DA-VAE: Plug-in Latent Compression for Diffusion via Detail Alignment — Supplementary Material

This supplementary material provides implementation details and additional analyses. In particular, it

- Sec. S1 summarizes the training and sampling hyperparameters used in all experiments;
- Sec. S2 describes how to instantiate DA-VAE on top of a pretrained VAE tokenizer, using SD3-VAE as a concrete example;
- Sec. S3 and Sec. S4 verify that the decoder and the diffusion backbone actually make use of the extra detail latent channels, rather than ignoring them;
- Sec. S5 provides a detailed comparison of DA-VAE against super-resolution post-processing baselines;
- Sec. S6 presents a frequency-domain analysis of the base and detail latents;
- Sec. S7 presents additional qualitative results for DA-VAE enhanced Stable Diffusion 3.5 Medium.

S1. Training and sampling hyperparameters

Tab. S1 lists the optimization and sampling configurations used in all our experiments.

For ImageNet class-to-image experiments with LightningDiT-XL, we largely follow the training recipe of Yao et al. [4], adjusting only the learning rate, batch size, and loss weights to accommodate our higher-compression DA-VAE latent space. DA-VAE is trained with AdamW and a relatively small KL weight, while λ_{align} is set to a moderate value to balance reconstruction and generation quality. For SD3.5-M, we use a smaller batch size and slightly different loss weights $(\lambda_L, \lambda_1, \lambda_{\text{adv}}, \lambda_{\text{KL}}, \lambda_{\text{align}})$ to stabilize high-resolution reconstruction. During DiT fine-tuning, the gradual loss scheduling down-weights the detail-latent loss for the first N_{warm} steps (10k for LightningDiT-XL; 5k for SD3.5-M), after which it is ramped up to full weight. We maintain an EMA of the DiT parameters with decay 0.999 throughout. For sampling, we use 250 diffusion steps with CFG scale 4.0 on ImageNet and 30 steps with guidance scale 2.5 for SD3.5-M.

Stage	Hyper-parameter	lightningDiT-XL [4], Class-to-image	SD3.5-M [3], Text-to-image
DA-VAE Training	learning rate	1e-4	1e-4
	batch size	128	16
	training steps	100K	10K
	optimizer	AdamW, betas=[0.5, 0.9]	AdamW, betas=[0.9, 0.999]
	loss weights $(\lambda_L, \lambda_1, \lambda_{\text{adv}}, \lambda_{KL}, \lambda_{\text{align}})$	(1.0, 1.0, 0.1, 1e-6, 0.5)	(1.0, 2.0, 0.1, 1e-7, 1.0)
DiT Fine-Tuning	learning rate	2e-4	1e-4
	Gradual loss scheduling steps	10K	5K
	batch size	640	128
	training steps	140K	10K
	optimizer	AdamW, betas=[0.9, 0.95]	AdamW, betas=[0.9, 0.999]
	EMA decay	0.999	0.999
Sampling for Generation	# sampling steps	250	30
	CFG / guidance scale	4.0	2.5
	CFG interval start	0.2	-
	timestep shift	0.3	-

Table S1. **Training and sampling hyperparameters for lightningDiT-XL and SD3.5-M.**

S2. DA-VAE architecture

Fig. S1 illustrates how we turn a SD3-VAE into DA-VAE. We *borrow* the overall encoder and decoder backbone architectures from SD3-VAE, but remove their original feature-to-latent and latent-to-feature heads and replace them with our own downsampling and upsampling blocks. The resulting encoder E_f and decoder D_f are therefore retrained as part of DA-VAE.

Concretely, the SD3-VAE encoder produces an intermediate feature map $F \in \mathbb{R}^{512\times H\times W}$. In the original SD3-VAE, a shallow head directly maps F to a latent $z_{\text{sd3}} \in \mathbb{R}^{16\times H\times W}$. In our design, we discard this head and instead attach a small downsampling module that further reduces the spatial resolution of F while keeping the channel dimension fixed (e.g., a stack of strided 3×3 conv blocks). This yields a more compressed detail latent $z_d \in \mathbb{R}^{16\times (H/s)\times (W/s)}$. We then concatenate z_d with the base latent z (the original SD3-VAE feature of the downsampled base image) to form the structured latent (z, z_d) used by our DiT.

The decoder side is modified symmetrically. Instead of feeding the original SD3-VAE latent z_{sd3} into a latent-to-feature stem, we concatenate our base and detail latents along the channel dimension and apply a lightweight upsampling block (e.g.,

pixel shuffle) that inverts the encoder's spatial downsampling. A 3×3 convolution then maps the upsampled latent back to a $512 \times H \times W$ feature map, which is passed through the SD3-VAE decoder backbone D_f for reconstruction.

In summary, DA-VAE keeps the deep convolutional backbone structure of SD3-VAE but replaces its shallow latent heads with our own downsampling/upsampling design, enabling a higher-compression latent space with an explicit separation between base and detail channels. All components, including the reused backbone blocks, are trained end-to-end under our DA-VAE objective.

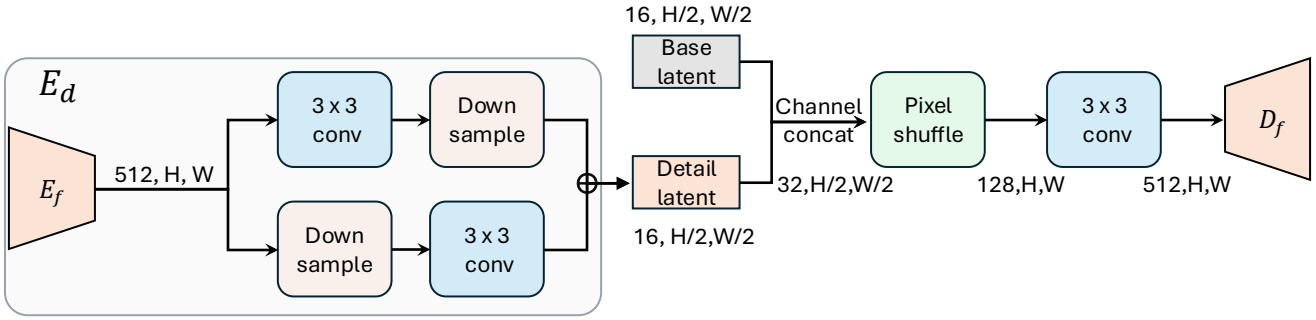
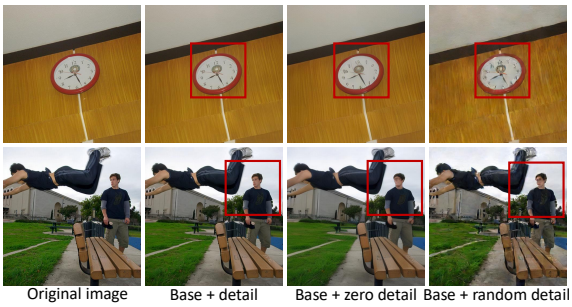


Figure S1. **DA-VAE architecture instantiated on SD3-VAE.** We reuse the convolutional encoder E_f and decoder D_f blocks from SD3-VAE, but remove its original feature-to-latent and latent-to-feature heads. Instead, a lightweight downsampling module maps the shared $512 \times H \times W$ feature map to a more compressed detail latent z_d and a parallel base latent z of the same shape, while a symmetric upsampling module concatenates (z, z_d), upsamples them back to $512 \times H \times W$, and feeds the result into the reused decoder backbone. This yields a higher-compression latent space with explicit base and detail channels, while keeping most of the VAE architecture intact.

S3. Decoder sensitivity to the detail latent

	Reconstruction (ImageNet val)			
Decoder variant	**rFID ↓**	**PSNR ↑**	**LPIPS ↓**	**SSIM ↑**
Full (base + detail)	**0.47**	**28.53**	**0.12**	**0.78**
Base + random detail	8.25	23.67	0.30	0.62
Base + zero detail	2.93	24.71	0.25	0.63

(a) Reconstruction metrics on the ImageNet validation set.



(b) Example reconstructions on ImageNet. Best for zoom-in view.

Figure S2. **Ablation on detail channels in the DA-VAE decoder on ImageNet.** (a) Reconstruction metrics for different decoder variants. (b) Visual examples showing that randomizing or zeroing the detail latent either destroys the image or removes fine-grained details such as faces and text. Please zoom in for best view.

We evaluate the sensitivity of the decoder to the detail latent on the ImageNet validation set. Starting from a trained DA-VAE, we fix the base latent z and modify the detail latent z_d in two ways: (i) we replace z_d with i.i.d. Gaussian noise $\mathcal{N}(0, I)$ (*Base + random detail*); and (ii) we set z_d to zero (*Base + zero detail*). The quantitative reconstruction metrics are summarized in Fig. S2a, and representative reconstructions are visualized in Fig. S2b.

Randomizing z_d leads to clearly invalid reconstructions with high rFID, low PSNR, and severe artifacts such as distorted faces and unreadable text, indicating that the decoder cannot simply ignore the detail channels. Zeroing z_d produces structurally plausible but over-smoothed images: edges become soft and fine textures disappear. In contrast, the full model using both z and z_d recovers both global structure and high-frequency details.

These observations confirm that the learned detail latent encodes semantically meaningful fine-grained information. Consequently, during fine-tuning the DiT must also learn to generate z_d correctly; otherwise the final high-resolution samples would lack sharp details even if the base latent is well modeled.

S4. Training dynamics of SD3.5-M fine-tuning

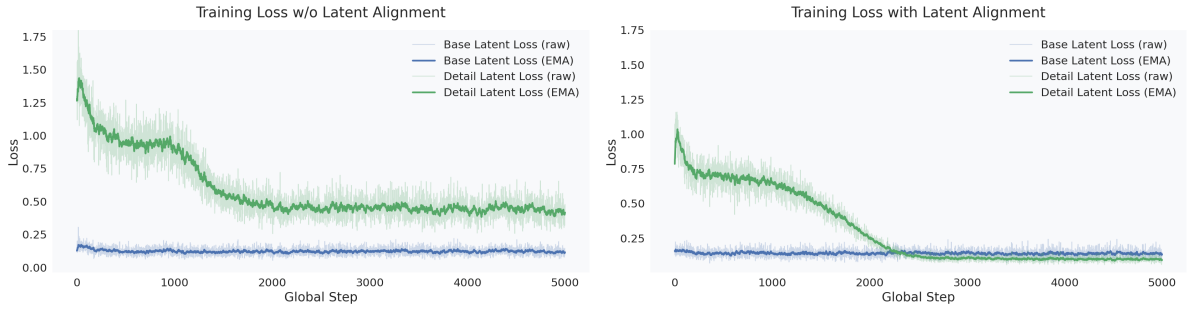


Figure S3. **Training loss curves for SD3.5-M fine-tuning with and without latent alignment.** We plot the unweighted diffusion loss on the base latent (blue) and the detail latent (green), showing both the raw loss (faint) and its EMA (solid). *Left:* without alignment, the detail-latent loss decreases slowly and stays significantly higher than the base-latent loss. *Right:* with alignment, optimization is more stable and the detail-latent loss eventually falls below the base-latent loss, indicating that the DiT has learned a well-structured distribution over the extra detail channels.

Fig. S3 visualizes the optimization behaviour when fine-tuning SD3.5-M from 512×512 to 1024×1024 resolution with our DA-VAE. We plot the *unweighted* diffusion loss on the base latent and on the detail latent, i.e., the true per-token MSE before applying the scheduling weight $w(n)$ described in the main paper. For each branch we show both the raw loss and its exponential moving average (EMA).

Two trends are worth noting. First, the base-latent loss stays relatively low throughout training, while the detail-latent loss starts much higher and gradually decreases—fine-tuning primarily teaches the model to predict the new detail channels, leveraging the well-trained prior in the base latent. Second, comparing the two plots shows the effect of latent alignment: without alignment the detail-latent loss plateaus at a high value, whereas with alignment it decreases steadily and eventually falls *below* the base-latent loss, confirming that aligned latents form a more learnable distribution that the DiT can effectively exploit.

S5. Comparison with Super-Resolution Post-Processing

A natural question is whether one could achieve similar results by first generating a low-resolution image and then applying a learned super-resolution (SR) model. We argue that DA-VAE is superior in two key aspects.

Joint modeling vs. conditional upsampling. A two-stage SR pipeline factorizes the high-resolution distribution as $P(x_{\text{high}}) \approx P(x_{\text{low}})\, P(x_{\text{high}} \mid x_{\text{low}})$. Once the 512px model has sampled x_{low}, the global composition (e.g., layout, object counts) is largely fixed; the SR model can only refine local appearance and cannot reliably correct missing objects or compositional errors. In contrast, DA-VAE models the joint distribution $P(x_{\text{high}})$ natively, yielding better structural fidelity and text alignment, as reflected by the higher GenEval-*Count* and CLIP-Score in Tab. S2.

Inference latency. SR requires a cascaded second-stage inference pass, adding non-trivial latency (e.g., SeedVR2 roughly doubles total inference time compared to the 512px baseline). DA-VAE generates high-resolution images in a single forward pass, matching the throughput of the 512px baseline.

Tab. S2 summarizes quantitative results and Fig. S4 shows qualitative examples. In the counting example, 512px generation produces an incorrect count that SR methods cannot fix, whereas DA-VAE generates the correct number of objects directly. In the scene example, SR sharpens local textures but preserves a simplified layout, while DA-VAE produces richer global structure.

Table S2. Comparison of DA-VAE with super-resolution post-processing baselines. All methods use the same 512×512 SD3.5-M backbone. Throughput in img/s on a single H100.

Method	FID↓	GenEval-*Count*↑	CLIP-Score↑	Throughput↑
512 + Bilinear	12.04	0.55	30.17	1.03
512 + SeedVR2	**10.48**	0.55	30.19	0.45
512 + FMBoost	11.02	0.55	30.16	0.52
DA-VAE (Ours)	10.91	**0.60**	**31.91**	**1.03**

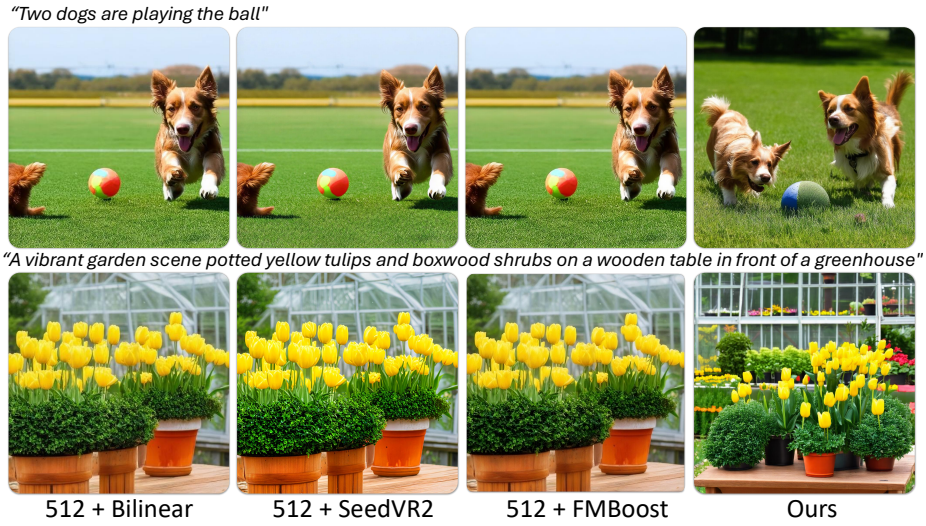


Figure S4. Qualitative comparison of DA-VAE vs. SR baselines. *Top:* a counting prompt where 512px generation gets the count wrong and SR cannot fix it. *Bottom:* a scene prompt where SR only sharpens local textures while DA-VAE produces richer global structure.

S6. Frequency-Domain Analysis of Base and Detail Latents

To verify that the detail latent $\mathbf{z}_d$ encodes genuinely complementary high-frequency information—rather than simply duplicating the base latent $\mathbf{z}$—we compute the radial power spectrum of each latent channel and average across channels and images from the ImageNet validation set.

Fig. S5 plots the resulting spectral energy as a function of spatial frequency. The base latent $\mathbf{z}$ concentrates energy at low frequencies, consistent with its role in capturing global structure, while $\mathbf{z}_d$ exhibits substantially higher energy in the mid-to-high frequency bands, confirming that it captures fine textures and edges absent from $\mathbf{z}$. This is consistent with Sec. S3: zeroing $\mathbf{z}_d$ produces over-smoothed reconstructions precisely because this high-frequency content is lost. Despite this complementarity, the alignment loss prevents $\mathbf{z}_d$ from collapsing into a trivial copy of $\mathbf{z}$: the two latents differ in both spectral content and spatial statistics, making them jointly necessary for full-resolution reconstruction.

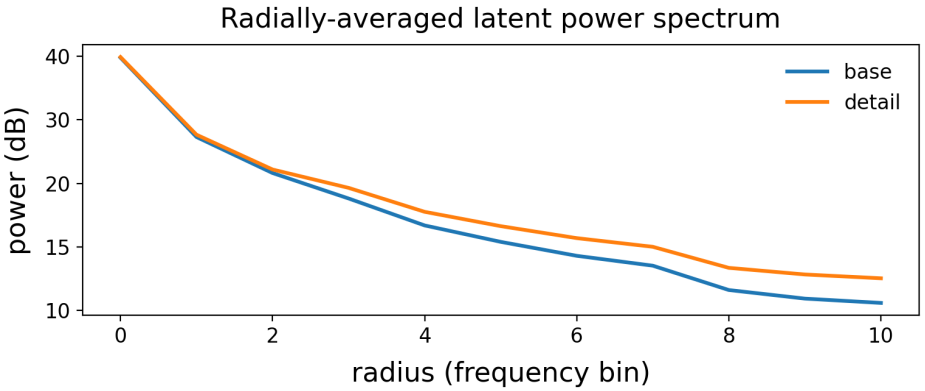


Figure S5. Radial power spectrum of the base latent $\mathbf{z}$ and detail latent $\mathbf{z}_d$ averaged over ImageNet validation images. The detail latent carries substantially more high-frequency energy, confirming that it encodes complementary fine-grained information rather than duplicating the base.

S7. Additional qualitative results

To further demonstrate the effectiveness of our method, Fig. S6 presents additional qualitative results of DA-VAE–enhanced Stable Diffusion 3.5 Medium (SD3.5-M) for text-to-image generation. To improve realism, we further fine-tune SD3.5-M with our model for 5K steps on 500K images generated by Flux [1] using prompts collected by [2].

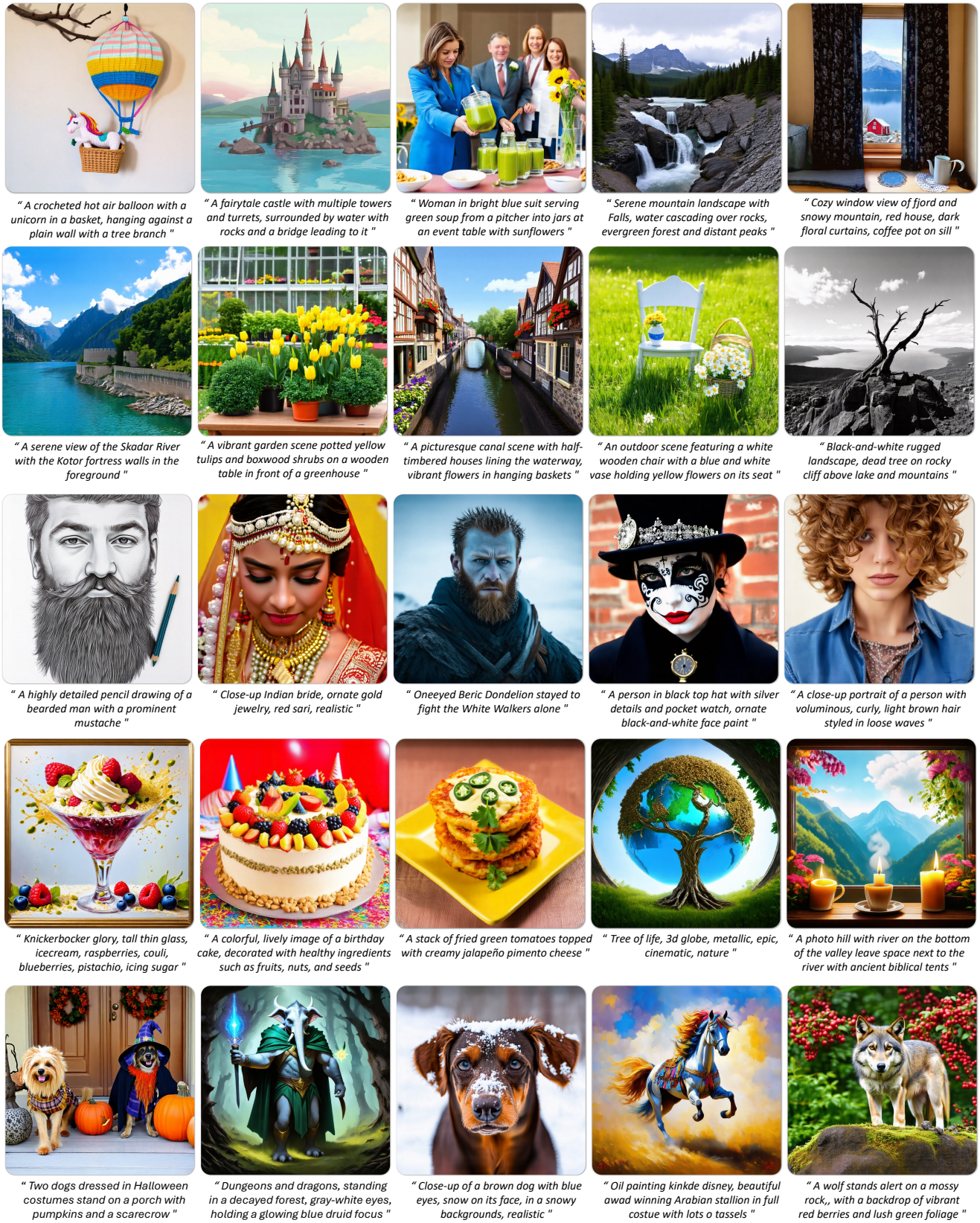


Figure S6. **Genearated examples by our DA-VAE enhanced SD3.5-M.** Please zoom in for best view.

References

[1] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. Accessed 2025-11-20. 4

[2] Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. pi-flow: Policy-based few-step generation via imitation distillation. *arXiv preprint arXiv:2510.14974*, 2025. 4

[3] Stability AI. Sd3.5. https://github.com/Stability-AI/sd3.5, 2024. 1

[4] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. Generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025. 1