SCALE-WISE DISTILLATION OF DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent diffusion distillation methods have achieved remarkable progress, enabling high-quality $\sim$4-step sampling for large-scale text-conditional image and video diffusion models (DMs). However, further reducing the number of sampling steps becomes more and more challenging, suggesting that efficiency gains may be better mined along other model axes. Motivated by this perspective, we introduce SwD, a scale-wise diffusion distillation framework that equips few-step models with progressive generation, avoiding redundant computations at intermediate diffusion timesteps. Beyond efficiency, SwD enriches the family of distribution matching distillation approaches by introducing a simple distillation objective based on kernel Maximum Mean Discrepancy (MMD). This loss significantly improves the convergence of existing distillation methods and performs surprisingly well in isolation, offering a competitive baseline for diffusion distillation. Applied to state-of-the-art text-to-image/video diffusion models, SwD approaches the sampling speed of two full-resolution steps and largely outperforms alternatives under the same compute budget, as evidenced by automatic metrics and human preference studies.

1 INTRODUCTION

Diffusion models (DMs) are the leading paradigm for visual generative modeling (Black Forest Labs, 2024; Wan et al., 2025; Esser et al., 2024; Polyak et al., 2024a; Zhou et al., 2025a). Since generating high-resolution images or videos (e.g., 1024×1024) becomes computationally prohibitive when operating directly in pixel space, state-of-the-art DMs leverage lower-resolution VAE (Kingma et al., 2013) latent spaces. However, the VAEs used in latent DMs typically employ an $8\times$ scaling factor, meaning the latent space still remains high-dimensional. Given the slow sequential diffusion process requiring $20-50$ steps, generation speed is a significant bottleneck, especially for recent large-scale models with >8 billion parameters (Sauer et al., 2024; Black Forest Labs, 2024; Cai et al., 2025; Polyak et al., 2024a; Wan et al., 2025).

Previous works have made substantial efforts in DM acceleration from different perspectives (Lu et al., 2022; Song et al., 2020a; Wimbauer et al., 2024; Li et al., 2023; Yin et al., 2024b). One of the most successful directions is distilling DMs into few-step generators (Song et al., 2023; Kim et al., 2024; Sauer et al., 2023; Yin et al., 2024b), aiming to achieve the inference speeds comparable to single-step generative models, such as GANs (Goodfellow et al., 2014). Notably, these approaches generally focus on reducing the number of sampling steps while freezing other promising degrees of freedom, such as model architectures or data dimensionality.

Recently, Rissanen et al. (2023); Dieleman (2024) has noticed the coarse-to-fine nature of the image diffusion generative process, drawing parallels to the implicit form of spectral autoregression. Specifically, low frequency image information is modeled at high noise levels, while higher frequencies are progressively produced over the reverse diffusion process. This observation establishes a connection to the next-scale prediction models (Tian et al., 2024; Voronov et al., 2024; Han et al., 2024), which predict higher frequency details at each step via upscaling. Despite this insight, state-of-the-art few-step DMs still operate within a fixed dimensionality throughout the diffusion process, highlighting an underexplored direction for improving their efficiency.

Contribution. Since most state-of-the-art DMs belong to the latent diffusion family (Rombach et al., 2021), firstly, we need to address whether the spectral autoregression perspective also applies to latent representations. In this work, we conduct a spectral analysis of existing VAE latent spaces

and also extend it to the video domain. Our findings confirm that both spatial and temporal latent resolutions implicitly increase over the diffusion process, similarly to the natural images. This suggests that latent DMs can avoid redundant computations at intermediate noisy timesteps, where high frequencies are largely suppressed.

Motivated by this observation, we introduce a ***S**cale-**w**ise **D**istillation* (SwD) framework, which transforms an arbitrary pretrained DM into a single few-step model that progressively increases spatial and temporal sample resolutions at each generation step. SwD integrates seamlessly with existing distribution matching distillation approaches (Sauer et al., 2023; Yin et al., 2024b) and leverages their few-step sampling algorithms, which appear naturally aligned with progressive generation.

In addition to the scale-wise distillation framework, we present a simple yet surprisingly effective diffusion distillation objective that minimizes kernel Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) in the feature space of a pretrained DM. The proposed objective complements state-of-the-art distillation methods and achieves strong performance even in isolation, establishing a competitive baseline for DM distillation. Importantly, it requires no additional trainable models, making it computationally efficient and easy to combine with existing distillation pipelines.

We apply SwD to state-of-the-art text-to-image and video DMs and show that our models compete or even outperform their teachers being more than $10\times$ faster. Compared to full-resolution few-step models, SwD significantly surpasses them under a similar computational budget. For the same number of sampling steps, SwD provides $\sim 2\times$ speedup in text-to-image generation and $\sim 3\times$ in text-to-video generation, without compromising quality.

2 Related Work

Diffusion distillation into few-step models. Diffusion distillation methods aim to reduce generation steps to $1-4$ while maintaining teacher model performance. These methods can be largely grouped into two categories: *teacher-following* methods (Meng et al., 2023; Song et al., 2023; Luo et al., 2023a; Huang et al., 2023; Song & Dhariwal, 2024) and *distribution matching* (Yin et al., 2024b;a; Sauer et al., 2023; 2024; Luo et al., 2023b; Zhou et al., 2024b;a).

Teacher-following methods approximate the teacher's noise-to-data mapping by integrating the diffusion ODE in fewer steps than numerical solvers (Song et al., 2020a; Lu et al., 2022). Distribution matching methods relax the teacher-following constraint, focusing instead on aligning student and teacher distributions without requiring exact noise-to-data mapping. State-of-the-art approaches, such as DMD2 (Yin et al., 2024a) and ADD (Sauer et al., 2023; 2024), demonstrate strong generative performance in ~ 4 steps. However, they still exhibit noticeable quality degradation at $1-2$ step generation, leaving room for further improvement. Recently, DMD has been successfully adopted for video diffusion models (Yin et al., 2025; Huang et al., 2025).

Progressive generation with DMs. The idea of progressively increasing resolution during diffusion generation was initially exploited in hierarchical or cascaded DMs (Ho et al., 2021; Saharia et al., 2022; Ramesh et al., 2022; Kastryulin et al., 2024; Gu et al., 2023b), which are strong competitors to latent DMs (Rombach et al., 2021) for high-resolution generation. Cascaded DMs consist of multiple DMs operating at different resolutions, where each model performs a diffusion sampling from scratch, conditioned on the lower-resolution sample. To bridge progressive generation with diffusion processes, several works (Gu et al., 2023a; Teng et al., 2023; Atzmon et al., 2024; Jin et al., 2025) have presented multi-stage pipelines, where DMs are trained to smoothly transition to higher-resolution noisy samples during diffusion sampling. SwD follows up this line of research by proposing a framework that readily integrates into existing diffusion distillation procedures and adapts arbitrary pretrained DMs into unified progressive few-step models.

Maximum Mean Discrepancy in generative modeling. Maximum Mean Discrepancy (MMD) is a metric between two distributions P and Q, widely explored in early GAN works (Bińkowski et al., 2018; Wang et al., 2018; Dziugaite et al., 2015; Bellemare et al., 2017; Sutherland et al., 2016).

Given a positive-definite *kernel function* $k(\mathbf{x}, \mathbf{y})$, MMD can be defined as

$$\mathrm{MMD}^2(P, Q) = \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim P}[k(\mathbf{x}, \mathbf{x}')] + \mathbb{E}_{\mathbf{y},\mathbf{y}'\sim Q}[k(\mathbf{y}, \mathbf{y}')] - 2\,\mathbb{E}_{\mathbf{x}\sim P,\,\mathbf{y}\sim Q}[k(\mathbf{x}, \mathbf{y})]\,, \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote samples from the generated and target distributions, respectively.

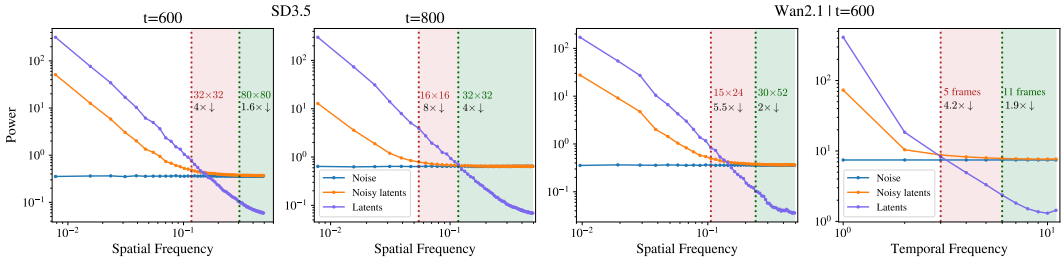

Figure 1: **Spectral analysis** of SD3.5 VAE latents (128×128) (*Left*) and Wan2.1 (21×60×104) for spatial and temporal dimensions (*Right*). Vertical lines mark the frequency boundaries for which the frequency components to the right are not present in lower resolution latents. Noise masks high frequencies, suggesting that latent DMs can operate at lower latent resolutions for high noise levels. Green area indicates the allowed latent resolution at corresponding timestep, while Red area shows that further resolution reduction would lead to noticeable information loss.

Generative Moment Matching Networks (GMMNs) (Li et al., 2015) employ MMD with a fixed Gaussian kernel (RBF) directly in data space. GANs, in contrast, typically consider learnable kernels, designed as the composition of a discriminator with a fixed kernel.

In diffusion modeling, MMD has been explored for DM training (Bortoli et al., 2025) or finetuning (Aiello et al., 2023). DMMD (Galashov et al., 2025) employs noise-adapted discriminators for MMD gradient flows (Arbel et al., 2019). Recently, IMM (Zhou et al., 2025b) leveraged MMD for consistency distillation (Song et al., 2023), computing the MMD with a fixed kernel between raw generator predictions at different timesteps. In our work, we adopt MMD between student and teacher distributions in the feature space of a pretrained DM, yielding a powerful and effective distribution matching objective.

3 Latent Space Spectral Analysis

Rissanen et al. (2023) and Dieleman (2024) showed that, in pixel space, diffusion models approximate spectral autoregression for natural images. Since state-of-the-art text-conditional diffusion models operate on VAE latent representations (Rombach et al., 2021), we first investigate this spectral perspective for various latent spaces, including temporal dimension.

Following Dieleman (2024), we evaluate *radially averaged power spectral density* (RAPSD), i.e., the averaged spectra power across different spatial frequency components, and its one-dimensional analogue for temporal frequencies.

We examine the latent spaces of image and video diffusion models, specifically Stable Diffusion 3.5 (SD3.5) (Esser et al., 2024) and Wan2.1 (Wan et al., 2025). The SD3.5 VAE maps 3×1024×1024 images into 16×128×128 latents, while the Wan2.1 VAE encodes 81×3×480×832 video inputs into 21×16×60×104 latents. Both models use a flow-matching process (Lipman et al., 2023).

Figure 1 shows the RAPSD of Gaussian noise (blue), clean latents (purple) and noisy latents (orange) at different timesteps. Figure 1 (Left) provides the results for SD3.5 VAE latents. Figure 1 (Right) shows RAPSD across both spatial and temporal frequencies of Wan2.1 latents. Vertical lines indicate frequency boundaries: the components to the right correspond to high frequencies absent at lower resolutions, while those to the left align with the full latent resolution (128×128).

Additional results, including more timesteps and SDXL (Podell et al., 2024) latents under a variance-preserving diffusion process (Ho et al., 2020; Song et al., 2020b), are provided in Appendix F.

Observations. First, we note that the latent frequency spectrum approximately follows a power law, similar to natural images (van der Arjen Schaaf & van Johannes Hateren, 1996). In contrast, however, highest frequency components in latent space exhibit slightly greater magnitude. We attribute this to the VAE regularization terms, which may cause "clean" latents to appear slightly noisy.

We also observe that the noising process progressively filters out high frequencies, thereby determining the safe downsampling range without noticeable information loss. Figure 1 (Left) shows that at t=800, noise masks high frequency components emerging at resolutions above 32×32. This

allows for $4\times$ downsampling of 128×128 latents (green area). On the other hand, $8\times$ downsampling would corrupt the data signal (red area), as the noise does not fully suppress those frequencies.

A similar effect is observed along the temporal dimension in Wan2.1 latents, see Figure 1 (Right). At t=600, the effective signal can be represented with $\sim$11 latent frames instead of the original 21.

Practical implication. Based on this analysis, we suppose that latent diffusion models may operate at lower resolution at high noise levels without losing the data signal. In other words, modeling high frequencies at timestep t is unnecessary if those frequencies are already masked at that noise level. Note that this holds true for both spatial and temporal axes for video DMs. We summarize this conclusion as follows:

> Diffusion process allows lower-resolution modeling at high noise levels in both spatial and temporal dimensions.

4 METHOD

This section introduces a *scale-wise distillation*, SwD, framework for diffusion models. We begin by describing the SWD pipeline, highlighting its key features and challenges. Then, we present our distillation objective based on Maximum Mean Discrepancy (MMD).

4.1 SCALE-WISE DISTILLATION OF DMS

The core design principle of SWD is to unify multi-scale generation within a *single distilled model* and *single diffusion process*, in contrast to cascaded approaches. To this end, we define a few-step *timestep schedule*, $[t_1, \dots, t_N]$, and pair each diffusion timestep t_i with a latent resolution s_i from a non-decreasing *scale schedule*, $[s_1, \dots, s_N]$. Therefore, starting the generation with Gaussian noise at the lowest scale, s_1, the resolution of intermediate noisy latents $\mathbf{x}_{t_i}$ is progressively increased over sampling steps.

Upsampling strategy. Before discussing the method details, an important question needs to be addressed: *how to upsample $\mathbf{x}_{t_i}$ to obtain faithful noisy latents?* A naive approach would be to directly upscale $\mathbf{x}_{t_i}$. However, we find it essential to first upscale a $\hat{\mathbf{x}}_0$ prediction and then noise it according to the forward diffusion process. We believe noise injection mitigates the upscaling artifacts and thus ensures closer alignment with the distribution of true noisy latents.

Configuration	$t=400$	$t=600$	$t=800$
A $\mathbf{x}_0 \xrightarrow{\text{noise}} \mathbf{x}_t$	9.2	9.8	12.9
B $\mathbf{x}_0^{\text{down}} \xrightarrow{\text{upscale}} \mathbf{x}_0 \xrightarrow{\text{noise}} \mathbf{x}_t$	32.4	17.3	13.0
C $\mathbf{x}_0^{\text{down}} \xrightarrow{\text{noise}} \mathbf{x}_t^{\text{down}} \xrightarrow{\text{upscale}} \mathbf{x}_t$	129.7	235.0	340.2

Table 1: Comparison of noisy latent upscaling strategies (**B**, **C**) for $64 \rightarrow 128$ in terms of generation quality (FID-5K) against the real noisy latents (**A**). Upscaling $\mathbf{x}_0^{\text{down}}$ before noise injection (**B**) aligns better with full-resolution noisy latents.

To validate this intuition, we generate images with Stable Diffusion 3.5 (Esser et al., 2024) from intermediate noisy latents, $\mathbf{x}_t$, obtained with different upscaling strategies. Specifically, given a full-resolution (128×128) real image latent, $\mathbf{x}_0$, and its downscaled version (64×64), $\mathbf{x}_0^{\text{down}}$, we consider the reference setting (**A**), where noise is added to full-resolution $\mathbf{x}_0$, and two upscaling strategies: (**B**) first upscale $\mathbf{x}_0^{\text{down}}$ and then inject noise; (**C**) first inject noise to $\mathbf{x}_0^{\text{down}}$ and then upscale.

As shown in Table 1, strategy (**B**) substantially outperforms (**C**), and the upscaling artifacts diminish at higher noise levels, e.g., at t=800, the performance gap is negligible.

To summarize, at each timestep t, SWD transitions to a higher resolution by upscaling the $\hat{\mathbf{x}}_0$ prediction using *bicubic interpolation* for spatial dimensions and *adjacent frame blending* for the temporal ones, followed by noise injection to produce $\hat{\mathbf{x}}_t$.

Sampling. This upsampling strategy favors the *stochastic multistep sampling*, widely used in state-of-the-art diffusion distillation approaches (Sauer et al., 2023; Yin et al., 2024a; Luo et al., 2023a; Sauer et al., 2024). SWD adapts it for multi-scale generation, i.e., given the intermediate noisy latent $\hat{\mathbf{x}}_{t_{i-1}}$ at resolution s_{i-1}, the model produces a prediction $\hat{\mathbf{x}}_0^{i-1}$. To proceed to the next timestep t_i,

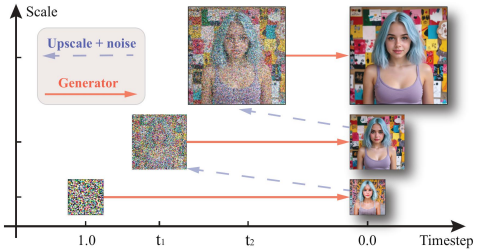

Figure 2: **SwD sampling.** The distilled model starts sampling from noise at the low resolution s_1 and gradually increases it over generation steps. At each step, the previous denoised prediction at the scale s_{i-1} is upscaled and noised according to the *timestep schedule*, t_i. Then, the generator predicts a clean image at the current resolution s_i.

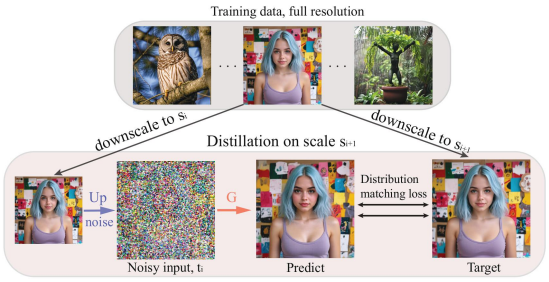

Figure 3: **SwD training step. i)** Sample a pair of adjacent resolutions $[s_i, s_{i+1}]$ from *scale schedule*. **ii)** Downscale the training images to s_i and s_{i+1}. **iii)** The lower scale versions are upsampled and noised to a timestep t_i with the forward process. **iv)** Given the noised images, the model G predicts clean data at target scale s_{i+1}. **v)** Distribution matching loss is calculated between predicted and target images.

$\hat{\mathbf{x}}_0^{i-1}$ is upscaled to s_i and noised according to the forward diffusion process, resulting in the less noisy latent $\hat{\mathbf{x}}_{t_i}$. Then, the model predicts next $\hat{\mathbf{x}}_0^i$. Figure 2 illustrates this sampling process.

Though this procedure can be directly applied to already pretrained distilled models, in practice, we notice that noise injection alone is not sufficient to completely mitigate upscaling artifacts or requires using very high noise levels, reducing the effectiveness of such models. Therefore, we aim to train a few-step generator that also serves as a robust upscaler.

Training. We train a single model across multiple resolutions, iterating over pairs of adjacent scales $[s_i, s_{i+1}]$ from the *scale schedule*. At each training step, we sample a batch of full-resolution images or videos, downscale them to the source and target resolutions in pixel space, according to the s_i and s_{i+1} scales, and then encode them into the VAE latent space. Notably, we find that downscaling in pixel space before the VAE encoding largely outperforms latent downscaling in our experiments.

Next, we upsample the lower resolution latents from s_i to s_{i+1} and add noise according to the *timestep schedule*, t_i. The noised latents are then fed into the scale-wise generator, which predicts $\hat{\mathbf{x}}_0$ at the target scale s_{i+1}.

Finally, we calculate a distillation loss between the predicted and target latents at s_{i+1}. In our work, we use distribution matching, motivated by ADD (Sauer et al., 2023; 2024) and DMD (Yin et al., 2024b;a), achieving state-of-the-art performance in diffusion distillation.

The schematic illustration of this training procedure is provided in Figure 3. Further implementation details and discussions are in Appendix A.

Discussion on the timestep and scale schedules. Following Section 3, the emergence of higher-frequency components at lower noise levels can provide useful initial assumptions for designing the schedules. However, since the analysis provides only averaged results and does not account for upscaling artifacts, the schedules ultimately remain hyperparameters.

In practice, we find it beneficial to use higher timesteps than in the default schedules, aligning with the intuition that noise injection mitigates upscaling artifacts.

4.2 Diffusion Distillation with Maximum Mean Discrepancy

In addition to the proposed scale-wise distillation framework, we extend the family of distribution matching distillation methods with a MMD loss, calculated on the intermediate features of the pre-trained DMs. Below, we discuss the loss computation for the transformer-based DMs (Peebles & Xie, 2022) as the most widely used architecture in state-of-the-art DMs, while keeping in mind that the loss is applicable to arbitrary architectures.

First, we leverage the ability of DMs to operate at different noise levels, enabling the extraction of structured signal at high noise levels and finer-grained feedback at low ones. Accordingly, before feature extraction, we noise both generated and target samples within a predefined timestep interval.

Then, we extract feature maps $\mathbf{F} \in \mathbb{R}^{N \times L \times C}$ from the middle transformer block of the teacher DM for generated and target images/videos and denote them as $\mathbf{F}^{\text{fake}}$ and $\mathbf{F}^{\text{real}}$, respectively. N is a batch size, L is a number of spatial tokens, and C is a hidden dimension of the transformer.

For MMD computation, we consider two kernels: linear ($k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$) and the radial basis function (RBF) (Chang et al., 2010). The former aligns feature distribution means, while the latter also matches all higher-order moments. In our experiments, both kernels perform similarly, so we simplify $\mathcal{L}_{\text{MMD}}$ using the linear kernel, i.e., calculate MSE between spatial token means *per image*:

$$\mathcal{L}_{\text{MMD}} = \sum_{n=1}^{N} \left\| \frac{1}{L} \sum_{l=1}^{L} \mathbf{F}^{\text{real}}_{n,l,\cdot} - \frac{1}{L} \sum_{l=1}^{L} \mathbf{F}^{\text{fake}}_{n,l,\cdot} \right\|^2 . \tag{2}$$

Note that the feature means computed across the entire batch rather than per image tend to mitigate condition-specific information, resulting in lower text relevance in our experiments.

Discussion. $\mathcal{L}_{\text{MMD}}$ with a linear kernel can be considered as a diffusion distillation adaptation of the feature matching loss, proposed for improved GAN training (Salimans et al., 2016). To our knowledge, such losses have not been explored in the context of diffusion distillation. The notable differences are: i) $\mathcal{L}_{\text{MMD}}$ leverages a pretrained DM instead of a learnable discriminator; ii) it uses the feedback from different noise levels; iii) the feature means are computed per image rather than across the entire batch.

Overall objective. We incorporate $\mathcal{L}_{\text{MMD}}$ as an additional loss to $\mathcal{L}_{\text{DMD}}$ and $\mathcal{L}_{\text{GAN}}$ in our scale-wise framework: $\mathcal{L}_{\text{SwD}} = \mathcal{L}_{\text{MMD}} + \alpha \cdot \mathcal{L}_{\text{DMD}} + \beta \cdot \mathcal{L}_{\text{GAN}}$. Interestingly, despite its simplicity, $\mathcal{L}_{\text{MMD}}$ proves to be a highly competitive standalone distillation objective.

5 Experiments

Models. We validate our approach in text-to-image generation by distilling SD3.5 Medium, SD3.5 Large (Esser et al., 2024) and FLUX.1-dev (Black Forest Labs, 2024). We also apply SwD to the recent text-to-video model, Wan2.1-1.3B (Wan et al., 2025).

Data. To remain in the isolated distillation setting and avoid biases from external data, we train all models exclusively on synthetic data generated by their teacher, rather than on real data.We note that this step does not pose a bottleneck for training, as the distillation process itself converges relatively fast ($\sim$3K iterations) and requires significantly less data than the DM training. The synthetic data generation settings for each model are provided in Appendix A.

Metrics. For text-to-image models, we use 30K text prompts from the COCO2014 and MJHQ sets (Lin et al., 2015; Li et al., 2024) and evaluate the automatic metrics: FID (Heusel et al., 2017), HPSv3 (Ma et al., 2025), ImageReward (IR) (Xu et al., 2023), and PickScore (PS) (Kirstain et al., 2023) and GenEval (Ghosh et al., 2023). Note that FID was shown to correlate poorly with human perception (Kirstain et al., 2023) for text-to-image assessment but we report it here for completeness.

Also, we conduct a user preference study via side-by-side comparisons evaluated by professional assessors. We select 128 text prompts from the PartiPrompts dataset (Yu et al., 2022), following (Yin et al., 2024a; Sauer et al., 2024), and generate 2 images per prompt. More details are in Appendix H.

For T2V models, we evaluate VBench-2.0 (Zheng et al., 2025), and VisionReward (Xu et al., 2024) and VideoReward (Liu et al., 2025) on 1003 prompts from MovieGenBench (Polyak et al., 2024b).

Setup. In our main experiments, we distill the models to 4 or 6 steps. For text-to-image models, the scale schedules begin at image resolutions of 256×256 or 512×512 and progress to 1024×1024. For text-to-video, we start with $21\times160\times272$ and achieve the $81\times480\times832$ resolution. We chose such starting points as lower resolutions provide only marginal speed improvements. The exact timestep and scale schedules for each model are in Appendix D.

Baselines. For text-to-image, we mainly compare with the teacher models and their publicly available distilled versions, e.g., SD3.5-Turbo (Sauer et al., 2024), FLUX-Schnell (Black Forest Labs, 2024). Also, we evaluate other fast state-of-the-art models, such as distilled SDXL models (DMD2-SDXL (Yin et al., 2024a), SDXL-Turbo (Sauer et al., 2023)) and next-scale prediction models (Switti (Voronov et al., 2024) and Infinity (Han et al., 2024)). For the text-to-video task, we compare with the teacher model (Wan2.1-1.3B (Wan et al., 2025)).

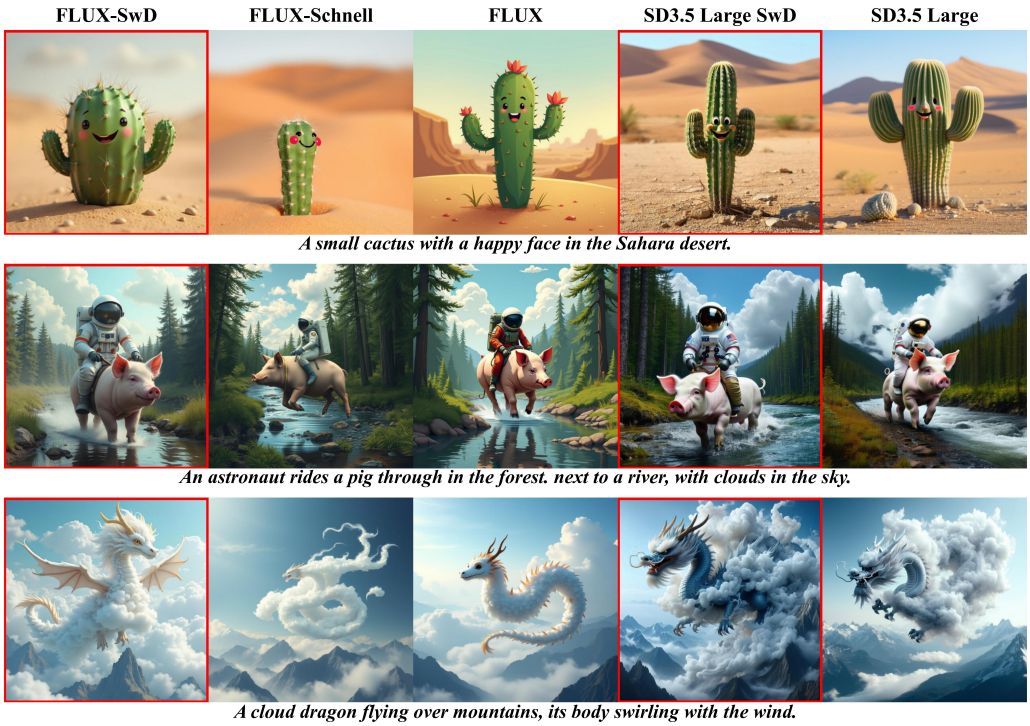

Figure 4: Qualitative results of FLUX-SwD and SD3.5 Large SwD. More examples are in Figure 15.

Figure 5: Qualitative results of Wan2.1-SwD. More examples are in Figure 14.

5.1 Main results

Text-to-image. Table 3 and Figure 6 present the comparisons of SwD with the baselines in terms of generation quality and speed. The results are split into subsections denoting different model sizes.

We find that SwD models achieve the best performance in terms of PS, HPSv3, IR and GenEval within their model families and outperform other models in most cases.

According to the human study, SwD outperforms most other models, including the more expensive teachers and their distilled variants, in terms of *image complexity* and *image aesthetics*, while maintaining comparable levels of *text relevance* and *defects*. Qualitative comparisons are presented in Figure 4, with additional results in Figure 18 and Figure 15.

Model	Latency, s/video	Vision Reward ↑	Video Reward ↑	VBench2 Overall ↑
Wan 2.1	137	0.038	5.43	51.59
Spatial SwD	2.1	**0.063**	5.92	53.39
SwD	**1.8**	**0.063**	**6.00**	**53.50**

Table 2: Comparison of 4-step SwD variants with the 50-step teacher model, Wan 2.1.

Text-to-video. The results in Table 2 show that SwD achieves slightly better performance than the teacher model, while being 72× faster. Visual examples are provided in Figures 5 and 14.

Also, we find that SwD, when applied across both temporal and spatial dimensions, yields results similar to the spatial-only variant.

Model	Latency, s/image	Model size, B	PS ↑	HPSv3 ↑	IR ↑	FID ↓	PS ↑	HPSv3 ↑	IR ↑	FID ↓	GenEval ↑
			COCO 30K				MJHQ 30K				
Switti	0.44	2.5	22.6	11.1	0.98	20.0	21.6	9.8	0.84	8.9	0.62
Infinity	0.80	2.0	22.7	11.8	0.94	28.1	21.5	10.5	0.98	12.9	0.69
SDXL	1.72	2.6	22.4	8.9	0.77	14.2	21.5	9.0	0.78	8.4	0.55
SDXL-Turbo	0.20	2.6	22.6	10.0	0.83	17.5	21.3	9.6	0.84	15.4	0.55
SDXL-DMD2	0.20	2.6	**22.8**	**12.0**	0.87	**14.1**	21.6	10.1	0.86	**8.3**	0.58
SD3.5-M	4.8	2.0	22.4	10.2	1.00	16.3	21.6	9.9	0.97	9.5	0.69
SD3.5-M-Turbo	0.96	2.0	22.2	9.6	0.83	17.6	21.3	9.3	0.74	13.6	0.59
SD3.5-M-SwD	**0.19**	2.0	**22.8**	11.7	**1.12**	23.1	**21.8**	**10.7**	**1.10**	13.4	**0.70**
SD3.5-L	8.3	8.0	**22.8**	11.3	1.06	**16.5**	**21.8**	10.4	1.04	**10.7**	0.70
SD3.5-L-Turbo	0.63	8.0	**22.8**	10	0.93	22.6	21.7	9.9	0.9	13.5	0.70
SD3.5-L-SwD	**0.39**	8.0	**22.8**	**12.8**	**1.20**	20.6	**21.8**	**11.1**	**1.22**	13.9	**0.71**
FLUX	10.0	12.0	22.9	12.4	1.03	23.6	21.7	10.7	0.93	13.0	0.66
FLUX-Turbo-Alpha	2.75	12.0	**23.1**	13.4	1.08	21.2	21.5	11.2	0.97	11.3	0.66
FLUX-Schnell	1.41	12.0	22.6	11.2	1.01	**16.5**	21.5	10.3	0.96	**9.8**	0.69
FLUX-SwD	**0.72**	12.0	**23.1**	**14.6**	**1.14**	26.4	**21.9**	**11.6**	**1.06**	14.4	**0.71**

Table 3: Quantitative comparison of SwD against other leading open-source models. **Bold** denotes the best performing configuration, while underline the 2nd one.

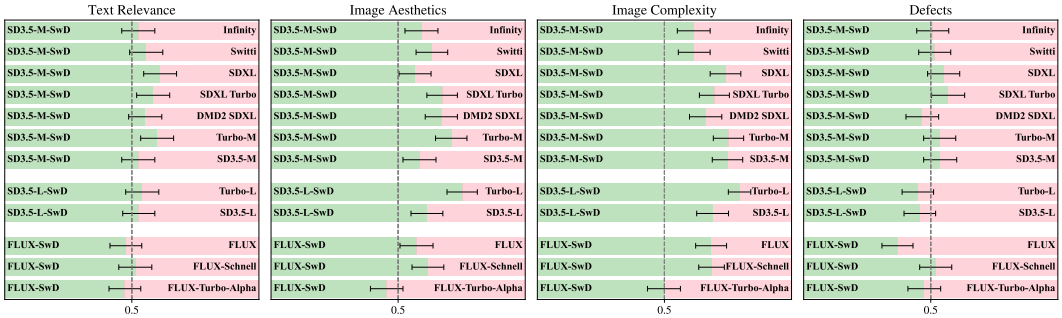

Figure 6: Human preference study for SwD against the baseline models.

5.2 Scale-wise versus Full-resolution

Next, we compare SwD against their full-resolution counterparts. The full-scale baselines use the same timestep schedules but operate at a fixed target latent resolution.

We provide the quality comparisons for the SD3.5 Medium and provide the FLUX results in Appendix E. Comparing the settings for the same number of steps (4 vs 4, 6 vs 6), human evaluation (Figure 7, Right) does not reveal any noticeable quality degradation. Qualitative examples (Figure 7, Left) further confirm this. Interestingly, automatic metrics (Tables 7 and 8) indicate that the scale-wise variants can even outperform their full-resolution counterparts, while being more efficient.

Then, we align generation times of scale-wise and full-resolution setups (4 vs. 2, 6 vs. 2 steps) to assess quality differences. Human evaluation reveals a clear advantage for the scale-wise setup, particularly in reducing *defects* and improving *image complexity*. Examples in Figure 7 (Left) highlight the high defect rates of the 2-step full-resolution baseline. Consistently, automatic metrics also show notable gains in HPSv3 and PS.

Runtime. Table 4 reports per-image generation latency (including VAE decoding and text encoding), and Table 5 shows average training iteration time. Compared to the full-resolution setting with the same number of steps (4 steps), the scale-wise setup achieves $\sim 2\times$ speedup in both training and sampling across text-to-image models, and $\sim 3\times$ for text-to-video.

5.3 Ablation study of MMD loss

Here, we study the role of the L_{MMD} loss and its design choices. Most experiments are conducted with the 6-step SD3.5-M setup from Section 5.1, with the MJHQ results reported in Table 6.

We first assess the L_{MMD} contribution to L_{SwD}. We observe that training with L_{MMD} alone underperforms the full L_{SwD} but remains effective as an independent distillation method, whereas removing

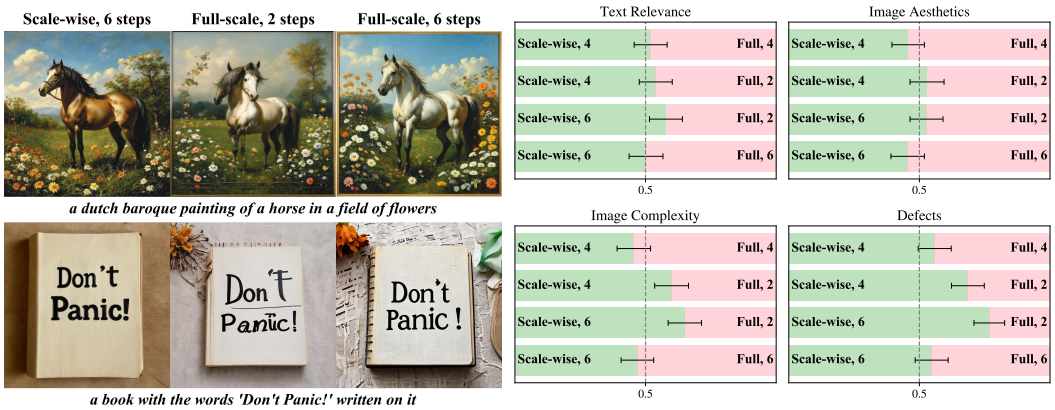

Figure 7: Visual examples (Left) and human preference study (Right) of the scale-wise and full-resolution settings within SD3.5 Medium. The numbers indicate the sampling steps.

Setup	Steps	SD3.5-M	SD3.5-L	FLUX	Wan2.1
Full-scale	4	0.29	0.63	1.41	5.51
Full-scale	2	0.16	0.33	0.72	2.97
Scale-wise	6	0.19	0.41	0.97	2.61
Scale-wise	4	0.17	0.32	0.72	1.84

Table 4: Sampling times (sec / image) of scale-wise and full-resolution setups. The measurement setting is described in Appendix G.

Setup	Loss	SD3.5-M	SD3.5-L	FLUX	Wan2.1
Full-scale	L_{SwD}	7.5	13.4	22.8	70.6
Full-scale	$L_{\text{SwD-MMD}}$	1.0	1.7	2.9	12.7
Scale-wise	L_{SwD}	3.2	7.8	11.3	23.9
Scale-wise	$L_{\text{SwD-MMD}}$	0.4	0.9	1.4	4.4

Table 5: Training times (sec / iteration) for scale-wise and full-resolution 4-step setups using the full objective (L_{SwD}) and MMD only ($L_{\text{SwD-MMD}}$).

L_{SwD} setup	PS ↑	HPSv3 ↑	IR ↑	FID ↓
L_{SwD} (Main)	**21.8**	10.7	1.11	**13.6**
L_{MMD} only	21.5	10.5	**1.15**	13.8
L_{SwD} w/o L_{MMD}	21.2	9.7	0.91	19.5
L_{MMD} Ablation				
A: RBF kernel	**21.8**	**10.8**	1.09	13.7
B: Batch averaging	21.5	10.5	0.97	16.4
C: w/o noising	21.3	10.2	1.01	16.6

Table 6: Ablation study of the L_{MMD} objective for SD3.5-Medium SwD on MJHQ30K.

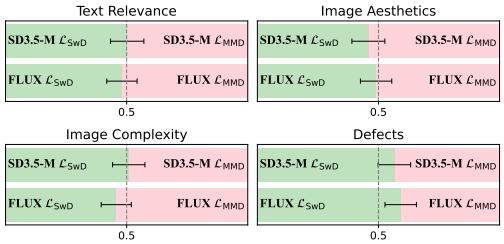

Figure 8: Comparison of the main SwD models against the ones distilled with L_{MMD} alone.

it from L_{SwD} leads to a significant drop in performance. Human evaluation (Figure 8) show that L_{MMD}-only models exhibit noticeable degradation in defects, though not severe. Visual comparisons (Figure 16) confirm that they provide comparable performance to the full L_{SwD}. Moreover, as shown in Table 5, L_{MMD}-only training enables more than 7× faster iterations since it avoids training extra models.

Finally, we examine several L_{MMD} variants. The L_{MMD} with the RBF kernel (**A**) shows similar results. Referring to the feature matching (Salimans et al., 2016), we consider two changes, **B**: the feature tokens in Equation (2) are averaged across the entire batch instead of per image, and **C**: extracting DM features only from clean samples, rather than noising them with the diffusion process. We observe that both **B** and **C** make L_{MMD} less effective.

6 Conclusion

We introduced SwD, a scale-wise diffusion distillation framework equipped with a novel MMD-based distillation technique. We show that both components can be readily combined with existing state-of-the-art distillation methods and lead to further efficiency and quality improvements for few-step models. We believe the proposed loss for DM distillation offers substantial potential for further development to pave the way toward a highly effective, self-contained distillation pipeline that eliminates the need for additional trainable models.

REFERENCES

Emanuele Aiello, Diego Valsesia, and Enrico Magli. Fast inference in denoising diffusion models via mmd finetuning. *arXiv preprint arXiv:2301.07969*, 2023.

Michael Arbel, Anna Korba, Adil SALIM, and Arthur Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/944a5ae3483ed5c1e10bbccb7942a279-Paper.pdf.

Yuval Atzmon, Maciej Bala, Yogesh Balaji, Tiffany Cai, Yin Cui, Jiaojiao Fan, Yunhao Ge, Siddharth Gururani, Jacob Huffman, Ronald Isaac, et al. Edify image: High-quality image generation with pixel space laplacian diffusion models. *arXiv preprint arXiv:2411.07126*, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Black Forest Labs. Flux.1. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024.

Ollin Boer Bohan. Taehv: Tiny autoencoder for hunyuan video. https://github.com/madebyollin/taehv, 2025.

Valentin De Bortoli, Alexandre Galashov, J Swaroop Guntupalli, Guangyao Zhou, Kevin Patrick Murphy, Arthur Gretton, and Arnaud Doucet. Distributional diffusion models with scoring rules. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=N82967FcVK.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. Training and testing low-degree polynomial data mappings via linear svm. *Journal of Machine Learning Research*, 11(4), 2010.

Sander Dieleman. Diffusion is spectral autoregression, 2024. URL https://sander.ai/2024/09/02/spectral-autoregression.html.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *CoRR*, abs/2403.03206, 2024.

Alexandre Galashov, Valentin De Bortoli, and Arthur Gretton. Deep MMD gradient flow without adversarial training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pf85K2wtz8.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Miguel Ángel Bautista, and Joshua M. Susskind. f-DM: A multi-stage diffusion model via progressive signal transformation. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=iBdwKIsg4m.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua M Susskind, and Navdeep Jaitly. Matryoshka diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023b.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2412.04431.

Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *arXiv preprint arXiv:2106.15282*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=08hStXdT1s.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Sergey Kastryulin, Artem Konev, Alexander Shishenya, Eugene Lyapustin, Artem Khurshudov, Alexander Tselousov, Nikita Vinokurov, Denis Kuznedelev, Alexander Markovich, Grigoriy Livshits, Alexey Kirillov, Anastasiia Tabisheva, Liubov Chubarova, Marina Kaminskaia, Alexander Ustyuzhanin, Artemii Shvetsov, Daniil Shlenskii, Valerii Startsev, Dmitrii Kornilov, Mikhail Romanov, Artem Babenko, Sergei Ovcharenko, and Valentin Khrulkov. Yaart: Yet another art rendering technology, 2024.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Pagoda: Progressive growing of a one-step generator from a low-resolution diffusion teacher, 2024. URL https://arxiv.org/abs/2405.14822.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17535–17545, 2023.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International conference on machine learning*, pp. 1718–1727. PMLR, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023a.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=MLIs5iRq4w.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score, 2025. URL https://arxiv.org/abs/2508.03789.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu,

Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2024a. URL https://arxiv.org/abs/2410.13720.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024b.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *International Conference on Learning Representations (ICLR)*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WNzy9bRDvG.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Danica J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350*, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. 2024.

van der Arjen Schaaf and van Johannes Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36:2759–2770, 1996. URL https://api.semanticscholar.org/CorpusID:18823051.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wei Wang, Yuan Sun, and Saman Halgamuge. Improving mmd-gan training with repulsive loss function. *arXiv preprint arXiv:1812.09916*, 2018.

Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, et al. Cache me if you can: Accelerating diffusion models through block caching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6211–6220, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=JVzeOYEx6d.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation, 2024. URL https://arxiv.org/abs/2412.21059.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025a.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=pwNSUo7yUb.

Mingyuan Zhou, Zhendong Wang, Huangjie Zheng, and Hai Huang. Long and short guidance in score identity distillation for one-step text-to-image generation. *ArXiv 2406.01561*, 2024a. URL https://arxiv.org/abs/2406.01561.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *International Conference on Machine Learning*, 2024b. URL https://arxiv.org/abs/2404.04057.

APPENDIX

A IMPLEMENTATION DETAILS

We combine SWD with DMD2 (Yin et al., 2024a), achieving state-of-the-art performance in diffusion distillation. We adapt DMD2 for transformer-based text-to-image DMs, whereas the original implementation is designed for the UNet-based models, such as SDXL (Podell et al., 2024).

Specifically, the generator consists of the pretrained DM with trainable LoRA adapters (Hu et al., 2022). The model is trained to minimize the *reverse KL-divergence* using the *scores* of the real and fake probability distributions. The real score is modeled using the pretrained DM, while the fake one is modeled by training a separate "fake" DM on the generated samples during distillation. The fake model is parameterized with other LoRA adapters added to the teacher DM.

GAN. Following DMD2, we also include a GAN loss. The discriminator is a small MLP (Haykin, 1994), which operates on the intermediate features extracted from the middle transformer block of the fake DM. The LoRA adapters of the fake DM are also updated using the discriminator loss.

The LoRA adapters are added to the attention and MLP layers, with a rank of 64 (SD3.5-M, SD3.5-L) and 128 (FLUX, Wan 2.1). The models are trained with a learning rate of $4e{-}6$ and batch sizes of 64 (SD3.5-M) and 24 (FLUX, SD3.5-L, Wan 2.1) for $3{-}4K$ iterations on a single node with 8 A100 GPUs. In the reverse KL-divergence, we set the guidance for the real score to 4.5 and 0.0 for the fake one. To train the discriminator, we use 4-layer MLP head including LayerNorm (Ba et al., 2016) and GELU (Hendrycks & Gimpel, 2023). The MLP head processes features extracted from the 11-th (SD3.5-M), 20-th (SD3.5-L), 15-th (FLUX, Wan2.1) transformer blocks of the fake DM.

$L_{\mathbf{MMD}}$. We use the timestep interval $[0, 600]$ to noise input samples prior to the feature extraction. The transformer blocks for feature extraction are the same as those used in the GAN setting.

Data. Similarly to LADD (Sauer et al., 2024), we train the models on the teacher synthetic data, prepared prior to distillation. The samples are generated using the standard teacher configuration. For SD3.5 Medium, we use 40 sampling steps with a guidance scale of 4.5. For SD3.5 Large, we use 28 sampling steps with a guidance scale of 4.5. For FLUX, we use 30 sampling steps with a guidance scale of 4.5. For Wan2.1, we use 50 steps with a guidance scale of 5.0.

B IMPORTANCE OF A SCALE-ADAPTED TEACHER MODEL

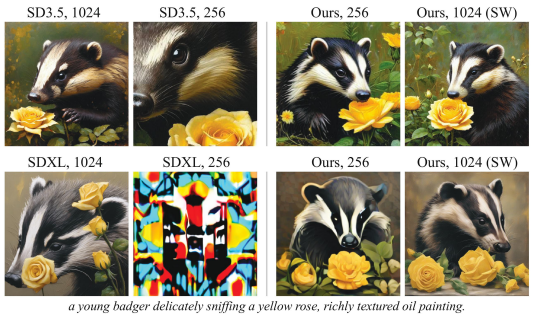

Figure 9: SD3.5 generates cropped images at low-resolutions ($256{\times}256$), while SDXL does not produce meaningful images at all. SWD is able to perform successful distillation for such cases and corrects these limitations.

We believe it is also important to address the following question: *Does the teacher model need to be capable of generating images at low scales prior to scale-wise distillation?* The teacher model may not inherently handle low scales effectively, making the scale-wise distillation a more challenging task compared to the full-scale distillation. If this is the case, additional pretraining of the teacher on small scales might be required, which could compromise the efficiency of the proposed approach.

To address this question, we evaluate the ability of SD3.5 Medium and SDXL to generate images at lower scales ($256{\times}256$). The results are presented in Figure 9. We find out that SD3.5 produces cropped and simplified images, but the overall quality remains acceptable due to its pretraining on $256{\times}256$ resolution. SWD effectively distills this model, correcting the cropped images and enhancing their complexity. However, more crucially, SDXL is unable to generate plausible images at $256{\times}256$ resolution. Interestingly, SWD can still produce a plausible generator even with such a poor starting point for distillation.

Setup	# steps	PS ↑	HPSv3 ↑	IR ↑	FID ↓
SD3.5 Medium					
Scale-wise	6	**22.8**	**11.7**	1.10	23.1
Scale-wise	4	22.7	**11.7**	**1.12**	23.7
Scale-wise	2	22.6	10.6	1.09	22.3
Full-scale	6	22.5	11.2	1.08	20.4
Full-scale	4	22.5	11.3	1.09	21.2
Full-scale	2	22.3	10.8	1.03	**20.3**
FLUX					
Scale-wise	4	**23.1**	**14.6**	**1.14**	**26.4**
Scale-wise	2	23.0	14.1	1.12	26.5
Full-scale	4	**23.1**	14.0	1.13	28.5
Full-scale	2	23.0	13.8	1.13	26.9

Table 7: Quantitative comparison between scale-wise and full-scale setups in terms of automatic metrics on COCO30K.

Setup	# steps	PS ↑	HPSv3 ↑	IR ↑	FID ↓
SD3.5 Medium					
Scale-wise	6	**21.8**	**10.7**	1.10	13.4
Scale-wise	4	**21.8**	**10.7**	**1.13**	13.7
Scale-wise	2	21.7	10.3	1.10	**12.8**
Full-scale	6	21.6	10.3	1.09	13.4
Full-scale	4	21.7	10.4	1.10	13.5
Full-scale	2	21.5	10.0	1.04	13.1
FLUX					
Scale-wise	4	**21.9**	**11.6**	1.06	14.4
Scale-wise	2	**21.9**	11.5	**1.10**	14.0
Full-scale	4	21.8	11.3	1.09	14.4
Full-scale	2	21.8	11.2	1.08	**13.4**

Table 8: Quantitative comparison between scale-wise and full-scale setups in terms of automatic metrics on MJHQ30K.

C Architecture choice

Although SwD can be adapted to arbitrary DM architectures, we primarily focus on its application to latent- and transformer-based diffusion models, i.e., the variants of the DiT architecture (Peebles & Xie, 2022), which are most widely used in state-of-the-art text-conditional image and video models (Esser et al., 2020; Black Forest Labs, 2024; Polyak et al., 2024a; Wan et al., 2025).

A key characteristic of DiT-based models is their reliance on attention layers (Vaswani, 2017), which scale quadratically with spatial resolution. Additionally, these models maintain a constant number of tokens across layers without downscaling, unlike the UNet-based DMs (Podell et al., 2024; Rombach et al., 2021). These factors underscore the particular significance of SwD for such architectures.

D SwD setups in main experiments

Below, we provide the time and scale schedules used in our main experiments. The scale schedule shows the resolutions in the corresponding VAE latent spaces. Note that the schedules do not require extensive tuning.

SD3.5 Medium. Timesteps $t = [1000, 945, 896, 790, 737, 602]$. Scales $s = [32, 48, 64, 80, 96, 128]$.

SD3.5 Large. Timesteps $t = [1000, 896, 737, 602]$. Scales $s = [64, 80, 96, 128]$.

FLUX. Timesteps $t = [1000, 945, 790, 602]$. Scales $s = [32, 64, 96, 128]$.

Wan2.1. Timesteps $t = [1000, 896, 737, 602]$.
Scales $s = [6{\times}20{\times}34, 11{\times}30{\times}52, 16{\times}40{\times}70, 21{\times}60{\times}104]$.

E Scale-wise versus Full-resolution

Here, we provide additional results to compare the various scale-wise and full-resolution settings. Table 7 and Table 8 present the automatic metrics for SD3.5 Medium and FLUX on the COCO and MJHQ datasets. The visual examples for FLUX, SD3.5 Large, SD3.5 Medium are presented in Figures 10, 11, and 17, respectively.

Also, in Table 9, we evaluate the use of "constant" 6-step scale schedules s=$[64, 64, 64, 64, 64, 128]$ and s=$[32, 32, 32, 32, 32, 128]$ for SD3.5 Medium in contrast to the progressively growing schedule s=$[32, 48, 64, 80, 96, 128]$, used in our main setup. Note that the last step is required to be made at the target resolution.

The results show that it is important to gradually increase the resolution over sampling steps.

Figure 10: Visual examples of 4-step scale-wise and 2-step full-resolution SD3.5-Large settings.

Figure 11: Visual examples of 4-step scale-wise and 2-step full-resolution FLUX settings.

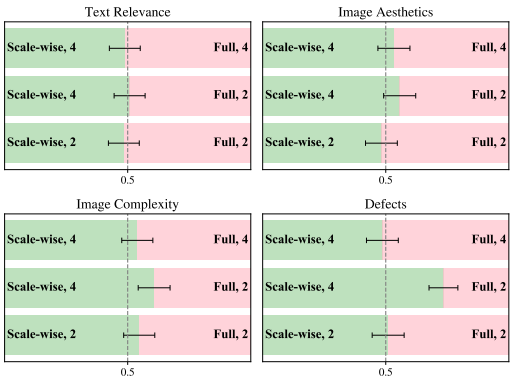

Figure 12: Human preference study comparing scale-wise and full-resolution FLUX setups.

Setup	PS ↑	HPSv3 ↑	IR ↑	FID ↓
COCO2014				
Main s=[32, 48, 64, 80, 96, 128]	**22.8**	**11.7**	**1.10**	**23.1**
s=[64, 64, 64, 64, 64, 128]	22.4	10.3	1.02	23.7
s=[32, 32, 32, 32, 32, 128]	22.3	9.8	0.97	23.8
MJHQ				
Main s=[32, 48, 64, 80, 96, 128]	**21.8**	**10.7**	**1.11**	**13.6**
s=[64, 64, 64, 64, 64, 128]	21.3	9.8	1.06	14.6
s=[32, 32, 32, 32, 32, 128]	21.2	9.4	0.99	15.7

Table 9: Comparisons to the "constant" scale schedules for SD3.5-Medium SwD.

F Extended latent space spectral analysis

Figure 13 provides more results for the SD3.5 and Wan2.1 models and also includes the analysis for FLUX (Black Forest Labs, 2024) and SDXL (Podell et al., 2024). In contrast to other models, the SDXL model (Podell et al., 2024) uses a variance-preserving (VP) diffusion process (Ho et al., 2020; Song et al., 2020b). The SDXL latent space has C=4 channels and 128×128 spatial resolution.

G Runtime measurement setup

In our experiments, we measure runtimes in half-precision (FP16), using *torch.compile* for all models: VAE decoders, text encoders, and generators. Note that, under our very fast sampling settings, the computational costs of the text encoder and VAE start to account for a noticeable portion of the overall runtime. Thus, we replace original VAEs with TinyVAEs (Boer Bohan, 2025) for all models.

The measurements are conducted in an isolated environment on a single A100 GPU. We use a batch size of 8 for all runs, and each measurement is averaged over 100 independent runs. The latency is then obtained by dividing the average runtime by the batch size.

H HUMAN EVALUATION

The evaluation is conducted using Side-by-Side (SbS) comparisons, where assessors are presented with two images alongside a textual prompt and asked to choose the preferred one. For each pair, three independent responses are collected, and the final decision is determined through majority voting.

The human evaluation is carried out by professional assessors who are formally hired, compensated with competitive salaries, and fully informed about potential risks. Each assessor undergoes detailed training and testing, including fine-grained instructions for every evaluation aspect, before participating in the main tasks.

In our human preference study, we compare the models across four key criteria: relevance to the textual prompt, presence of defects, image aesthetics, and image complexity. Figures 19, 22, 20, 21 illustrate the interface used for each criterion. Note that the images displayed in the figures are randomly selected for demonstration purposes.

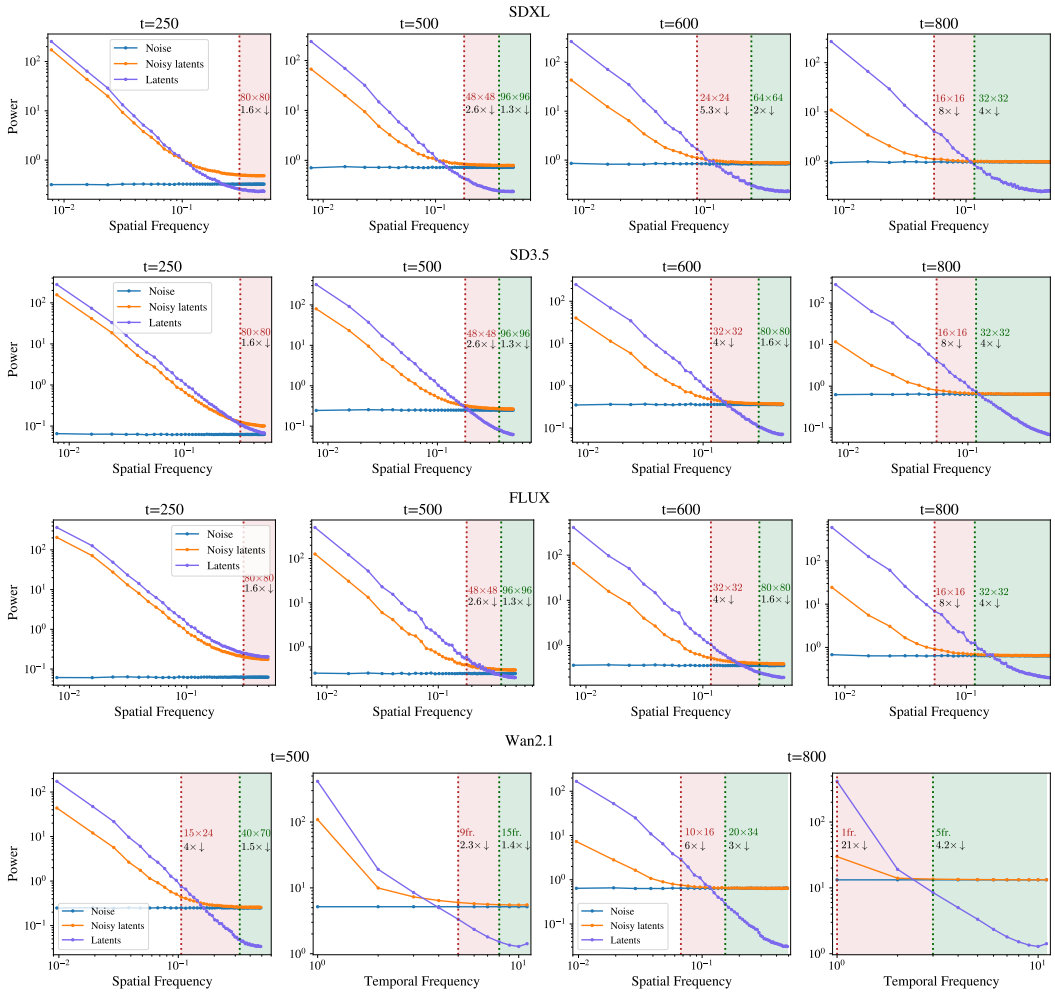

Figure 13: Extended spectral analysis from Section 3 to more timesteps and models (FLUX, SDXL).

Method	Total Score	Creativity Score	Commonsense Score	Controllability Score	Human Fidelity Score	Physics Score	Human Anatomy	Human Clothes	Human Identity	Composition	Diversity	Mechanics	Material	Thermotics	Multi-View Consistency	Dynamic Spatial Relationship	Dynamic Attribute	Motion Order Understanding	Human Interaction	Complex Landscape	Complex Plot	Camera Motion	Motion Rationality	Instance Preservation
Wan 2.1	51.59	**53.75**	57.06	22.65	83.03	41.45	87.00	91.24	70.85	42.56	**64.94**	59.16	36.58	57.89	12.15	26.08	15.01	21.21	46.33	**19.77**	**11.02**	**19.13**	28.16	85.96
Spatial SwD	53.39	47.62	57.96	22.76	**92.94**	45.68	91.89	**100.00**	**86.93**	**54.31**	40.93	**59.66**	49.05	55.22	18.81	24.15	**15.75**	**21.35**	**67.33**	16.00	7.09	7.71	25.28	**90.64**
SwD	**53.50**	44.78	**61.40**	**22.80**	91.93	**46.61**	**92.69**	99.10	84.02	48.05	41.52	54.33	**52.04**	**60.14**	**19.96**	**33.81**	14.28	15.15	56.33	17.55	8.94	13.58	**32.75**	90.05

Table 10: Full comparison on VBench2.0 using all 18 metrics.

Figure 14: Qualitative results of Wan2.1-SwD.

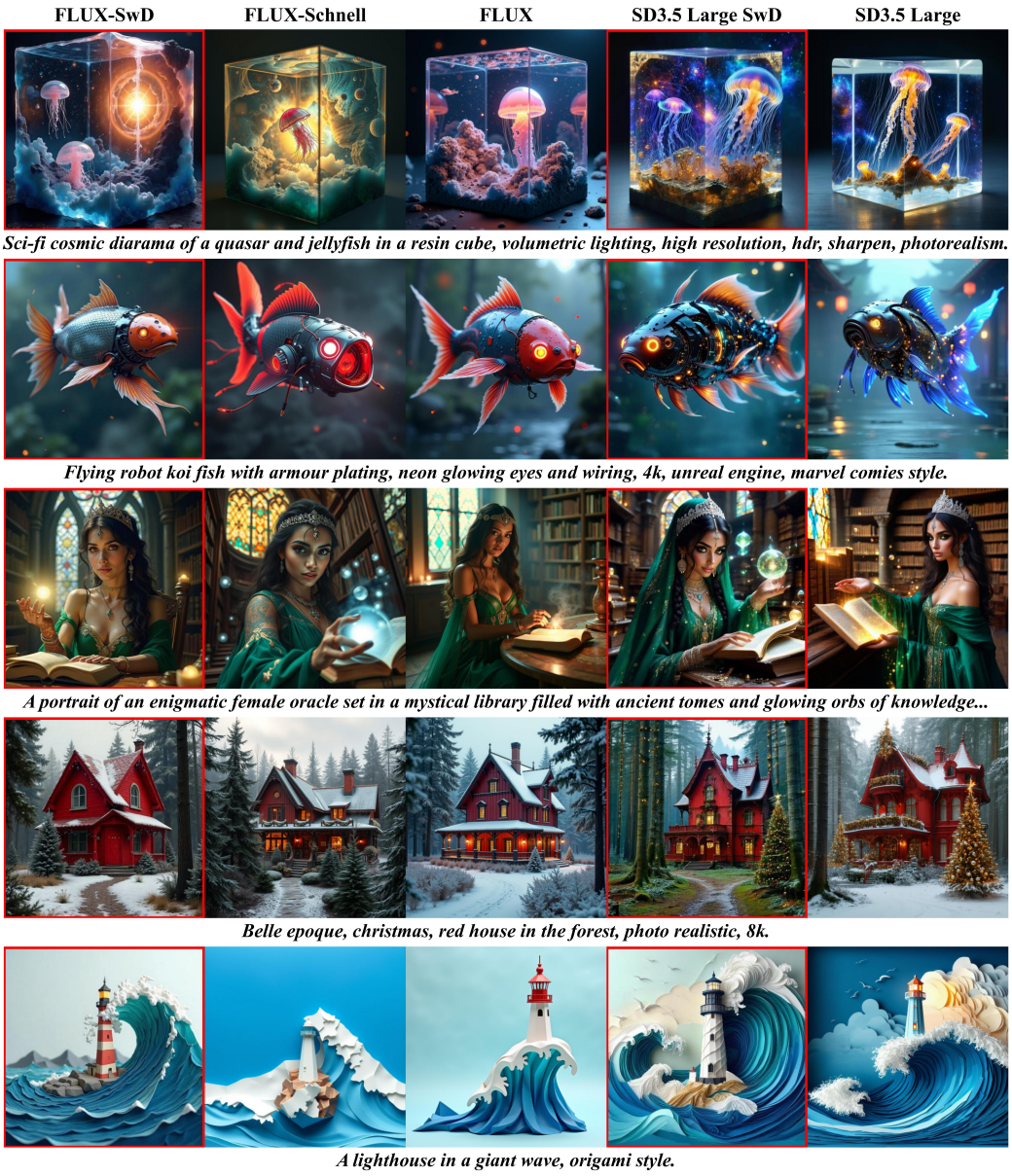

Figure 15: Qualitative results of FLUX-SwD and SD3.5 Large SwD.

Figure 16: Qualitative comparisons of SwD trained with the full L_{SwD} loss against the ones trained with L_{MMD} alone. L_{MMD} in isolation produces competitive few-step models.

Figure 17: Qualitative examples of image generations using scale-wise and full-resolution SD3.5 Medium SwD variants for different generation steps.

Figure 18: Qualitative comparison of SD3.5 Medium SwD against the models of the similar size.

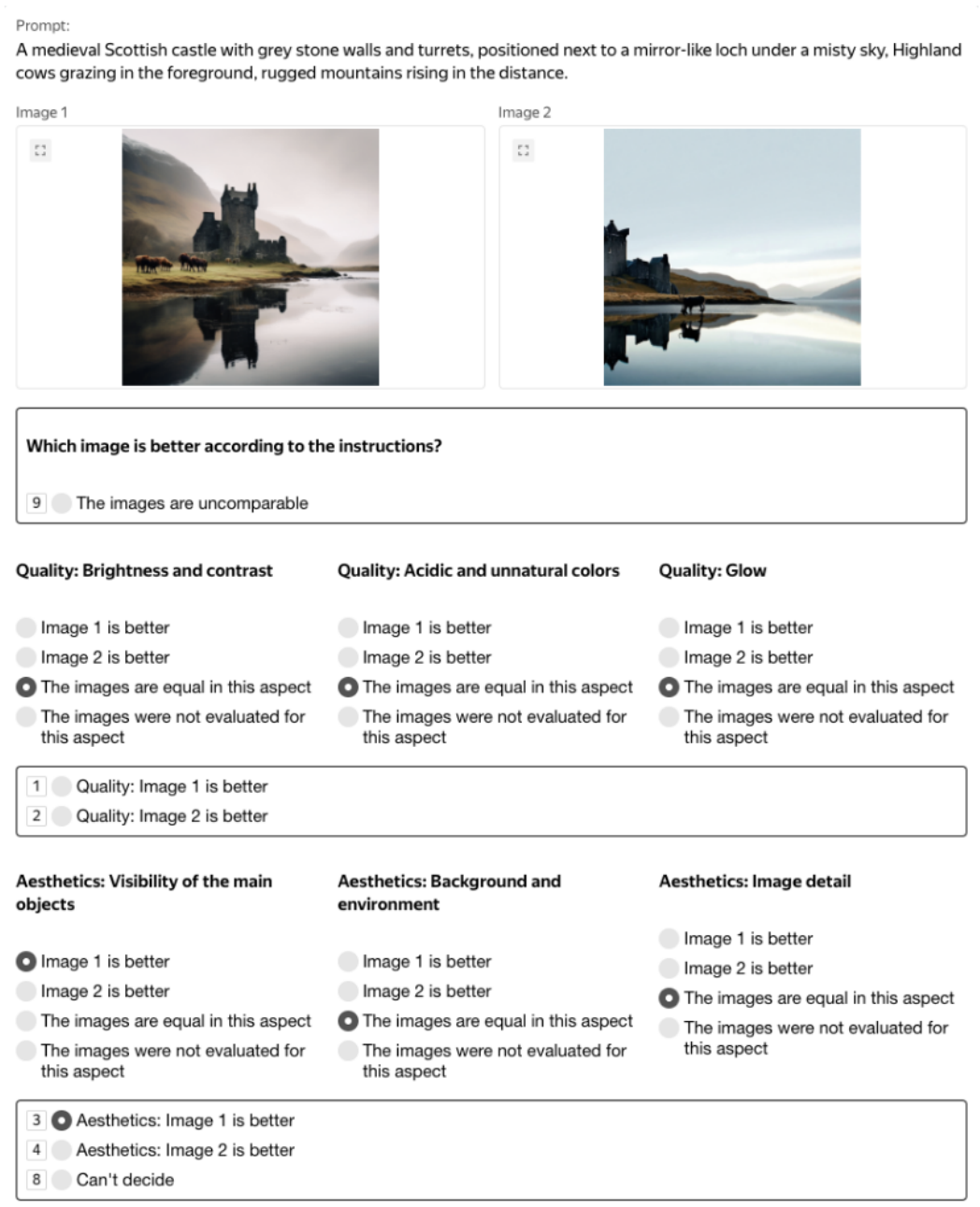

Figure 19: Human evaluation interface for aesthetics.

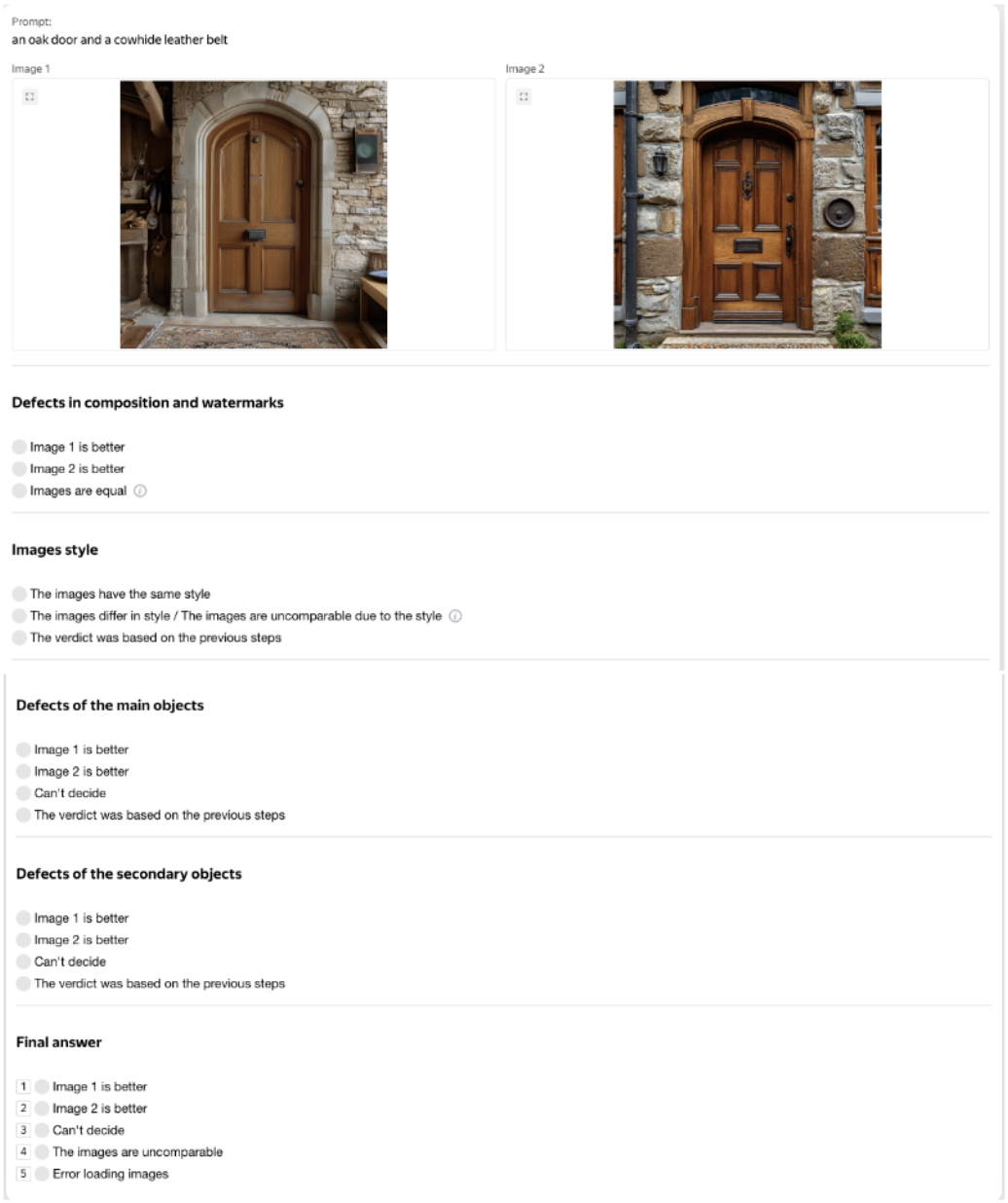

Figure 20: Human evaluation interface for defects.

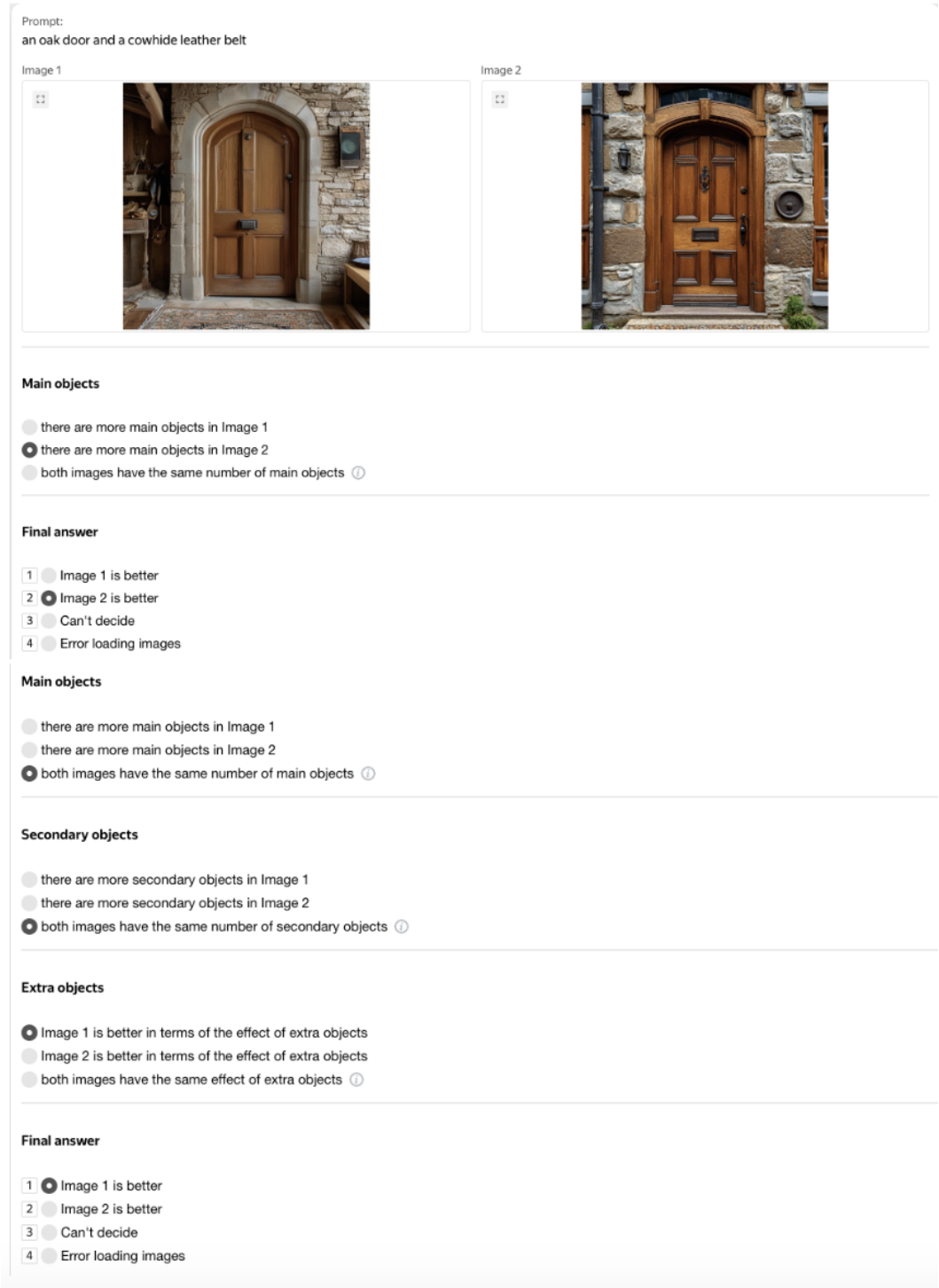

Figure 21: Human evaluation interface for relevance.

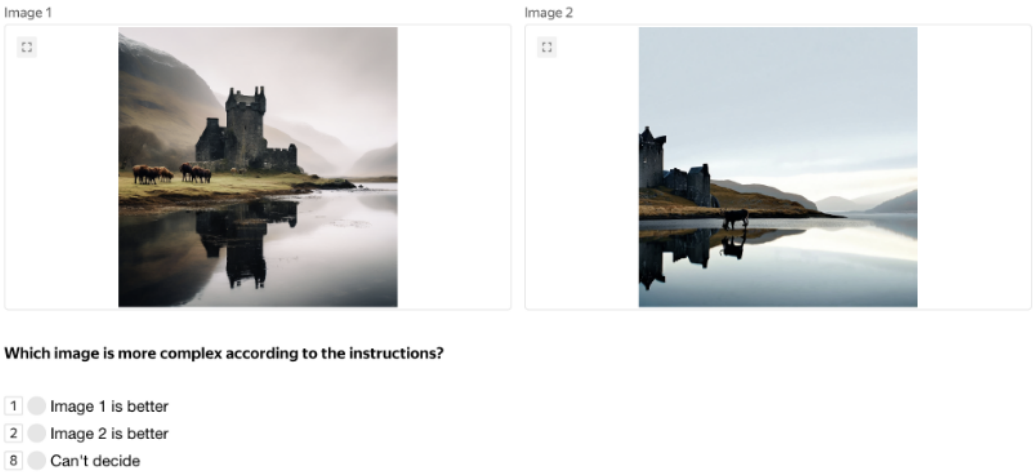

Figure 22: Human evaluation interface for complexity.