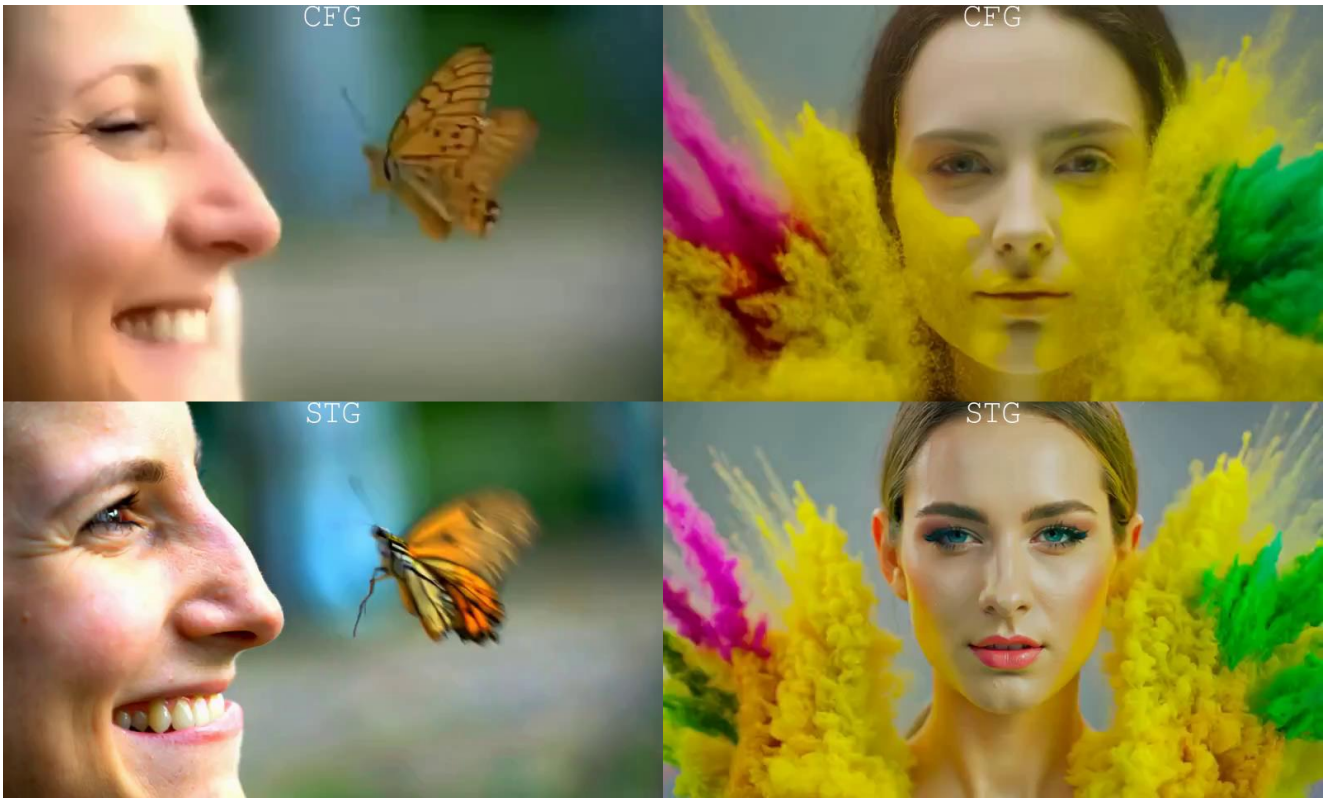
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Spatiotemporal Skip Guidance for Enhanced Video Diffusion Sampling

Junha Hyung*[1] Kinam Kim*[1] Susung Hong[2] Min-Jung Kim[1] Jaegul Choo[1]

[1]KAIST AI [2]University of Washington

{sharpeeee, kinamplify, emjay73, jchoo}@kaist.ac.kr, susung@cs.washington.edu



"A close-up shot of a butterfly landing on the nose of a woman, highlighting her smile and the details of the butterfly's wings."

"A close-up of a woman's face with colored powder exploding around her, creating an abstract splash of vibrant hues."

Figure 1. Visual comparison of video quality between CFG (top row) and our STG method (bottom row). Best viewed in Acrobat Reader; click on the images to watch the videos.

Abstract

Diffusion models have emerged as a powerful tool for generating high-quality images, videos, and 3D content. While sampling guidance techniques like CFG improve quality, they reduce diversity and motion. Autoguidance mitigates these issues but demands extra weak model training, limiting its practicality for large-scale models. In this work, we introduce Spatiotemporal Skip Guidance (STG), a simple training-free sampling guidance method for enhancing transformer-based video diffusion models. STG employs an implicit weak model via self-perturbation, avoiding the need for external models or additional training. By selectively skipping spatiotemporal layers, STG produces an aligned, degraded version of the original model to boost sample quality without compromising diversity or dynamic degree. Our contributions include: (1) introducing STG as

* indicates equal contribution.

an efficient, high-performing guidance technique for video diffusion models, (2) eliminating the need for auxiliary models by simulating a weak model through layer skipping, and (3) ensuring quality-enhanced guidance without compromising sample diversity or dynamics unlike CFG. For additional results, visit https://junhahyung.github.io/STGuidance/.

1. Introduction

Diffusion models [11, 30–32] are a successful class of generative models known for their flexibility in modeling complex data distributions, achieving impressive results in image, video, and 3D generation. By progressively denoising random noise, they enable robust generalization, making them a leading choice for realistic content generation and often surpassing GAN-based methods [5, 9, 18, 29]. Building on this success, video diffusion models [4, 12, 27, 33, 38] generate high-quality videos by using temporal or 3D attention layers to handle sequential frames.

Meanwhile, to enhance sample quality, sampling guidance techniques such as Classifier-Free Guidance (CFG) [10] and Autoguidance [19] have been introduced to guide the denoising process. These techniques employ *weak models* to predict poor trajectories, steering the main model away from them and pushing samples toward high-quality regions on the data manifold. However, CFG often reduces diversity, leading to saturated or overly simplified results [6, 19].

Autoguidance [19] addresses this issue by using a weak model trained on the same task, conditioning, and data distribution as the main model. The main drawback of this approach, however, is the need to train an additional weak model, which is impractical for large-scale models. Alternative methods such as PAG [1] and SEG [13] use self-perturbation to implicitly mimic a weak model, avoiding the need for extra training. Yet, these methods are designed specifically for image generation diffusion models, applying self-perturbation to 2D spatial attention maps.

In this work, we propose Spatiotemporal Skip Guidance (STG), a simple and effective sampling guidance method for video diffusion models that significantly enhances the performance of any transformer-based video diffusion model without additional training. Specifically, we use an implicit weak model for guidance through self-perturbation, eliminating the need for explicit weak models and their associated training costs. This is especially crucial for video diffusion models, where training costs are high.

Our implicit weak model, a key component of our framework, is deliberately designed as a degraded but "aligned" version of the original video generation model. As demonstrated in Autoguidance [19], having the weak model share the same task, conditioning, and data distribution as the main model is essential for quality improvement, as we expect both models to exhibit similar, aligned errors in the same parts of samples. We conceptually illustrate this alignment in Fig. 2, where our weak model and main model are arranged in a direction of increasing quality. To achieve this, we apply spatiotemporal perturbation to both spatial and temporal attention layers—or, in the case of 3D attention, to the entire layer—by selectively *skipping* certain layers. This straightforward approach effectively nullifies specific residual or attention layers, generating a lower-quality version of the model that simulates an aligned weak model.

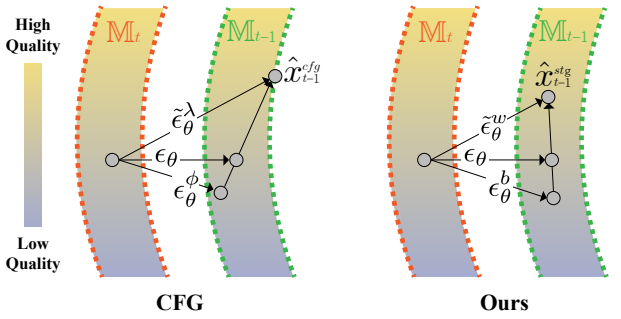


Figure 2. Comparison between CFG and STG, with the band conceptually representing the noisy data manifold. In STG, the weak model and the main model are aligned along the direction of increasing quality. In contrast, the two models in CFG differ not only in quality but also in aspects such as diversity and prompt alignment capabilities.

To ensure samples remain on the data manifold, even when using large guidance scales, we employ optional techniques such as the rescaling [21] and restart [36] methods. Rescaling the latent code constrains its variance, addressing the issue of larger variance causing saturation in the results [21]. Meanwhile, the restart sampling technique leverages the error contraction property of forward SDEs [36] to keep the sampling trajectory on the manifold. These techniques help prevent overshooting in the sampling trajectory, which could otherwise push samples off the manifold and result in saturated or distorted outputs.

Our key contributions are as follows:

- We propose STG—a surprisingly simple sampling guidance framework that significantly boosts the performance of video diffusion models.
- Our method introduces an implicit weak model by skipping spatiotemporal layers in video diffusion models, eliminating the need for additional training or external models.
- Our method enhances sample quality during guidance without reducing diversity or limiting the dynamics of generated videos.

2. Related Work

Guidance with trained weak model Classifier-Free Guidance (CFG) [10] improves conditional generation in Diffusion Models by using an implicit unconditional model as a weak model. However, differences in tasks between the unconditional and conditional models can reduce sample diversity [6, 19] and increase sampling trajectory curvature [6], leading to overshooting the data manifold and producing skewed or oversaturated images.

Autoguidance [19] mitigates these issues by employing a *bad version of the main model* as a weak model, trained with reduced capacity and compute to ensure alignment. This alignment allows the guidance algorithm to correct errors by analyzing prediction differences. While effective, it requires additional training, which is challenging for large-scale video diffusion models.

Guidance with training-free weak model Another line of work avoids additional training by using self-perturbation of the main model to mimic a weak model. Self-Attention Guidance (SAG) [14] blurs high-attention regions, Perturbed Attention Guidance (PAG) [1] replaces attention maps with identity matrices, and Smoothed Energy Guidance (SEG) [13] applies Gaussian blur to the attention weights to smooth the energy landscape. These methods guide sampling toward high-quality outputs by leveraging differences in predictions from the weakened model. While effective, they are primarily designed for image diffusion models with 2D self-attention. We aim to extend this approach to video diffusion models, which require handling temporal dynamics with additional temporal or 3D spatiotemporal attention layers.

3. Preliminaries

3.1. Diffusion models

Diffusion models generate samples by progressively removing noise from noisy data, restoring the original data distribution. Song et al. [32] defined the process of adding noise to the data using a stochastic differential equation (SDE):

$$dx = -\frac{\beta(t)}{2}x\,dt + \sqrt{\beta(t)}\,dw, \tag{1}$$

where $\beta(t)$ is a time-dependent noise schedule, and w represents the standard Wiener process. Corresponding reverse-time SDE is:

$$dx = \left[-\frac{\beta(t)}{2}x - \beta(t)\nabla_x \log p_t(x)\right] dt + \sqrt{\beta(t)}\,d\bar{w} \tag{2}$$

where $d\bar{w}$ is the Wiener process running backward in time. Here, the score function $\nabla_x \log p_t(x)$ is approximated by a neural network $s_\theta(x(t))$ trained using denoising score matching [35]:

$$\theta^* = \arg\min_\theta \mathbb{E}_t \left\{\lambda(t)\mathbb{E}_{x_0}\mathbb{E}_{x_t|x_0}\left[\|s_\theta(x_t) - \nabla_{x_t}\log p_t(x_t|x_0)\|_2^2\right]\right\}. \tag{3}$$

3.2. Classifier Guidance

Classifier Guidance (CG) [25] reformulates the reverse process of a diffusion model by incorporating an external classifier p_ϕ as

$$\tilde{p_\theta}(x_t|y) \propto p_\theta(x_t)p_\phi(y|x_t)^\lambda, \tag{4}$$

where y is the desired class label and λ controls the guidance strength. The score function is then derived as:

$$\nabla_{x_t}\log\tilde{p}_\theta(x_t|y) = \nabla_{x_t}\log p_\theta(x_t) + \lambda\nabla_{x_t}\log p_\phi(y|x_t). \tag{5}$$

By substituting the score function in Eq. 2 with Eq. 5, sampling from the desired class condition y becomes possible.

3.3. Classifier-Free Guidance

Classifier-Free Guidance (CFG) [10] uses Bayes' rule to replace a classifier-guided score with a linear combination of conditional and unconditional score estimates:

$$\tilde{\epsilon}_\theta^\lambda(x_t) = \epsilon_\theta(x_t) + \lambda\left(\epsilon_\theta(x_t) - \epsilon_\theta(x_t|\phi)\right). \tag{6}$$

CFG jointly trains the unconditional model $\epsilon_\theta(x_t|\phi)$ and the conditional model $\epsilon_\theta(x_t|c)$ $(= \epsilon_\theta(x_t))$ within a single model by setting the condition c to a null token ϕ. Using this guided denoising process, the model better captures the conditions and often generates high-fidelity outputs.

4. Method

4.1. Optimal Weak Model Design

Our goal is to construct an *aligned* weak model that captures a distribution similar to the original model while generating slightly lower-quality samples. This ensures that the guidance gradient points toward improved quality, as shown in Fig. 2. Misaligned models, as seen in CFG [10], may lead to unintended outcomes like reduced diversity, while Autoguidance [19] requires additional weak model training, making it impractical for large-scale models.

To address these limitations, we design an *implicit* weak model by leveraging the original model itself, eliminating the need for external training. This training-free approach ensures computational efficiency and better alignment, as the weak model is derived directly from the original, preserving most network weights.

To achieve this, we apply perturbation methods directly to the main model's forward pass, creating an implicit weak

Models	Imaging Quality	Aesthetic Quality	Motion Smoothness	Dynamic Degree	Temporal Flickering
Mochi (CFG)	0.524	0.507	0.985	**0.87**	0.976
Mochi (STG)	**0.628**	**0.554**	**0.988**	0.86	**0.978**
Open-Sora (CFG)	0.561	0.493	0.982	**0.902**	0.975
Open-Sora (STG)	**0.606**	**0.509**	**0.987**	0.895	**0.976**

Table 1. Quantitative results for Mochi [33] and Open-Sora [38] on VBench [15] T2V benchmarks.

Models	FVD (↓)	IS	Imaging Quality	Aesthetic Quality	Motion Smoothness	Dynamic Degree
SVD (CFG)	151.3	38.0	0.687	0.637	0.966	0.562
SVD (STG)	**128.7**	**38.5**	**0.694**	**0.639**	**0.968**	**0.694**

Table 2. Quantitative results for SVD [4] on FVD, IS, and VBench [15] I2V benchmarks.

model. Specifically, we use *spatiotemporal* perturbation, a key component for aligning the weak model with video diffusion models, which will be elaborated upon in detail.

4.2. Sampling from High Quality Samples

Similar to CG [25], we define our goal as conditioning the model on an imaginary label y_g that represents high-quality samples, leading to the sampling distribution

$$\tilde{p}_\theta(x_t|y_g) \propto p_\theta(x_t)p_\phi(y_g|x_t)^w. \tag{7}$$

Using $w > 0$ sharpens the distribution, promoting the generation of high-quality samples. From Eq. 7, the score is derived as

$$\nabla_{x_t} \log \tilde{p}_\theta(x_t|y_g) = \nabla_{x_t} \log p_\theta(x_t) + w\nabla_{x_t} \log p_\phi(y_g|x_t). \tag{8}$$

Rather than using an external classifier p_ϕ, we propose using our model p_θ as an implicit classifier. We design this classifier to be inversely proportional to the probability of an imaginary "bad" label y_b, expressed as

$$p_\phi(y_g|x_t) \propto \frac{1}{p_\theta(y_b|x_t)}. \tag{9}$$

Using Bayes' rule, the function can be expressed in terms of the marginal posterior as follows:

$$\begin{aligned}\nabla_{x_t} \log \frac{1}{p_\theta(y_b|x_t)} &= \nabla_{x_t} \log \frac{p_\theta(x_t)}{p_\theta(y_b)p_\theta(x_t|y_b)} \\ &= \nabla_{x_t}(\log p_\theta(x_t) - \log p_\theta(x_t|y_b)),\end{aligned} \tag{10}$$

leading to the score of the target distribution as:

$$\nabla_{x_t} \log \tilde{p}_\theta(x_t|y_g) = \nabla_{x_t} \log p_\theta(x_t) + w\nabla_{x_t}(\log p_\theta(x_t) - \log p_\theta(x_t|y_b)). \tag{11}$$

We can sample from $\tilde{p}_\theta(x_t|y_g)$ by substituting the score function in Eq. 2 with Eq. 11, resulting in:

$$\begin{aligned}dx = \Big[&- \frac{\beta(t)}{2}x - \beta(t)(\nabla_{x_t} \log p_t(x_t) \\ &+ w\nabla_{x_t}(\log p_\theta(x_t) - \log p_\theta(x_t|y_b))\Big]dt + \sqrt{\beta(t)}d\bar{w},\end{aligned} \tag{12}$$

and solving the reverse SDE. Since the score function is approximated using the neural network, we can generate samples using

$$\tilde{\epsilon}^w_\theta(x_t) = \epsilon_\theta(x_t) + w(\epsilon_\theta(x_t) - \epsilon^b_\theta(x_t)). \tag{13}$$

An interesting approach here is to model $\nabla_{x_t} \log p_\theta(x_t|y_b)$ by perturbing the forward pass of $\epsilon_\theta(x_t)$, denoted as $\epsilon^b_\theta(x_t)$. Now the main focus is designing a perturbation that effectively yields the weak model capable of estimating $\epsilon^b_\theta(x_t)$, aligning closely with $\epsilon_\theta(x_t)$, as discussed in Sec. 4.1. Ideally, we want a distribution that deviates minimally from $\epsilon_\theta(x_t)$ while producing slightly lower-quality samples.

4.3. Spatiotemporal Skip Guidance (STG)

We introduce Spatiotemporal Skip Guidance (STG), a simple guidance method designed for video diffusion models that generates diverse, high-fidelity samples while addressing the limitations of existing methods. STG implicitly simulates an aligned weak model through spatiotemporal perturbation, capturing the spatiotemporal dynamics of video data.

An interesting discovery of this paper is that *skipping* layers within the network is an effective approach for constructing an aligned weak model. Modern neural network architectures for video diffusion models, such as ADM [25], DiT [26], and SiT [23], are partially or fully transformer-based and contain attention layers and residual blocks. These architectures are well-suited for layer skipping, as



Figure 3. Selected frames from videos generated by Mochi [33] with increasing STG scales.

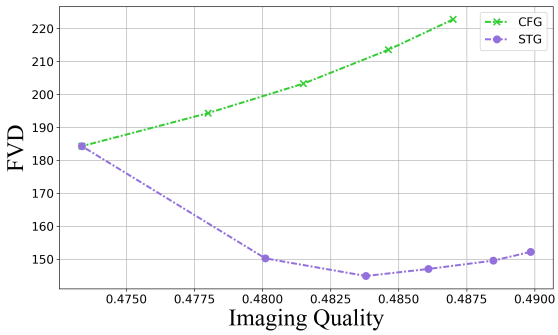


Figure 4. Comparison of CFG and STG across varying scales in terms of Imaging Quality and FVD.

they can still produce plausible outputs without significant degradation, even when a few layers are removed.

Layer skipping is advantageous for generating an aligned weak model because it retains most of the neural network's weights and forward pass, resulting in similar predictions and distributions. This approach offers a clear advantage over methods that rely on external models, such as Autoguidance [19], or alternative objectives like CFG [10], where forward passes differ significantly.

Moreover, our method applies perturbations to both spatial and temporal layers (or spatiotemporal layers), unlike existing image-based perturbation methods [13, 14] limited to 2D spatial attention layers. This dual-layer perturbation is essential for aligning the weak model with the main video diffusion model. We now discuss various STG configurations that can be applied across different video diffusion model architectures.

Residual skip To skip an entire residual block, we modify $\text{Res}(z_l)$ to $\text{Res}'(z_l)$,

$$\text{Res}(z_l) = z_{l+1} = f_l(z_l) + z_l, \tag{14}$$

$$\text{Res}'(z_l) = z_{l+1} = z_l, \tag{15}$$

where z_l and z_{l+1} represent the features at the l^{th} and $(l+1)^{\text{th}}$ layers, respectively, and f_l denotes the l^{th} neural net layer. The residual layers, which add small residuals $f_l(z_l)$ to the original feature z_l, ensure that the perturbed z_{l+1} does not deviate significantly from the original z_{l+1}. This reduces out-of-distribution issues in consecutive layers, generating perturbed but aligned samples.

Attention skip Self-attention computes a linear combination of value tensors as

$$\text{SA}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V = \mathbf{A}V, \tag{16}$$

where $Q \in \mathbb{R}^{(h\times w\times f)\times d}$, $K \in \mathbb{R}^{(h\times w\times f)\times d}$, $V \in \mathbb{R}^{(h\times w\times f)\times d}$ are the query, key, and value matrices, respectively. Here, h, w, f, and d represent the height, width, frame number, and channel dimensions.

We can skip this layer partially by passing the value matrix directly to the next layer without computing its linear combination. This is equivalent to replacing the attention matrix $\mathbf{A}$ with an identity matrix $\mathbf{I} \in \mathbb{R}^{hwf\times hwf}$, resulting in

$$\text{SA}'(Q, K, V) = \mathbf{I}V. \tag{17}$$

This represents a 3D extension of PAG [1].

Factorized attention While recent models like Movie Gen [27] and Mochi [33] utilize full 3D spatiotemporal attention layers, many models, such as SVD [4] and Open-Sora [38], still use factorized attention layers for efficiency. These models employ sequential 2D spatial attention and 1D temporal attention to approximate 3D spatiotemporal attention. For factorized models, we apply skip perturbation of Eq. 17 to spatial and temporal layers separately and use their linear combination for the final guidance.

We denote the spatial and temporal perturbation labels as y_{sb} and y_{tb}, respectively. In practice, y_{sb} and y_{tb} are rarely independent, as they can influence each other. For instance, random spatial perturbations that create varied color adjustments across frames may disrupt temporal continuity. Similarly, random temporal perturbations can affect spatial consistency, leading to distortions in specific frames. However, for our *skip* perturbations, we can loosely assume independence, as skipping layers primarily reduces details in residual networks, which may not inherently affect temporal consistency.

Therefore, to simplify the derivation, we initially assume independence between spatial and temporal perturbations.

(1) A romantic scene of a couple dancing under string lights in a backyard, with warm, golden tones highlighting their laughter.
(2) An animation showing a floating castle drifting above the clouds, with birds flying around it and sunlight casting golden rays ...
(3) A realistic documentary-style video of artisans crafting pottery, with the scene unfolding and transforming as hands shape clay ...
(4) A ghost in a white bedsheet faces a mirror. The ghost's reflection can be seen in the mirror. The ghost is in a dusty attic ...



Figure 5. Qualitative comparison between CFG and STG on videos generated by Mochi [33].

We will revisit this assumption and propose an adjustment afterward. Under this independence assumption, the joint distribution can be expressed as

$$p_\theta(x_t|y_b) = p_\theta(x_t|y_{sb})p_\theta(x_t|y_{tb}). \tag{18}$$

Following the same approach as in Eq. 9 and using Eq. 11, we modify the score in Eq. 8 as follows:

$$\begin{aligned}\nabla_{x_t} \log\tilde{p}_\theta(x_t|y_g) = \nabla_{x_t} \log p_\theta(x_t) \\ + w\nabla_{x_t}(\log p_\theta(x_t) - \log p_\theta(x_t|y_s)) \\ + w\nabla_{x_t}(\log p_\theta(x_t) - \log p_\theta(x_t|y_t)).\end{aligned} \tag{19}$$

By replacing the score with the estimated denoiser and utilizing separate scales for spatial and temporal terms, we obtain the following equation:

$$\begin{aligned}\tilde{\epsilon}^w_\theta(x_t) = \epsilon_\theta(x_t) + w_1(\epsilon_\theta(x_t) - \epsilon^{\text{s}}_\theta(x_t)) \\ + w_2(\epsilon_\theta(x_t) - \epsilon^{\text{t}}_\theta(x_t))\end{aligned} \tag{20}$$

where $\epsilon^{\text{s}}_\theta(x_t)$ and $\epsilon^{\text{t}}_\theta(x_t)$ represent spatially and temporally perturbed models, respectively.

Next, we revisit our assumption of independence between spatial and temporal perturbations. While Eq. 20 works well in practice, we can derive an alternative STG formulation that uses orthogonalization to isolate the independent components of the spatial and temporal guidance. Inspired by a negative prompting technique [3], we can modify the Eq. 20 as follows:

$$\begin{aligned}\tilde{\epsilon}^w_\theta(x_t) = \epsilon_\theta(x_t) + w_1\Delta_{\text{s}} \\ + w_2\left(\Delta_{\text{t}} - \frac{\langle\Delta_{\text{s}}, \Delta_{\text{t}}\rangle}{\|\Delta_{\text{s}}\|^2}\Delta_{\text{s}}\right),\end{aligned} \tag{21}$$

where $\Delta_{\text{s}} = \epsilon_\theta(x_t) - \epsilon^{\text{s}}_\theta(x_t)$ and $\Delta_{\text{t}} = \epsilon_\theta(x_t) - \epsilon^{\text{t}}_\theta(x_t)$.

Manifold Constrained Guidance Even with a well-aligned weak model, larger guidance scales w can drive samples off the data manifold, resulting in poor quality and oversaturation. To address error accumulation at high guidance scales, we explore optional techniques to keep samples constrained to the manifold. Rescaling the latent code [21] helps mitigate this issue by constraining its variance, as larger variance is known to cause saturation in the results. Additionally, Restart sampling [36] demonstrates that introducing stochasticity can correct off-manifold deviations. Building on this idea, we incorporate stochastic forward processes into our sampling guidance framework as an optional method. While this approach modestly improves final sample quality and reduces saturation, it introduces additional computational overhead. Further details are provided in Appendix A5.1.

5. Experiments

5.1. Overview

We employ three models for our experiments:

- **Mochi** [33] is a text-to-video model built on AsymmDiT blocks, containing a total of 10 billion parameters and utilizing 3D self-attention in its spatiotemporal layers.
- **Open-Sora** [38] is a text-to-video model built on STDiT blocks with 1.1 billion parameters, employing factorized spatial and temporal attention layers.
- **SVD** [4] is an image-to-video model with 1.5 billion parameters that leverages factorized spatial and temporal attention within a UNet architecture.

We evaluate the proposed method using widely adopted datasets and metrics to ensure a comprehensive performance analysis.

- **UCF-101** The UCF-101 dataset comprises 13,320 videos organized into 101 action classes. Using this dataset, we assess the Fréchet Video Distance (FVD) and Inception Score (IS) of the image-to-video model SVD [4]. Following DIGAN [37], we calculate FVD and IS on 2,048 and 10,000 samples, respectively. For conditioning, initial frames from UCF-101 videos are used as inputs to the SVD model, with each input frame serving as the starting frame of the generated videos.
- **VBench** We use VBench [15] for automatic evaluation across various video metrics. SVD is evaluated using the image-to-video (I2V) framework on 355 samples from the VBench I2V dataset with 5 random seeds. Mochi and Open-Sora are evaluated with the text-to-video (T2V) framework: Open-Sora uses the standard VBench prompt list, while Mochi uses 100 randomly selected prompts due to computational limits.
- **EvalCrafter** We perform human evaluations using 700 prompts from the EvalCrafter dataset [22].
- **LLM-Generated Prompts** We use a set of 100 selected prompts generated by Claude 3.5 Sonnet [2] for our demo and qualitative comparisons.

For evaluation, CFG is used together with STG for our evaluation. We did not use Restart sampling or guidance orthogonalization; additional results with these techniques are in the Appendix sections A5.1 and A5.2.

5.2. Results

Qualitative Comparison Fig. 5 compares CFG and our method. Videos from the naive CFG model often show blurry objects with indistinct shapes, while STG produces clearer, more vivid frames with sharper image quality. STG also reduces temporal inconsistency and flickering, especially in dynamic videos with large motion where CFG frequently fails. Spatial guidance enhances object structure, and temporal guidance improves consistency. Please refer to the Appendix sections A5 and A6 for additional results.

Models	FVD (↓)	IS	Imaging Quality	Aesthetic Quality	Motion Smoothness	Dynamic Degree	Temporal Flickering
Mochi (STG-R)	-	-	**0.628**	**0.554**	**0.988**	0.86	**0.978**
Mochi (STG-A)	-	-	0.555	0.541	0.987	0.86	0.976
Open-Sora (STG-R)	-	-	0.550	0.474	0.981	0.894	**0.977**
Open-Sora (STG-A)	-	-	**0.606**	**0.509**	**0.987**	**0.895**	0.976
SVD (STG-R)	155.9	**39.3**	0.687	0.637	0.965	0.641	-
SVD (STG-A)	**128.7**	38.5	**0.694**	**0.639**	**0.968**	**0.694**	-

Table 3. Comparison of STG-R (residual skip) and STG-A (attention skip) across Mochi [33], Open-Sora [38], and SVD [4]. STG-R shows stronger performance on Mochi, while STG-A yields better results on Open-Sora and SVD.

Models	FVD (↓)	IS	Imaging Quality	Aesthetic Quality	Motion Smoothness	Dynamic Degree
CFG	151.3	38.0	0.687	0.637	0.966	0.562
+ Spatial	133.8	38.3	0.691	**0.639**	0.967	0.659
+ Temporal	**128.7**	**38.5**	**0.694**	0.638	**0.968**	**0.694**

Table 4. Ablation study results on SVD [4] factorized attention, showing the impact of adding spatial and temporal guidance.

Quantitative Comparison We compare CFG and STG using the FVD-Imaging Quality (VBench) plot across different scales in Fig. 4. FVD measures video distribution, while Imaging Quality assesses frame clarity. Higher CFG scales improve Imaging Quality but reduce diversity, as reflected in higher FVD. STG avoids this trade-off, maintaining diversity at increased scales.

T2V and I2V VBench metrics in Tab.1 and Tab.2 show notable improvements in frame-level quality (Imaging and Aesthetic Quality). Temporal quality improves qualitatively, though metrics like Motion Smoothness and Temporal Flickering show marginal gains, as these scores are near saturation (~ 0.9x).

For Dynamic Degree, we aim to keep it unchanged, as it may not correlate with video quality. This is achieved in the T2V metric, but in the I2V model, CFG increases the influence of the conditioned image, reducing motion. STG, while not directly affecting Dynamic Degree, mitigates CFG's impact, leading to increased Dynamic Degree in I2V when used together.

Human Evaluation We provide human evaluation results on 700 prompts from the EvalCrafter dataset [22] in the Appendix A4.

5.3. Ablation Study

Fig. 3 displays selected frames from videos generated by Mochi using various STG scales, where $w = 0$ corresponds to the CFG-only model without STG. Increasing the STG scale results in more vivid colors and finer details compared to the monotonous colors at $w = 0$. Notably, unlike with CFG, sampling diversity is preserved as the STG scale increases, as shown in Fig. 4.

We performed an ablation study on SVD to evaluate spatial and temporal guidance (Tab. 4). Spatial guidance notably reduced FVD and improved metrics, while temporal guidance further boosted performance, confirming their combined effectiveness for spatiotemporal generation.

We also test two STG variants—residual skip (STG-R) and attention skip (STG-A)-on SVD, Open-Sora, and Mochi. As shown in Tab. 3, STG-R performs well with Mochi, while STG-A is more effective for the other models. This is likely due to Mochi's higher parameter count and layer depth, enabling more extensive skipping without triggering out-of-distribution (OOD) issues in consecutive layers. Furthermore, Mochi's spatiotemporal layers consist of a single 3D attention layer, so STG-R skips a residual layer with just one attention layer. In contrast, SVD and Open-Sora use factorized attention, meaning STG-R skips a residual layer with two attention layers, which may cause excessive perturbation and lead to OOD issues in subsequent layers. More results are available in the Appendix A5.

6. Conclusion

We proposed Spatiotemporal Skip Guidance (STG), a simple, training-free method for video diffusion models. By simulating an aligned weak model via spatiotemporal skipping, STG offers strong guidance for high-fidelity video generation, achieving notable qualitative and quantitative improvements. We hope this work advances video diffusion models and inspires further research in the field.

Limitation and Ethical Considerations STG's performance depends on scale and layer selection, with the optimal configuration varying across models, requiring users to set these through heuristic tuning. While video quality improvements are notable, they also raise ethical concerns about misuse, underscoring the importance of using this technology responsibly and constructively.

Acknowledgements This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-00555621).

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. *arXiv preprint arXiv:2403.17377*, 2024. 2, 3, 5, 7

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku. In -, 2024. 7

[3] Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 7

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 4, 5, 7, 8, 1, 9, 11, 13, 14, 17, 18, 23

[5] Andrew Brock. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[6] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024. 2, 3

[7] Zihan Ding, Xiao-Yang Liu, Miao Yin, and Linghe Kong. Tgan: Deep tensor generative adversarial nets for large image generation. *arXiv preprint arXiv:1901.09953*, 2019. 1

[8] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7277–7288, 2024. 1

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 5

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[13] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *arXiv preprint arXiv:2408.00760*, 2024. 2, 3, 5, 8, 18

[14] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023. 3, 5

[15] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 4, 7

[16] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014. 1

[17] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 1

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[19] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024. 2, 3, 5

[20] Black Forest Labs. Flux: Github repository. https://github.com/black-forest-labs/flux, 2024. Accessed: 2024-11-22. 5, 6

[21] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024. 2, 7, 5, 10

[22] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 7, 8, 5, 6

[23] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024. 4

[24] Xinyin Ma, Gongfan Fang, Michael Bi Mi, and Xinchao Wang. Learning-to-cache: Accelerating diffusion transformer via layer caching. *arXiv preprint arXiv:2406.01733*, 2024. 4

[25] Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification. *arXiv preprint arXiv:2307.08702*, 2023. 3, 4

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

[27] 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*, 2024. 2, 5

[28] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020. 4

[29] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[32] 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*, 2020. 2, 3

[33] Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024. 2, 4, 5, 6, 7, 8, 1, 9, 10, 15, 19, 20

[34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[35] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 3

[36] Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. *Advances in Neural Information Processing Systems*, 36:76806–76838, 2023. 2, 7, 5

[37] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022. 7

[38] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. https://github.com/hpcaitech/Open-Sora, 2024. 2, 4, 5, 7, 8, 1, 12, 16, 21, 22