weight decay 10^{-4}, label smoothing 0.05, and cosine decay to 10^{-6}. They run for 40 epochs with AMP and without enforced deterministic execution. Checkpoints are selected by maximum validation BA over epochs 10–40, with the earliest epoch used to break ties. The test set is used to evaluate the selected checkpoints. For the earlier Real-only runs, archived prediction exports do not provide the same checkpoint-hash linkage as the later runs.

Proposed generator. The final diffusion model uses U-Net widths 128/256/256/512, two layers per block, context width 256, eight latent channels at 64×64, and 23 conditioning tokens. Training uses 100 epochs/50,000 updates, batch size 16, AdamW at 2×10^{-4} with betas (0.95,0.999), weight decay 10^{-6}, 1,000 warm-up updates, FP16, and velocity MSE with 1,000 cosine diffusion steps. Validation diffusion loss selects epoch 94. The squared-cosine schedule uses zero-terminal-SNR rescaling, uniform timestep sampling, and mean reduction over latent entries. No MONET-consistency or downstream-classification loss is added. DDIM100 with trailing timestep spacing, $\eta = 0$, and no classifier guidance generates one fixed 1,800-image pool. The autoencoder is frozen. Concept selection uses $k = 5$, anchor fraction 0.4 and a maximum of 20 replicas, with class budgets and noise recorded in the generation plan. Training runs use an RTX 5060 Ti with 16 GB memory.

Synthesis baselines. Vanilla LDM combines an ordinary autoencoder with class-label conditioning and is trained for 100 diffusion epochs, matching the proposed generator in epoch count. Epoch count is distinct from the validation-selected checkpoint epoch. The ordinary autoencoder was trained on the earlier 8,036-image training split, whereas downstream classifiers use the retained 8,012-image training split. This baseline follows a complete-pipeline comparison under the retained downstream protocol.

DA-Fusion uses six learned class tokens with SD1.4 image-to-image synthesis. DistDiff uses distribution guidance and existing real-data classifier/prototype assets. Diff-II uses SD1.5, 35,000 adaptation updates, 1,647 source-image inversions and within-class interpolation. Its two generic suffixes come from BLIP training captions rather than the original GPT-4 route. All use the fixed downstream split and synthetic class budget, while retaining their method-specific generative training and adaptation settings.

Reconstruction evaluation. For AE reconstruction, images are converted to RGB, bilinearly resized on the shorter side and centre-cropped to 256. We use posterior mode, FP32 without AMP/TF32, and decoded values clipped to $[-1, 1]$ then mapped to $[0, 1]$. PSNR is averaged per image. SSIM uses an 11-pixel Gaussian window ($\sigma = 1.5$), data range 1, population covariance and channel averaging. LPIPS uses VGG v0.1 on $[-1, 1]$ inputs. Detail MAE averages the LH, HL and HH errors from orthonormal Haar transforms of the reconstructed RGB images; it is not the raw decoder-subband training loss. Saved examples were fixed by ID before viewing reconstructions. Per-image variation is not training-seed variation.

A.2 Additional Classification Results

Table 7: Primary seeds 42, 1042, and 2026. Mean ± sample SD, percentages; all runs are retained.

Method	BA	Macro-F1	AUROC	Accuracy
Real-only	68.78 ± 0.90	71.40 ± 0.86	96.07 ± 0.08	82.87 ± 0.07
Vanilla LDM	68.44 ± 1.06	70.08 ± 0.68	95.83 ± 0.23	82.23 ± 0.49
DA-Fusion	67.84 ± 1.40	70.34 ± 1.02	96.02 ± 0.64	82.12 ± 0.70
DistDiff	70.22 ± 0.88	71.43 ± 0.51	95.96 ± 0.21	82.67 ± 0.30
Diff-II	68.67 ± 1.73	70.94 ± 1.48	95.98 ± 0.45	82.89 ± 1.16
Uniform	69.63 ± 1.36	71.91 ± 1.25	96.36 ± 0.26	82.63 ± 0.64
CoSTA-Diff	71.48 ± 1.71	73.73 ± 1.64	96.12 ± 0.26	83.64 ± 0.46

Table 8: Four-seed summary for methods with available runs at 42, 1042, 2048, and 2026. Vanilla LDM has only the three reported seeds and is not included. Mean ± sample SD, percentages; all runs are retained.

Method	BA	Macro-F1	AUROC	Accuracy
Real-only	68.36 ± 1.12	70.94 ± 1.15	95.96 ± 0.21	82.85 ± 0.06
DA-Fusion	67.97 ± 1.17	70.46 ± 0.87	96.09 ± 0.55	82.08 ± 0.58
DistDiff	69.53 ± 1.56	71.14 ± 0.72	95.99 ± 0.18	82.74 ± 0.28
Diff-II	69.07 ± 1.63	71.31 ± 1.41	96.01 ± 0.37	82.99 ± 0.96
Uniform	70.17 ± 1.55	72.66 ± 1.80	96.27 ± 0.28	83.05 ± 0.99
CoSTA-Diff	70.66 ± 2.14	73.29 ± 1.60	96.17 ± 0.23	83.47 ± 0.52

Table 9: All R50 classifier runs. Seed 2048 remains explicitly reported.

Method	Seed	BA	Macro-F1	AUROC	Accuracy
Vanilla LDM	42	67.23	69.93	95.91	82.41
Vanilla LDM	1042	69.16	69.49	96.01	81.68
Vanilla LDM	2026	68.95	70.82	95.57	82.61
Real-only	42	68.00	70.75	96.06	82.80
Real-only	1042	68.58	71.08	95.99	82.87
Real-only	2048	67.10	69.58	95.66	82.80
Real-only	2026	69.76	72.37	96.15	82.94
DA-Fusion	42	69.14	70.64	95.28	82.01
DA-Fusion	1042	68.01	71.18	96.42	82.87
DA-Fusion	2048	68.35	70.83	96.33	81.94
DA-Fusion	2026	66.36	69.21	96.35	81.48
DistDiff	42	70.65	72.02	95.72	82.61
DistDiff	1042	69.21	71.18	96.05	82.41
DistDiff	2048	67.45	70.26	96.07	82.94
DistDiff	2026	70.80	71.09	96.10	83.00
Diff-II	42	69.97	71.16	95.54	83.86
Diff-II	1042	69.32	72.30	95.96	83.20
Diff-II	2048	70.28	72.42	96.10	83.27
Diff-II	2026	66.71	69.37	96.44	81.61
Uniform	42	68.18	70.68	96.10	81.94
Uniform	1042	69.84	71.89	96.61	82.74
Uniform	2048	71.78	74.89	96.00	84.33
Uniform	2026	70.88	73.17	96.37	83.20
CoSTA-Diff	42	71.99	74.10	96.21	84.13
CoSTA-Diff	1042	69.57	71.93	96.32	83.20
CoSTA-Diff	2048	68.23	71.98	96.30	82.94
CoSTA-Diff	2026	72.87	75.16	95.83	83.60

Table 10: Full cross-backbone metrics, mean ± sample SD over seeds 42, 1042, and 2026. The generator and image pool are fixed across classifier runs.

Backbone	Method	BA	Macro-F1	AUROC	Accuracy
ResNet-18	Real-only	68.81 ± 1.66	70.82 ± 1.46	95.33 ± 0.45	82.10 ± 0.80
ResNet-18	CoSTA-Diff	70.94 ± 1.04	71.72 ± 1.20	95.30 ± 0.32	82.14 ± 0.40
ResNet-50	Real-only	68.78 ± 0.90	71.40 ± 0.86	96.07 ± 0.08	82.87 ± 0.07
ResNet-50	CoSTA-Diff	71.48 ± 1.71	73.73 ± 1.64	96.12 ± 0.26	83.64 ± 0.46
DenseNet-121	Real-only	69.91 ± 0.99	72.64 ± 1.11	95.58 ± 0.06	83.05 ± 0.36
DenseNet-121	CoSTA-Diff	71.40 ± 1.06	73.49 ± 1.09	95.21 ± 1.17	83.20 ± 0.97

A.3 Comparison with Related Approaches

Table 11: Positioning against recent concept-based and targeted synthesis methods. Entries describe the cited mechanisms, not a common-protocol performance comparison.

Method	Conditioning information	Generation or targeting mechanism	Relation to CoSTA-Diff
LesionGen (Fayyad et al., 2026)	Concept-rich captions	Class-balanced prompt sampling	Medical concepts enter through text
DiDGen (Shentu et al., 2026)	Clinical attribute captions	Region-aware training; optional layout-guided pair synthesis	Text and spatial supervision
CPDE (Sun et al., 2026)	Source/target attribute descriptions	MONET-supervised, mask-guided attribute editing	MONET acts in an editing loss
BOB (Yang et al., 2026)	Class, background and pose	Context marginalization during synthesis	Targets class–context associations
TADA (Nguyen et al., 2026)	Conditions of selected examples	Selection using early learning behavior	Targets classifier learning difficulty
MeDi (Drexlin et al., 2026)	Class and metadata	Targeted subpopulation synthesis	Targets metadata-defined groups
CoSTA-Diff	Class, concept evidence and fixed text directions	Same-class concept-neighborhood selection; fresh-noise synthesis	Shared concept coordinates for conditioning and target selection

A.4 Wavelet Autoencoder Design

Table 12: Autoencoder structural comparison on 193 validation images. These runs precede the saved-checkpoint evaluation in Table 15 and use separate trained checkpoints.

Structure	Low+high	Global MAE ↓	PSNR ↑	Lesion MAE ↓	Boundary MAE ↓
Ordinary AE	8	0.004527	44.567	0.005887	0.004470
Independent	6+2	0.004975	43.643	0.006563	0.004974
Encoder-only	6+2	0.005048	43.527	0.006672	0.005036
Decoder-only	6+2	0.004427	44.781	0.005700	0.004413
Decoder-only	7+1	0.004651	44.266	0.006090	0.004611
Full	4+4	0.004709	44.128	0.006389	0.004605
Full	7+1	0.004348	44.995	0.005509	0.004355
Full (selected)	6+2	0.004297	45.057	0.005478	0.004290

A.5 Conditioning Interfaces

Table 13: Conditioning exploration using seeds 42, 1042, and 2026. Values report mean ± sample SD across the three classifier seeds; only completed interfaces are shown.

Conditioning interface	BA	Macro-F1	AUROC	Accuracy
FiLM with constant concept slots	68.83 ± 1.02	70.91 ± 0.53	95.87 ± 0.34	82.43 ± 0.40
FiLM with image-specific concept evidence	69.68 ± 0.64	71.69 ± 0.49	95.70 ± 0.20	82.76 ± 0.33
Legacy additive global-and-concept conditioning	69.63 ± 0.98	71.96 ± 0.81	95.73 ± 0.10	82.41 ± 0.81
FiLM concept-only conditioning	69.61 ± 0.54	72.19 ± 0.69	96.04 ± 0.15	83.18 ± 0.43
Additive class-and-concept (Ours)	**71.48 ± 1.71**	**73.73 ± 1.64**	**96.12 ± 0.26**	**83.64 ± 0.46**

A.6 Anchor Selection

Table 14: Concept versus Uniform under the primary three-seed summary (seeds 42, 1042, and 2026). Full seed records and the four-seed summary are supplied above.

Method	BA	Macro-F1	AUROC	Accuracy
Uniform	69.63 ± 1.36	71.91 ± 1.25	96.36 ± 0.26	82.63 ± 0.64
CoSTA-Diff	71.48 ± 1.71	73.73 ± 1.64	96.12 ± 0.26	83.64 ± 0.46

Anchor-count-matched random control. Uniform permits all training anchors; the random control instead matches the number of selected anchors and per-class generation budget to concept targeting. The paired runs use the primary classifier seeds 42, 1042, and 2026. Concept targeting obtains BA 69.93 ± 1.91 and macro-F1 72.67 ± 1.24, versus 70.09 ± 1.48 and 72.57 ± 1.38 for matched random selection. The respective mean differences are -0.16 and $+0.10$ percentage points; this near-tie provides a complementary control for interpreting the selection effect.

A.7 Reconstruction Results

Table 15: Saved-checkpoint reconstruction comparison for the corresponding reconstruction models.

Images	Model	PSNR ↑	SSIM ↑	LPIPS ↓	Detail MAE ↓
All 2,196	Ordinary	42.406	0.97002	0.09140	0.004679
All 2,196	Full	44.263	0.97730	0.05703	0.004156
Common-unseen 575	Ordinary	42.517	0.97052	0.09124	0.004596
Common-unseen 575	Full	44.410	0.97773	0.05675	0.004078

A.8 Synthesis Budget

Table 16: Synthetic-count sensitivity with ResNet-50, seed 42 only. Each point uses 40 epochs and validation-selected checkpoints; no training-seed error bars are available.

Images	BA	Macro-F1	AUROC	Accuracy
600	69.57	71.13	96.39	83.47
1200	70.31	71.36	96.06	83.07
1800	71.99	74.10	96.21	84.13

A.9 Concept Response and Qualitative Sampling

For the final additive class-and-concept generator, six anchors, three concepts, and two noise seeds are fixed. Each concept is set to its within-class 10th percentile, observed value, or 90th percentile while other scores and noise remain fixed (36 triplets). Low/high pairs differ at the pixel level, but mostly show small texture, colour, or boundary changes without clear attribute progression; noise changes are larger.

Figure 4 uses existing images from the fixed synthetic pool. Within each class, records were ordered by the stored anchor sparsity score with synthetic ID as a tie-breaker, and records nearest the 10th, 50th, and 90th percentile positions were selected deterministically. Each is paired with its recorded real anchor, illustrating the fixed-pool sampling procedure.

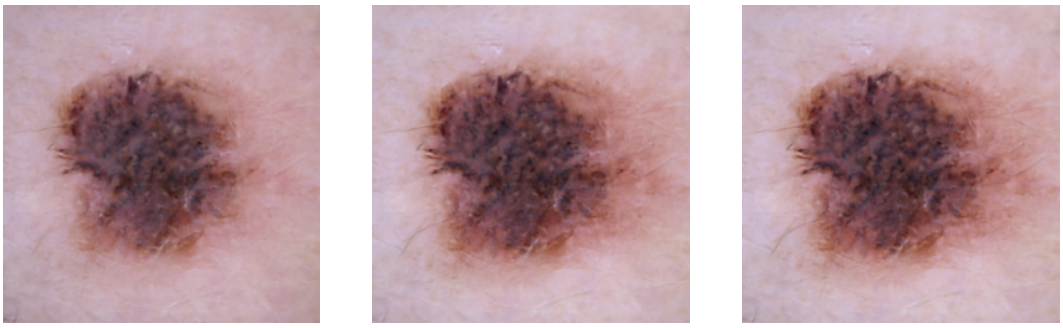

Figure 5: First predefined pigment-network triplet (MEL anchor ISIC_0024545, noise seed 42): low, observed, and high concept score, from left to right; all three images are generated. The triplet illustrates the range of visible responses under the fixed generation procedure.

A.10 Implementation Details

The MONET similarity scale is fixed at $\lambda = \exp(4.5944)$. The conditioner uses a shared $1 \to 64 \to 256$ score MLP with SiLU and biases, and a shared bias-free $768 \to 256$ projection for the text directions. Concept identity embeddings are concept-specific; type embeddings and LayerNorms are shared within their token type. LayerNorm acts on each token's 256-dimensional feature axis. The class embedding has seven entries, and the conditioner uses dropout zero. The directions and normalization statistics are fixed; a zero standardized concept score corresponds to its training-cache mean.

The Haar transform is the fixed orthonormal one-level transform applied independently to RGB channels. The low and high encoder features have 128 and 64 channels at 64×64 resolution; their 192-channel concatenation is fused to 256 channels. The complete six-plus-two latent enters the low decoder, and the projected two-channel high pathway plus the low-decoder feature enter the detail decoder. The high decoder predicts the nine detail-band channels. The diagonal-Gaussian KL used in Eq. 6 is summed over all $8 \times 64 \times 64$ latent entries before averaging over images.

A.11 Scope and Limitations

The AE structural study and saved-checkpoint reconstruction comparison address complementary questions: architecture design and reconstruction quality. Classifier repetitions use a fixed generator and synthetic pool, so their standard deviations summarize downstream-training variation. The reported comparisons include the completed runs in the final evaluation plan.

A.12 Normalization, Diffusion, and Allocation Details

Latent normalization and image synthesis. Per-channel means μ_z and standard deviations σ_z are estimated from training-image posterior-mode latents and broadcast over spatial positions:

$$\widetilde{z}_0 = (z - \mu_z) \oslash \sigma_z, \qquad z = \widetilde{z}_0 \odot \sigma_z + \mu_z, \tag{9}$$

For each allocated job j, the anchor label and evidence are retained with the fixed direction bank $\mathbf{d}$:

$$\widehat{x}_{i,j} = D_\omega(\mathrm{DDIM}_\theta(\xi_{i,j}; \mathcal{C}_\psi(y_i, c_i; \mathbf{d})) \odot \sigma_z + \mu_z), \quad \xi_{i,j} \sim \mathcal{N}(0, I). \tag{10}$$

Here DDIM_θ returns the normalized final latent and D_ω is the frozen decoder. The synthetic set is $\mathcal{D}_{\mathrm{syn}} = \{(\widehat{x}_{i,j}, y_i) : i \in \cup_c \mathcal{A}_c, 1 \le j \le n_i\}$.

For image domain $\mathcal{X}$, the concept representation is defined below. Conditioning uses training-cache mean and standard deviation; selection uses separately fitted training-set statistics with its stated small-variance fallback. The forward schedule has $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, with Gaussian noise ϵ. The autoencoder uses $\mu = [\mu_L, \mu_H]$ and log variance $\ell = [\ell_L, \ell_H]$; MAE averages over tensor entries. A fixed-seed shuffle followed by round-robin allocation assigns generation jobs to selected anchors. For selection, $\mathcal{A}_c$ contains the indices of the highest-ranked profiles. Classes with

$B_c = 0$ have an empty anchor set; allocation requires $B_c \le 20|\mathcal{A}_c|$. Tied sparsity scores receive midpoint percentile ranks.

$$\mathcal{S}_{\text{med}} = \{c(x) : x \in \mathcal{X}\} \subseteq \mathbb{R}^{22}, \tag{11}$$

$$\bar{c}_{i,m} = \frac{c_{i,m} - \mu_m^{\text{cond}}}{\max(\sigma_m^{\text{cond}}, 10^{-6})}. \tag{12}$$

$$\begin{aligned} \widetilde{z}_t &= \sqrt{\bar{\alpha}_t}\,\widetilde{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \\ v_t &= \sqrt{\bar{\alpha}_t}\,\epsilon - \sqrt{1-\bar{\alpha}_t}\,\widetilde{z}_0. \end{aligned} \tag{13}$$

$$q_\phi(z \mid x) = \mathcal{N}(\mu, \operatorname{diag}(\exp \ell))\,. \tag{14}$$

$$\begin{aligned} \mathcal{L}_{\text{image}} &= \text{MAE}(x, \widehat{x}), \\ \mathcal{L}_{\text{high}} &= \frac{1}{3} \sum_{b \in \{LH, HL, HH\}} \text{MAE}(x_b, \widehat{x}_b). \end{aligned} \tag{15}$$

$$s_{i,m} = \frac{c_{i,m} - \mu_m^{\text{sel}}}{\delta_m}, \qquad \delta_m = \begin{cases} \sigma_m^{\text{sel}}, & \sigma_m^{\text{sel}} > 10^{-6}, \\ 1, & \text{otherwise.} \end{cases} \tag{16}$$

$$M_c = \min(B_c, \lceil \rho N_c \rceil)\,, \qquad \mathcal{A}_c = \text{TopIndices}_{M_c}\{r_i : y_i = c\}. \tag{17}$$

$$\sum_{i \in \mathcal{A}_c} n_i = B_c, \qquad 1 \le n_i \le 20, \qquad |n_i - n_j| \le 1 \quad (i, j \in \mathcal{A}_c). \tag{18}$$

$$\pi_i = \begin{cases} n_i / B_c, & i \in \mathcal{A}_c, \\ 0, & i \in \mathcal{I}_c \setminus \mathcal{A}_c. \end{cases} \tag{19}$$

A.13 Concept Vocabulary and Prompt Bank

The fixed vocabulary is an engineering concept bank used for reproducible scoring. Each of the three synonyms for a concept is inserted into each of three target/reference pairs, giving nine pairs per concept. Eq. 1 averages these pairs. The templates are:

1. Target: "This is a dermoscopic image of a skin lesion showing [concept]." Reference: "This is a dermoscopic image of a skin lesion."
2. Target: "This dermoscopic photograph shows a skin lesion with [concept]." Reference: "This dermoscopic photograph shows a skin lesion."
3. Target: "Dermoscopy of a skin lesion showing [concept]." Reference: "Dermoscopy of a skin lesion."

Table 17: Fixed concept identifiers and exact synonym descriptions.

Concept	Synonym descriptions
pigment_network	pigment network; reticular pigment network; melanocytic pigment network
atypical_pigment_network	atypical pigment network; irregular pigment network; pigment network with thick lines and irregular holes
negative_network	negative pigment network with pale serpiginous lines; reverse pigment network with light grid lines and darker holes; inverse pigment network with pale branching lines around darker globules
irregular_peripheral_streaks	irregular peripheral streaks at the lesion edge; asymmetric peripheral streaks at the lesion edge; uneven peripheral streaks at the lesion border
regular_dots_globules	regularly distributed brown dots and globules; uniform brown dots and globules; evenly distributed pigmented dots and globules
atypical_dots_globules	atypical dots and globules; irregular dots and globules; unevenly distributed dots and globules of variable size and shape
lesion_asymmetry	asymmetric lesion; asymmetric lesion shape; overall asymmetric lesion outline
color_variegation	asymmetric color variation within the lesion; uneven color variation within the lesion; nonuniform color variation within the lesion
blue_white_veil	blue-white veil; confluent blue-white structureless area; bluish-white veil
scar_like_depigmentation	scar-like depigmentation within the lesion; white regression-like depigmented area within the lesion; focal porcelain-white regression-like area within the lesion
blue_gray_peppering	blue-gray peppering; fine blue-gray granules; pepper-like blue-gray granularity
atypical_vascular_pattern	atypical vascular pattern with irregular vessels; polymorphous atypical vessels; irregular dotted and linear vessels
milia_like_cysts	milia-like cysts; small white or yellow milia-like cysts; multiple white or yellow round cyst-like structures
comedo_like_openings	comedo-like openings; dark keratin-filled comedo-like openings; multiple comedo-like openings
scale_hyperkeratosis	surface scale or hyperkeratosis; hyperkeratotic scale; keratin crust on the lesion surface
arborizing_vessels	arborizing vessels; tree-like branching vessels; large branching telangiectatic vessels
blue_gray_ovoid_nests	blue-gray ovoid nests; ovoid blue-gray nests; multiple blue-gray ovoid nests
leaf_like_areas	pigmented leaf-like areas; maple leaf-like pigmented areas; leaf-like pigmented structures
ulceration_erosion	surface ulceration or erosion; superficial erosion; ulcerated lesion surface
strawberry_pattern_red_pseudonetwork	strawberry pattern with a red pseudonetwork around follicular openings; pink-red pseudonetwork surrounding hair follicles; erythematous strawberry pattern with prominent follicular openings
central_white_scar_like_patch	central white scar-like patch; central white scar-like area within the lesion; sharply demarcated white scar-like patch in the lesion center
vascular_lacunae	sharply demarcated red purple vascular lacunae; clustered red purple blue vascular lacunae; multiple round red purple vascular lacunae