Unified Backbone Refinement for Diffusion Models via Internal-Latent Analysis

Haksoo Lim[1], Myeongjin Lee[1], Wonjoon Chang[1], and Jaesik Choi[1,2]

[1] Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
[2] INEEJI, Seongnam 13558, Republic of Korea

Fig. 1: We propose DUNE, a method that freely improves image quality, meaning no additional training or increase of memory. DUNE mitigates visual artifact while preserving semantic consistency.

Abstract. Diffusion models have achieved remarkable success across diverse domains, with performance closely related to the denoising backbones that parameterize the score function. In this paper, we present a systematic, phase-aware analysis of diffusion components and show that abrupt, early-stage fluctuations in deep latents are strongly associated with artifacts. Guided by these findings, we introduce **DUNE** (**D**iffusion **U**nified Network refi**NE**r), a training-free refinement framework that detects abrupt deviations in deep low-noise internal latents using a shared EMA-based criterion, and applies backbone-specific suppression to the detector-selected entries. Although derived from U-Net, the same detect–suppress principle extends naturally to Transformer-based diffusion models by acting on the latents of deep self-attention blocks. Extensive experiments across multiple backbones indicate that

DUNE improves fidelity while reducing hallucinations, offering new insight into where and when diffusion backbones should be controlled.

Keywords: Diffusion models · Explainability

1 Introduction

Recently, diffusion models have achieved remarkable results across various computer vision domains, including image generation, super-resolution, and image editing [6–8]. Surpassing other generative models, diffusion models have also been successfully extended to other domains such as video generation, voice synthesis, and time series forecasting [20]. These models operate through specific procedures, namely a *forward process* and *reverse process* [6,18,20]. During the forward process, the model incrementally adds noise to an input image, finally transforming it into Gaussian noise. In the reverse process, the model learns a gradient of the log-likelihood, also called a *score function*, using specialized architectures, most commonly U-Net and Transformer variants [12,15]. Once trained, diffusion models generate new images by sampling from a Gaussian distribution and applying the learned score function to iteratively denoise the sample.

Across a wide range of applications, the U-Net architecture has typically been employed as the foundational framework for diffusion models [6, 13, 14]. A U-Net consists primarily of *downsampling blocks*, a bottleneck region (known as *h-space*), *skip connections*, and *upsampling blocks*. Recent work reveals that skip connections and upsampling blocks play distinct internal roles, emphasizing high-frequency and low-frequency components, respectively [16]. Although existing refinement methods manipulate internal activations or the output of the score network, improper control of these features can lead to deviations from the original image, intensifying visual artifacts—commonly referred to as hallucinations—as well as overly simplified backgrounds and exaggerated contrasts between light and dark regions (see Figure 1).

In this work, we observe that artifact-associated score dynamics are concentrated in early denoising stages and are reflected in deep internal latents. Through component-wise and phase-aware analysis, we find that detecting and correcting such anomalies in the deepest latent space of U-Net (h-space) substantially improves fidelity while preserving semantics (see Figure 4). Motivated by these insights, we introduce **DUNE**: **D**iffusion **U**nified Network refi**NE**r, a *training-free* refinement framework that suppresses detector-selected internal deviations during the detect phase and optionally adjusts perceptual details in a later phase.

Our analysis reveals that diffusion models already know which parts contain anomalies and can substantially enhance quality by explicitly addressing these areas (cf. Section 3, 4). We partition generation into two phases: a *detect phase*, where anomalies are identified and corrected in the latent representations, and an optional *detail phase*, where internal manipulations can be applied for user-controllable perceptual adjustments such as tone and saturation. We extend the detect–suppress framework to Transformer-based diffusion models by

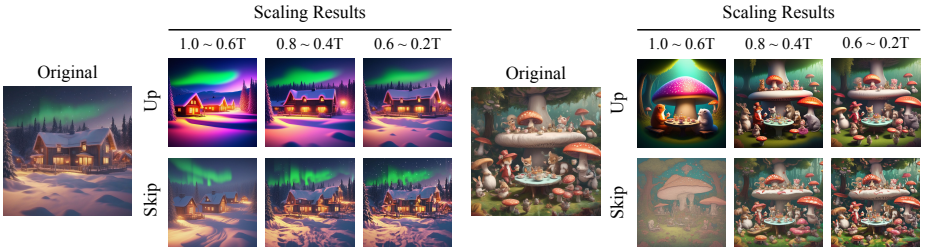

Fig. 2: Effects of scaling U-Net components across denoising steps. Upsampling blocks affect brightness and tone, while skip connections enhance edge details. In early phases, however, scaling primarily leads to semantic changes.

operating on deep self-attention latents that satisfy the same intervention criterion as U-Net h-space: semantic integration with reduced stochastic fluctuation. These improvements incur no additional training or model-specific finetuning, and deliver consistent gains across metrics and models (see Figure 6). Across extensive experiments, DUNE consistently improves fidelity and robustness over training-free refinement baselines.

Our contributions are summarized as follows:

- Through component-wise and phase-specific analysis of U-Net backbones, we show that artifact-associated dynamics are concentrated in early denoising stages and are reflected in deep internal latents, motivating targeted phase-aware control.
- Based on theoretical and empirical analysis of deep intervention points in U-Net h-space and Transformer self-attention latents, we propose DUNE, which applies a shared EMA-based detector and early masked-correction template with backbone-specific suppressors.

2 Background & Observation

We adopt the DDPM [6] formulation: a forward process corrupts clean data $X_0 \sim q(X_0)$ into $X_t = \sqrt{\bar{\alpha}_t}\, X_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and a score network $\epsilon_\theta(X_t, t)$ is trained to predict the injected noise, equivalently estimating the score function [6, 17]. Building on this foundation, we first review related training-free refinement methods, then systematically analyze how individual backbone components influence artifact emergence across denoising phases.

2.1 The Roles of U-Net Components across denoising steps

In this section, we explore the role of U-Net components considering phases of generation, i.e., denoising steps. Specifically, we scale the upsampling blocks and

skip connections during particular durations and analyze how each component, along with the timing of scaling, influences the generated images[3].

Figure 2 presents the scaling results which indicate three observations. First, scaling upsampling blocks improves the quality of the generated images with smoothened textures, consistent with prior observations in FreeU [16]. In addition, we can also observe that these blocks also contribute to increasing the brightness and tone of the images. Second, while FreeU reports that naively scaling the skip connections show minimal effects, we find that scaling them during mid-to-late phases (e.g., after 0.6T) enhances the detail information and clarifies edges by sharpening faint or indistinct contours in the generated images, suggesting that skip connections are in charge of detail information during mid-to-late phases. Lastly, it is important to note that scaling components during early phases (1.0T $\sim$ 0.6T and 0.8T $\sim$ 0.4T) has a more substantial impact on altering the semantics of the generated images than on fulfilling the specific roles of each component described above. This indicates that, during early phases, the semantic structure formation dominates over the individual functional contributions of the U-Net components.

Our hypothesis. Based on the above observations, unexpected semantic variations (e.g., artifacts) are associated with unstable early-phase latent dynamics in the score network, and naively scaling each latent might exaggerate anomalies and making visual artifacts. We therefore hypothesize that abrupt early-phase variations in component latents are strongly associated with artifact emergence in generated images.

3 Analysis of Internal Latents

In this section, we investigate the depth-wise characteristics of the U-Net architecture to support our use of the internal latent spaces—*h-space* for U-Net and the self-attention latents for Transformer backbones. In typical U-Net–based diffusion models, downsampling is performed by convolutional layers (CNNs), halving the spatial dimensions and increasing channel depth [3, 11, 13]. The following propositions formally describe how U-Net effectively reduces the noise in diffused images and concentrates the semantic content within the h-space.

Proposition 1. *Given an $n \times n$ data matrix $X_0 = \{X_0^{i,j}\}_{1 \le i,j \le n}$ and an $m \times m$ convolutional kernel $K = \{\lambda_{i,j}\}_{1 \le i,j \le m}$ used in a downsampling layer, the standard deviation of the noise added to diffused data X_t (for $t \in 0, 1, ..., T$) is scaled by a factor of $||K||_2 = \sqrt{\sum_{i,j} \lambda_{i,j}^2}$ through the convolutional operation.*

We empirically checked that the mean of kernel norms in each downsampling block are sufficiently small, effectively suppressing the added noise and confirming the theoretical prediction of Proposition 1 (e.g., SDXL shows kernel norms of 0.2036 and 0.1509 in its respective downsampling layers).

[3] In this experiment, we scale input feature maps of all upsampling blocks and do not scale the shallowest skip connections since they are too sensitive, only scaling the remaining skip connections.

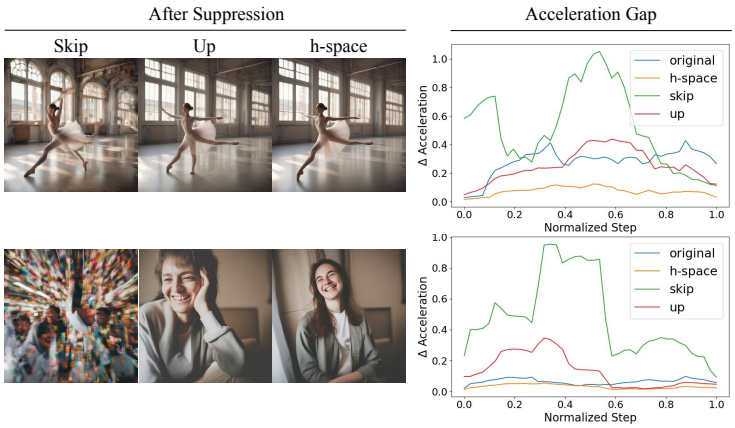

Fig. 3: Suppression Results with Different Components in U-net Controlling. The change of the score function according to denoising time step is called *acceleration* [4].

This inherent denoising capability early in the reverse trajectory clarifies *where* to intervene. However, directly correcting these anomalies in shallow layers poses challenges, as simultaneous adjustments of image content and added noise are needed to maintain the normal distribution of the score function outputs. Conversely, corrections applied in the h-space are less problematic, as this deeper latent space inherently condenses semantic information while effectively reducing noise. The first column of Figure 3 clearly illustrates that corrections in shallower layers, such as skip connections or upsampling blocks, increase score deviations in hallucinated regions, whereas corrections within the h-space significantly mitigate these deviations. Therefore, we operate our detect–suppress mechanism primarily in h-space.

Next, we extend our analysis into Transformer-based diffusion models. The following proposition shows that the latent of self-attention layer has innate denoising capability like h-space in U-Net.

Proposition 2. *Given tokens* $x_j = s_j + \varepsilon_j \in \mathbb{R}^d$ *with i.i.d.* $\varepsilon_j \sim \mathcal{N}(0, \sigma^2 I_d)$ *and attention weights* $w = \mathrm{softmax}((qK^\top)/\tau)$ *(*$w_j \geq 0$, $\sum_j w_j = 1$*), the attention output* $y = \sum_{j=1}^N w_j x_j$ *scales the isotropic noise standard deviation by* $\sqrt{\sum_{j=1}^N w_j^2}$*, i.e.,*

$$\mathrm{Cov}(y_{\mathrm{noise}}) = \sigma^2 \left(\sum_{j=1}^N w_j^2 \right) I_d,$$

so $\sum_j w_j^2 \in [1/N, 1]$ *plays the role of a* data-adaptive kernel norm *(cf. Proposition 1 for CNNs).*

Thus the standard deviation scales by $\sqrt{\sum_j w_{ij}^2} \in [1/\sqrt{N}, 1]$, exactly mirroring the role of the kernel norm in Proposition 1 but now *data–adaptive* via

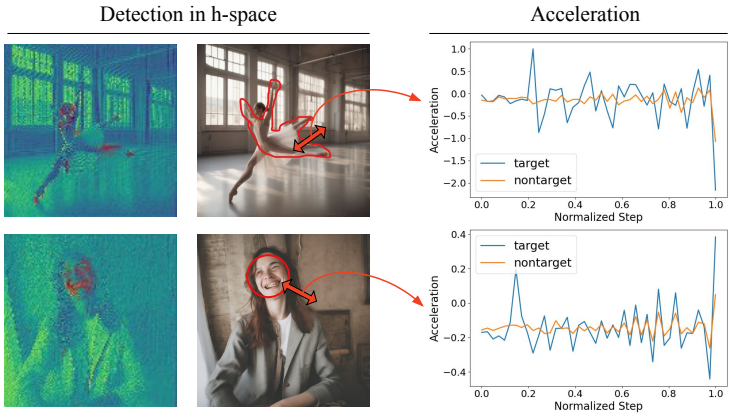

Fig. 4: We visualize how detector-selected regions align with abrupt score-acceleration patterns. The plot shows the averaged score function across diffusion steps, with "Target" and "Non-target" representing the inner and outer regions of the red bounding contour, respectively.

softmax. As the attention distribution becomes more diffuse (higher entropy; e.g., larger temperature), $\sum_j w_{ij}^2 \downarrow$ toward $1/N$ and the denoising effect strengthens. Stacking layers therefore yields a deep token space that is both semantically integrated and noise–reduced, making mid–to–deep self–attention representations the Transformer analogue of U-Net's h–space and a natural sweet spot for detection and suppression.

The two results above justify our design: (i) downsampling convolutions scale (and often contract) noise toward the U-Net bottleneck, and (ii) self-attention performs noise averaging over tokens before value projection. Operating our detect–suppress mechanism in these deepest latents stabilizes score dynamics (reducing abnormal spikes) and preserves semantics.

Transformer backbones lack explicit skip-/upsampling branches. We therefore define the extension through the intervention criterion. In both architectures, DUNE targets a deep semantically integrated latent where stochastic fluctuations are attenuated. For Transformers, Proposition 2 justifies mid-to-deep self-attention latents as this operating region, and we further validates this empirically through latent visualization and target-layer ablation in Appendix.

4 Proposed Method

In the previous section, we identified that abrupt changes in the h-space of U-Net components are strongly associated with the emergence of artifacts in generated images. This phenomenon mainly appears at early denoising phases, where semantic properties are determined. Based on these observations, we propose a component-aware refinement framework, **DUNE**: **D**iffusion **U**nified Network refi**NE**r. Leveraging insights from Section 2, 3, we divide the denoising process

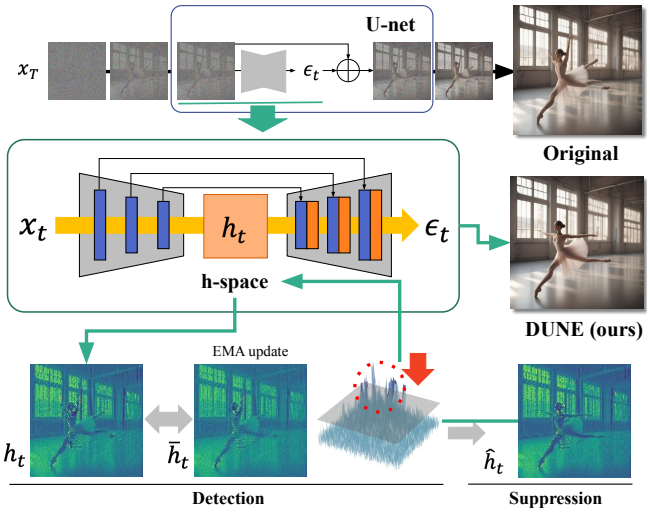

Fig. 5: Overall framework of DUNE. The h-space is directly matched with self-attention latent in Transformer-based diffusion models.

into two phases: the *detect phase* (1.0T–0.6T) and the *detail phase* (0.6T–0.0T). This division reflects that refinements in early and later stages predominantly affect global structure and local details, respectively. Note that the detect phase constitutes the core of DUNE and is sufficient for artifact suppression; the detail phase is an optional extension for stylistic control.

The key functionalities and usage of DUNE across the two phases are described below. In the detect phase, DUNE identifies and corrects potential anomalies within the h-space. The illustrative description of this process is shown in Figure 5. The detail phase is an optional, user-facing module that enables fine-grained control over brightness, tone, and saturation by selectively scaling skip connections and upsampling blocks during the later denoising steps. This step depends on the design choice of users, which properties they want to emphasize. In our implementation, we adopt the scaling methodology from FreeU [16], restricting it to the detail phase to preserve semantic content established during the detect phase. However, applying FreeU alone may result in overly saturated images, which deviate significantly from prompts intended to produce darker or moodier imagery. We therefore investigate each component in greater detail, enabling users to finely adjust the clarity and tonal qualities of the generated images. Also, since these adjustments are stylistic rather than corrective, we exclude the detail phase from all quantitative evaluations and present its effects qualitatively in Appendix.

4.1 Detection Analysis

Several recent studies link visual artifacts in diffusion models to abnormal temporal variations in the score function [2, 4]. The change of the score function

according to denoising time step is called *acceleration* [4]. Building on these findings, we detect abrupt internal changes in the score network (U-Net and Transformer) over the denoising trajectory. Masks are resized to match the spatial dimensions of the corresponding feature maps, and we compare acceleration between masked and unmasked regions.

During the reverse process to generate images, we obtain latent variable $\mathbf{h}_t$ at normalized timestep $t \in [1, T]$, where T denotes the total number of inference steps. To detect abrupt changes in latent features, we compute the exponential moving average (EMA) of a scaled latent variable, which represents the estimation. Motivated by prior observations that artifact regions exhibit abnormal score dynamics [2, 4], and given that the score function in diffusion models is defined as $s_\theta(t, \mathbf{x}_t) = -\frac{\epsilon_\theta(t, \mathbf{x}_t)}{\sigma_t}$, we apply detection to $z_t = -h_t/\sigma_t$, the score-consistent normalization of h_t, to factor out timestep-dependent amplitude drift and isolate abrupt internal deviations. Consequently, our EMA is defined as: $\bar{\mathbf{z}}_t = \gamma \bar{\mathbf{z}}_{t+1} + (1-\gamma)\mathbf{z}_t$. We use the standard value $\gamma = 0.7$ throughout all experiments without per-backbone tuning. The EMA is initialized as $\bar{\mathbf{z}}_T = \mathbf{z}_T$ at the first inference step.

We then measure deviations between each scaled latent $\mathbf{z}_t$ and its corresponding EMA $\bar{\mathbf{z}}_{t+1}$ at each step. Because these values differ according to the brightness of individual images, we adopt a log-ratio to facilitate consistent comparisons across images:

$$\Delta = \log \left| \frac{\mathbf{z}_t}{\bar{\mathbf{z}}_{t+1}} \right| .$$

Since Δ is less sensitive to image-specific scale and brightness variations, we identify a quantile threshold λ corresponding to the $p\%$ percentile, thereby isolating the top $(1-p)\%$ most deviated features. The resulting binary mask M is defined as $M = (\Delta > \lambda)$, highlighting anomalous regions (see second column in Figure 4). Figure 4 also confirms that detected regions exhibit strong acceleration fluctuations concentrated in the first 40% of steps (the detect phase), and that the same logic applies to Transformer self-attention latents.

4.2 Suppression Analysis

We use masked EMA suppression as the canonical correction objective for detector-selected low-noise latents. Transformer latents instantiate this objective through masked EMA blending. U-Net h-space contains low-resolution channels with specialized semantic activations, so DUNE uses channel-aware masked scaling to shrink detected anomalous channels while preserving channel specialization. We analyzes this U-Net instantiation on detector-selected low-SNR subsets and relates it to EMA-style suppression in Appendix.

Let $\mathbf{h}_t$ denote the target internal latent at timestep t (U-Net h-space or Transformer self-attention latent), and let M_t be the detected anomaly mask. We apply a masked correction in a unified gated form:

$$\hat{\mathbf{h}}_t = (1 - M) \odot \mathbf{h}_t \; + \; M_t \odot \mathcal{S}_{\text{bb}}(\mathbf{h}_t, \bar{\mathbf{h}}_t; \kappa), \tag{1}$$

where $\mathcal{S}_{\text{bb}}$ denotes a backbone-specific suppression operator, $\bar{\mathbf{h}}_t := -\sigma_t \bar{\mathbf{z}}_t$ is the de-scaled EMA reference used in detection and $\kappa \in [0, 1]$ controls correction strength.

For U-Net-based diffusion models, we suppress detector-selected h-space outliers because our analysis identifies this region as a stable intervention point where artifact-associated deviations are concentrated. Since each $\mathbf{h}_t$ consists of multiple channels with relatively low spatial resolution (e.g., $1280 \times 32 \times 32$ for SDXL), where each channel encodes distinct semantic information. Hence, channels must be treated independently. We compute the absolute value of the channel-wise mean of $\mathbf{h}_t$, denoted by $\mathbf{n}_t$ (shape of $1280 \times 1 \times 1$). Using a scaling factor κ, the corrected latent representation becomes $\mathcal{S}_{\text{UNet}}(\mathbf{h}_t, \bar{\mathbf{h}}_t; \kappa) = \kappa\,(n_t \cdot \mathbf{h}_t)$.

For Transformer-based methods, latents are patchified and self-attention mixes tokens, so simple masked scaling is less stable. We therefore blend masked features with the EMA reference: $\mathcal{S}_{\text{Tr}}(\mathbf{h}_t, \bar{\mathbf{h}}_t; \kappa) = \kappa \mathbf{h}_t + (1-\kappa)\bar{\mathbf{h}}_t$. We report the performance of our detect–suppression methods for both U-Net and Transformer backbones in Figure 6.

Although the algebraic forms differ, both suppressors implement the same local stability objective in detector-selected low-noise latents. The Transformer case admits direct EMA blending, whereas the U-Net case uses a bounded channel-aware instantiation of EMA-style suppression to avoid disrupting specialized h-space channels.

4.3 SNR Analysis of the Detect–Suppress Phase

We refine the theoretical role of the detect–suppress phase by explicitly analyzing the *signal-to-noise ratio (SNR)* in the internal latent that DUNE operates on. We model the scaled latent $\mathbf{z}_t := -\frac{\mathbf{h}_t}{\sigma_t} \in \mathbb{R}^d$. as a sum of a slowly-varying semantic component and a stochastic component:

Assumption 1 *For each timestep t, the scaled latent admits a decomposition $z_t = s_t + n_t$, where $s_t := \mathbb{E}[z_t \mid \mathcal{F}_t]$ and $n_t := z_t - s_t$, In here, s_t denotes a semantic signal, n_t denotes a stochastic component (noise / abnormal spikes) and $\mathcal{F}_t$ contains the conditioning, timestep, and slowly varying semantic state. Then $\mathbb{E}[n_t \mid \mathcal{F}_t] = 0$ by construction.*

Assumption 2 *There exists $\delta \geq 0$ such that $\|s_{t+1} - s_t\|_2 \leq \delta$ for all t in the detect phase.*

Definition 1. *We define the latent SNR at timestep t as*

$$\text{SNR}(z_t) := \frac{\mathbb{E}\|s_t\|_2^2}{\mathbb{E}\|n_t\|_2^2}. \tag{2}$$

Let $M_t \in \{0, 1\}^d$ denote a binary mask (broadcastable to the latent shape) produced by the detect step. Define masked/unmasked parts by $a_{t,M} := M_t \odot a_t$ and $a_{t,\neg M} := (1 - M_t) \odot a_t$. Then we analyze a unified suppression operator:

$$\hat{\mathbf{z}}_t = (1 - M_t) \odot \mathbf{z}_t + M_t \odot \big(\kappa \mathbf{z}_t + (1 - \kappa)\bar{\mathbf{z}}_t\big), \qquad \kappa \in [0, 1]. \tag{3}$$

In our U-Net implementation, the masked channels are shrunk with a channel-aware factor, which can be interpreted as a per-channel version.

Next, define the *noise concentration* and *signal concentration* within the detected mask by

$$\eta_t := \frac{\mathbb{E}\|n_{t,M}\|_2^2}{\mathbb{E}\|n_t\|_2^2} \in [0,1] \quad (4) \qquad \rho_t := \frac{\mathbb{E}\|s_{t,M}\|_2^2}{\mathbb{E}\|s_t\|_2^2} \in [0,1] \quad (5)$$

which capture how much noise and signal energy, respectively, fall within M_t. The following theorem shows SNR gain of EMA blending:

Theorem 3. *Consider Eq.* (3) *and define the EMA estimation error* $e_t := \bar{\boldsymbol{z}}_{t+1} - s_t$. *Under Assumption 1 and assuming* $\mathbb{E}\langle n_t, e_t\rangle = 0$, *the post-suppression SNR satisfies*

$$\frac{\mathrm{SNR}(\hat{z}_t)}{\mathrm{SNR}(z_t)} \;\geq\; \frac{1}{1-(1-\kappa^2)\eta_t + (1-\kappa)^2\varepsilon_t}, \qquad \varepsilon_t := \frac{\mathbb{E}\|e_{t,M}\|_2^2}{\mathbb{E}\|n_t\|_2^2}.$$

Consequently, $\mathrm{SNR}(\hat{\boldsymbol{z}}_t) > \mathrm{SNR}(\boldsymbol{z}_t)$ *whenever*

$$(1-\kappa^2)\eta_t \;>\; (1-\kappa)^2\varepsilon_t.$$

The following corollary shows that suppressing anomalies in deep latents bounds the resulting score perturbation:

Corollary 1. *Let the remaining mapping from the target latent to noise prediction be* $\epsilon_\theta(\cdot, t) = g_t(\cdot)$ *and assume* g_t *is* L_t*-Lipschitz. Then the change in the predicted score satisfies*

$$\| s_\theta(t, \hat{\mathbf{x}}_t) - s_\theta(t, \mathbf{x}_t) \|_2 \;\leq\; \frac{L_t}{\sigma_t} \|\hat{\mathbf{h}}_t - \mathbf{h}_t\|_2 \;\propto\; \frac{L_t(1-\kappa)}{\sigma_t} \|M_t \odot (\mathbf{z}_t - \bar{\mathbf{z}}_t)\|_2.$$

This provides a local stability argument for why suppressing detector-selected residuals can reduce score perturbations.

In the outlier regime selected by our detector, the following lemma shows that replacing masked EMA blending with a masked channel-wise scaling is justified: the only discrepancy consists of (i) a coefficient mismatch $\tilde{\kappa} - \alpha_t$, which we empirically confirm is near zero at our chosen hyperparameters, and (ii) a provably small EMA residual suppressed by the detection threshold λ.

Lemma 1. *Consider the naive EMA blending operator:* $\mathbf{u}_t^{\mathrm{EMA}} := \kappa \mathbf{z}_t + (1-\kappa)\bar{\mathbf{z}}_t, \quad \kappa \in [0,1]$. *For U-Net, define the channel statistic* $n_t \in \mathbb{R}^{C\times 1\times 1}$ *as in Sec. 4.2 and let the channel-aware scaling coefficient be* $\alpha_t := \kappa\, n_t$. *Define the U-Net scaling output in the scaled-latent space as* $\mathbf{u}_t^{\mathrm{UNet}} := \alpha_t \odot \mathbf{z}_t$. *Let* $\tilde{\kappa} := 1 - \gamma(1-\kappa) = \kappa + (1-\kappa)(1-\gamma)$. *Then for any* $p \in [1, \infty]$,

$$\|M_t \odot (\mathbf{u}_t^{\mathrm{EMA}} - \mathbf{u}_t^{\mathrm{UNet}})\|_p \;\leq\; \|M_t \odot (\tilde{\kappa} - \alpha_t) \odot \mathbf{z}_t\|_p \;+\; (1-\kappa)\gamma e^{-\lambda} \|M_t \odot \mathbf{z}_t\|_p.$$

Table 1: Quantitative comparison of Original vs. Fixed models. Arrows indicate the preferred direction for each metric. "–" denotes unsupported backbones; N/A indicates that the method produced pure noise (cf. Figure 7).

Methods	Metric	SDXL	LCM	Kandinsky 3	PixArt-Σ	Hunyuan-DiT
Original	FID ↓	18.86	22.53	21.36	28.75	30.82
	HADM ↓	0.227	0.152	0.250	0.316	0.294
	CLIP sim. ↑	1.0	1.0	1.0	1.0	1.0
FreeU	FID ↓	31.22	27.57	–	–	–
	HADM ↓	0.351	0.237	–	–	–
	CLIP sim. ↑	0.781	0.756	–	–	–
ASCED	FID ↓	22.07	27.60	22.44	N/A	N/A
	HADM ↓	0.227	0.151	0.252	N/A	N/A
	CLIP sim. ↑	0.945	0.995	0.924	N/A	N/A
PAG	FID ↓	21.72	35.87	–	32.61	–
	HADM ↓	0.210	0.123	–	0.284	–
	CLIP sim. ↑	0.821	0.362	–	0.840	–
DUNE	FID ↓	**18.75**	**22.41**	**20.76**	**27.80**	**30.67**
	HADM ↓	**0.074**	**0.052**	**0.095**	**0.254**	**0.283**
	CLIP sim. ↑	0.925	0.940	0.933	0.952	0.988

5 Experiments

5.1 Experimental Environments

We evaluate DUNE across various high-resolution diffusion models. We select widely-adopted U-Net-based models (Stable Diffusion XL (SDXL) [13], Latent Consistency Model (LCM) [11], and Kandinsky 3 [3]) and Transformer-based models (PixArt-Σ [5] and Hunyuan-DiT [9]). Our implementations use the publicly available `diffusers`[4] library. Both qualitative and quantitative analyses are provided to demonstrate that DUNE consistently refines and enhances generated images.

We compare against three representative training-free model-augment refinement methods: FreeU [16], ASCED [4] and PAG [1]. FreeU reweights backbone and skip features of a pre-trained diffusion U-Net at inference time. ASCED analyzes temporal score dynamics to detect artifact-prone regions and injects corrective noise during sampling. PAG controls internal latent of self-attention layer of diffusion backbones. We give detailed description of these models in Appendix. Note that we use official implementations in `diffusers` and the authors' code for ASCED[5], following the authors' published default settings.

For quantitative evaluation, we assess the overall generation quality by using the Fréchet Inception Distance (FID) metric[6] calculated on 5,000 images generated from COCO prompts [10]. We also report HADM-L [19], a detection-based metric that localizes fine-grained anatomical defects (e.g., extra fingers,

[4] https://github.com/huggingface/diffusers.git
[5] `https://github.com/YuCao16/ASCED.git`
[6] https://github.com/boomb0om/text2image-benchmark

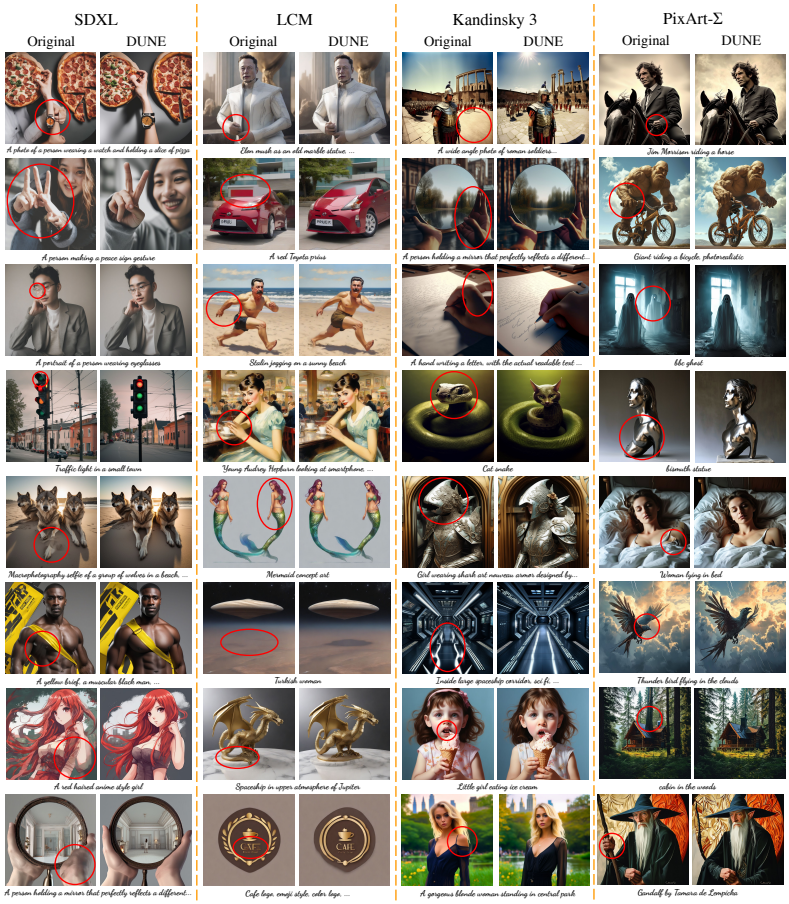

Fig. 6: An experimental result on high-resolution text to image generation. We give remained Hunyuan-DiT figures in Appendix.

distorted joints). For HADM analysis, we generate 5,000 images from human-related prompts by using the human-related annotation files from official COCO resource, following HADM paper [19]. Lastly, since we observed existing refinement methods fail to preserve consistent semantic (e.g., FreeU), we calculate clip similarity to check original-to-refined semantic consistency: cosine similarity of CLIP embeddings between original and refined images. Note that unless otherwise stated, all quantitative results report the detect phase only (detect–suppress without detail-phase reweighting), ensuring a fair comparison with baselines that do not employ component scaling [7].

[7] We adopt this over text–image CLIP score, as we found that overly saturated outputs can paradoxically achieve higher CLIP scores, making it unreliable for evaluating refinement quality.

Table 2: User survey result for qualitative comparison of Original (red) vs. DUNE outputs (blue).

Model	Dataset	Original	DUNE
DDPM	CelebA-HQ	6.60%	**93.40%**
	Bedroom	15.28%	**84.72%**
VE	FFHQ	11.51%	**88.49%**
	Church	16.67%	**83.33%**

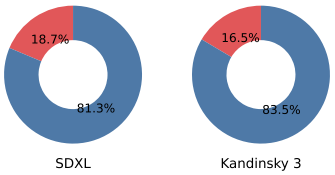

Fig. 7: Visual comparison of ASCED and DUNE on PixArt-Σ.

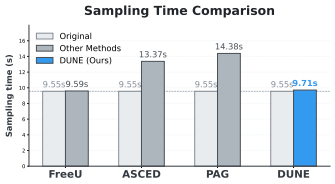

Fig. 8: Sampling time comparison on SDXL.

5.2 Enhancing Quality with DUNE

We first verify that DUNE suppresses detector-selected unstable regions. Figure 9 (extending Figure 4) depicts score-acceleration in target/non-target regions before/after DUNE. DUNE attenuates the acceleration spikes in the detector-selected target region, while the non-target region remains largely unchanged. We quantify this effect using normalized acceleration amplification (NAA) and excess acceleration gap reduction (EAGR):

$$\mathrm{NAA}(M) = \mathbb{E}_t\left[\frac{\sqrt{\mathbb{E}_{c,i\in M}\, a_t(c,i)^2}}{\sqrt{\mathbb{E}_{c,i\notin M}\, a_t(c,i)^2}+\epsilon}\right], \quad \mathrm{EAGR} = 1 - \frac{\mathrm{NAA}_D^{TopK} - \mathrm{NAA}_D^{Rand}}{\mathrm{NAA}_O^{TopK} - \mathrm{NAA}_O^{Rand} + \epsilon}.$$

Here a_t denotes score acceleration, D/O denote DUNE/original sampling, TopK is the detector mask, and Rand is a density-matched random mask. DUNE reduces the excess target-vs-random acceleration gap by 24.9% on SDXL and 18.1% on PixArt-Σ; random-mask control shows no comparable reduction.

These diagnostics are consistent with our intervention-location ablations. In U-Net backbones, Figure 3 shows that suppressing shallower skip/upsampling features can increase score deviations in hallucinated regions, whereas h-space suppression reduces them. For Transformer backbones, we shows that the 3/4-depth self-attention layer gives the best FID in Appendix. We further shows that randomly selected h-space suppression degrades FID, indicating that the gain comes from detector-selected intervention in deep low-noise latents in Appendix.

Table 1 shows that DUNE exhibits no artifact–fidelity trade-off in our evaluation: it improves both FID and HADM over the vanilla generator on all five tested backbones. On SDXL, every supported baseline worsens FID, whereas DUNE improves FID by 0.11 and gives the largest HADM drop ($0.227 \rightarrow 0.074$). Unsupported entries are reported explicitly because the corresponding official/default implementations are unavailable for those backbones: FreeU lacks official/default

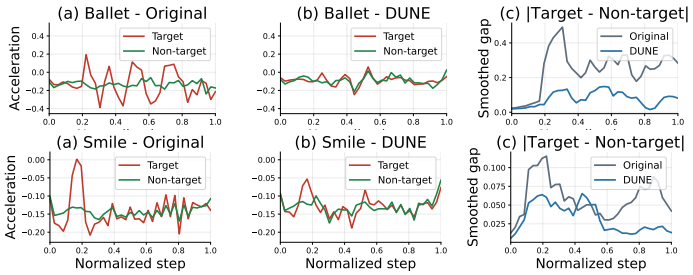

Fig. 9: Before/after score acceleration. Target denotes the detector-selected mask and non-target denotes its complement. Curves omit the final 10 denoising steps and are normalized only for visualization.

settings for Kandinsky 3, PixArt-Σ, and Hunyuan-DiT, while PAG lacks official/default settings for Kandinsky 3 and Hunyuan-DiT. DUNE also maintains high original-to-refined CLIP similarity, indicating low global semantic drift from the input generation. Overall, Table 1 shows that DUNE reduces localized artifacts while preserving the global content of the original sample.

Notably, ASCED, which edits the score trajectory directly, performs poorly on Transformer-based high-resolution text-to-image settings, introducing noisy artifacts (cf. Figure 7). This suggests that direct score correction can deviate the denoising path and that Transformers are sensitive to abrupt trajectory changes and channel-agnostic masking; DUNE's latent-level, backbone-aware suppression avoids these failure modes. Even though PAG achieves better HADM than original images, PAG shows overly saturated images with simplified edges, as reported in [1].

Finally, we summarize the practical settings of DUNE. We sweep the detection percentile p and suppression factor κ on LCM and Kandinsky 3, computing FID on 1,000 images for each pair $(p, \kappa) \in \{0.3, 0.5, 0.7, 0.9\} \times \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Across both backbones, $p = 0.9$ gives the best FID, so we fix $p = 0.9$ for all main experiments. The suppression factor κ is fixed once per backbone; no per-prompt tuning is used. For Transformer backbones, the target layer is fixed at 3/4 depth, following the depth ablation in Appendix.

5.3 Qualitative Evaluation

Figure 6 compares the generated images from the original generation process (left column) and the proposed DUNE (right column) across various diffusion models. In the original generations, diffusion models often struggle with fine structures (e.g., fingers, teeth) and may introduce extra parts, producing unrealistic complexity. DUNE suppresses these extraneous components and yields more coherent, detailed structures. When prompts are challenging or intentionally unusual (e.g., cat snake), DUNE effectively addresses hallucinations, removing artifacts while preserving semantic consistency with the original generation.

We provides qualitative comparisons with other baselines in Appendix. Even with the default hyperparameters reported in their original papers, FreeU and

PAG produce overly saturated images with simplified edges, losing the original semantic intent, as the authors mentioned [1, 16]. ASCED, which directly modifies the score trajectory, tends to inject noise that degrades local structure, particularly in high-resolution Transformer-based generations. In contrast, DUNE reduces artifacts without inducing semantic drift or over-saturation by operating on deep latents rather than the score output itself.

We also conduct a user study with 50 volunteers. For breadth, we include unconditional models (DDPM and VE) on CelebA-HQ, Bedroom, FFHQ, and Church, as well as high-resolution text-to-image models (SDXL and Kandinsky 3). Prompts and generated images are provided in Appendix. Users consistently favored DUNE-enhanced outputs, supporting the effectiveness of our approach in improving perceived quality and reducing hallucinations.

Finally, we conduct several additional experiments in Appendix. We selectively scale skip connections and upsampling blocks both upwards and downwards during the detail phase, contrasting with previous methods that predominantly amplify these components [16]. This enables precise control over brightness and tone, allowing both brighter and dimmer images.

6 Conclusion

In this work, we analyzed the functional roles of U-Net components throughout the diffusion process, clarifying where and when artifact-associated internal dynamics appear during diffusion sampling. Based on these findings, we introduced DUNE, a training-free, phase-aware refinement framework whose core detect–suppress mechanism identifies and corrects anomalous deep latents, yielding consistent quality gains across both U-Net and Transformer backbones. The framework additionally offers an optional detail phase for user-driven stylistic adjustments. We extend DUNE into Transformer-based methods and also achieve performance improvements. Extensive evaluations show that DUNE reduces artifacts while preserving semantic consistency.

Limitations Although we suggest some guideline of choosing hyperparameters in sensitivity analysis, our method requires minimal hyperparameter tuning (percentile threshold and scaling factor). Fully automatic threshold setting remains for future work.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation; No. RS-2024-00457882, AI Research Hub Project), and by the Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (No. RS-2025-25458052, Development of Core Technologies for Manufacturing Foundation Models).

References

1. Ahn, D., Cho, H., Min, J., Jang, W., Kim, J., Kim, S., Park, H.H., Jin, K.H., Kim, S.: Self-rectifying diffusion sampling with perturbed-attention guidance (2025), https://arxiv.org/abs/2403.17377
2. Aithal, S.K., Maini, P., Lipton, Z., Kolter, J.Z.: Understanding hallucinations in diffusion models through mode interpolation. Advances in Neural Information Processing Systems **37**, 134614–134644 (2024)
3. Arkhipkin, V., Vasilev, V., Filatov, A., Pavlov, I., Agafonova, J., Gerasimenko, N., Averchenkova, A., Mironova, E., Bukashkin, A., Kulikov, K., et al.: Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. arXiv preprint arXiv:2410.21061 (2024)
4. Cao, Y., Zhao, Z., Patras, I., Gong, S.: Temporal score analysis for understanding and correcting diffusion artifacts. arXiv preprint arXiv:2503.16218 (2025)
5. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation (2024), https://arxiv.org/abs/2403.04692
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
7. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)
8. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
9. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., Chen, D., He, J., Li, J., Li, W., Zhang, C., Quan, R., Lu, J., Huang, J., Yuan, X., Zheng, X., Li, Y., Zhang, J., Zhang, C., Chen, M., Liu, J., Fang, Z., Wang, W., Xue, J., Tao, Y., Zhu, J., Liu, K., Lin, S., Sun, Y., Li, Y., Wang, D., Chen, M., Hu, Z., Xiao, X., Chen, Y., Liu, Y., Liu, W., Wang, D., Yang, Y., Jiang, J., Lu, Q.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding (2024), https://arxiv.org/abs/2405.08748
10. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312
11. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference (2023), https://arxiv.org/abs/2310.04378
12. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023), https://arxiv.org/abs/2212.09748
13. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
14. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

16. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4733–4743 (2024)
17. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
18. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations (2021), `https://arxiv.org/abs/2011.13456`
19. Wang, K., Zhang, L., Zhang, J.: Detecting human artifacts from text-to-image models (2025), `https://arxiv.org/abs/2411.13842`
20. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications (2023)