TITAN-Guide: Taming Inference-Time AligNment for Guided Text-to-Video Diffusion Models

Christian Simon[†] Masato Ishii[♣] Akio Hayakawa[♣] Zhi Zhong[†]
Shusuke Takahashi[†] Takashi Shibuya [♣] Yuki Mitsufuji[†,♣]

[†]Sony Group Corporation [♣]Sony AI
{first_name.last_name}@sony.com

In this supplementary material, we present a comprehensive overview of our proposed method, including a detailed explanation of our proposed method and implementation. Additionally, we describe the experimental settings in depth, covering models and parameter configurations. Furthermore, we present additional results and analyses to support our findings.

1. The Details of Forward Gradient Descents

In this section, we provide detailed explanation on using forward AD to estimate gradients. Let $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$. The directional gradient along $\boldsymbol{V}$ evaluated at $\boldsymbol{X}$ can be defined as:

$$f'(\boldsymbol{X}) = \lim_{\delta \rightarrow 0} \frac{f(\boldsymbol{X} + \delta \boldsymbol{V}) - f(\boldsymbol{X})}{\delta}. \quad (1)$$

The forward gradient method approximates using this forward AD method with only a single forward run involved during one iteration. Initially, the estimation of a directional gradient or a gradient guess $\boldsymbol{V}$ is initialized with standard basis vectors. In forward gradient descents [1], the gradient guess $\boldsymbol{V}$can be initialized with random Gaussian noise, allowing for a single forward-mode run instead of multiple runs. In the modern deep learning framework (*e.g.*, PyTorch [2]), this operation can be performed with `torch.func.jvp`.

2. Experimental Settings

Text-to-Video (T2V) Models. In all experiments, we employ AnimateDiff [3] with epiCRealism [1] as the base text-to-image model to generate 16 frames 8fps. We set the denoising process to 20 iterations, as we found this to be sufficient for generating clear videos. Because we work on computing gradients and updating the latent noise, we opt to use FP32 that is more stable in optimization.

[1] https://huggingface.co/emilianJR/epiCRealism

Hyperparameters. For all experiments, we set $\lambda_t = 0.1$. For the sampled gradient guess technique, we set the number of sampled frames $\mathcal{F} = 2$, where we found that this number does not impact on significant memory usage.

Toy experiment. In this experiment, we aim to demonstrate the effectiveness of gradient approximation using forward gradient descent, which offers improved memory efficiency. The toy dataset consists of moon-shaped data, and we employ a diffusion model to fit the generated dataset. Our diffusion model comprises two layers of neural networks, each with a hidden dimension of 64. Each layer includes a linear transformation followed by a ReLU activation function. For the diffusion process, we use Denoising Diffusion Probabilistic Models (DDPM) [4] as the foundational equation to fit the toy dataset.

Details of guidance tasks. Given a text prompt, our guidance diffusion process incorporates three essential tasks: aesthetic score guidance, style guidance, and audio-video alignment. For **aesthetic score guidance**, we build upon the classifier model of DOVER [5], which is designed to enhance the semantic and compositional alignment of generated content. In **style guidance**, we introduce a reference image that serves as a target for guiding the visual style of the generated videos. This approach ensures that the output retains a consistent artistic or cinematic look. To achieve this, we utilize Style-CLIP [6], a powerful model that enables classifier-based guidance for style adaptation. For **multi-modal alignment** (*i.e.*, audio-video alignment), we employ ImageBind [7], which facilitates the alignment of target audio representations with the generated videos. This step is crucial for ensuring that the visual elements correspond appropriately to the accompanying sounds, creating a seamless and natural viewing experience. By leveraging ImageBind, we enhance the coherence between auditory and visual modalities, improving the overall realism and immersive quality of the generated videos. Another task is

frame interpolation with the first and end frames are provided and the model needs to interpolate the frames initialized by the latent noise. For aesthetic score and style guidance tasks, we make use of TFG-1000Prompt dataset [8] and for multi-modal alignment and frame interpolation, we use VGG-Sound [9]. To effectively optimize all these tasks, we design our guidance mechanisms using a loss function based on the cosine similarity between the feature representations of the generated video and the target. This ensures that the generated output closely aligns with the intended aesthetic, style, and multi-modal synchronization. For all experiments, we use 16 frames with 8 fps.

Frame interpolation experiments. Frame interpolation includes filling in missing segments of a video. In our experiment, we retain the first and last frames while reconstructing the intermediate frames. The objective is to generate a video that maintains the continuity and overall quality of the target videos. The task in the guided diffusion process is to produce a clean video:

$$\max_{\boldsymbol{X}_0} p(\boldsymbol{X}_0) = \max_{\boldsymbol{X}_0} \exp(-\|\mathcal{A}(\boldsymbol{X}_0) - \boldsymbol{y}\|), \qquad (2)$$

where $\mathcal{A}(.)$ is an operator for interpolating the missing segments and $\boldsymbol{y}$ is an original video and $\boldsymbol{X}_0$ is the corrupted visual input.

3. Additional Results on Video Generation

In this section, we present our qualitative results, with additional videos available in the supplementary material. We also provide our generated video samples at our project page[2].

Qualitative results. We present qualitative results illustrating the generated samples in ig.A1 andA2. In the frame interpolation task, the model is given only the first and last frames of a video and have to generate the missing intermediate frames. As shown in Fig.A2, our proposed method effectively completes this task, demonstrating the capability in temporal consistency. Additionally, for style guidance, we showcase two distinct visual styles in Fig.A1, further highlighting the efficacy of our approach. We also provide generated video results of 384×384 resolution in Fig. A3 and A4.

Time and memory consumption. Below, we discuss the time require to process a video on 16 frames with 256×256 resolution. We use H200 to evaluate the processing time. TITAN-Guide requires 2 minutes to perform guidance in diffusion models, and DOODL [10] requires a similar time with ours. As we know the other methods *e.g.*, TFG [8], MPGD [11], FreeDoM [12], Seeing and Hearing [13] do not traverse to $t = 0$, they require about 40 seconds for video generation.

[2]Project page: https://titanguide.github.io/

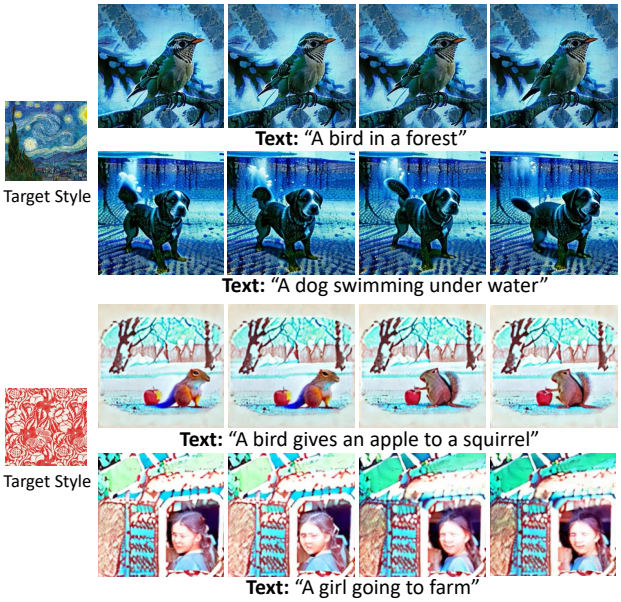

Figure A1. Qualitative results of TITAN-Guide on style guidance.

Figure A2. Qualitative results of TITAN-Guide on frame interpolation. The given frames are indicated by red and interpolated frames are in the between these two frames.

4. Additional Results on Image Generation

We also demonstrate that our proposed approach has the potential to be effectively applied to the image diffusion models [4, 14]. However, since our primary focus is text-to-video tasks, we do not explore this aspect in depth but

Method	Super Resolution		CelebA		Deblurring	
	LPIPS↓	FID↓	Acc.↑	KID↓	LPIPS↓	FID↓
FreeDoM [12]	0.191	74.5	68.7	3.89	0.245	87.4
MPGD [11]	0.283	82.0	68.6	4.79	0.177	69.3
TFG [8]	0.190	**65.9**	75.2	3.86	**0.150**	64.5
TITAN-Guide (ours)	**0.180**	75.9	**77.2**	**3.61**	0.222	**63.5**

Table A1. Evaluation results across various metrics for assessing the quality of generated images.

provide evidence of its applicability.

Settings. In all image generation experiments, we use 256×256 image resolution. We use three different tasks: 1) Super resolution, 2) CelebA (guided by gender and age specification), and deblurring. Following [8], for super resolution, and deblurring tasks, we use the CAT-DDPM diffusion model trained on the CAT dataset [15]. While, we use CelebA-DDPM trained on the CelebA dataset [16] for gender[3] and age[4] guidance.

Evaluation. For the experiment in image domain, we evaluate based on Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Image Distance (FID) to assess the quality of generated images. For the CelebA (gender and age) task, we measure classification accuracy (Acc.) and Kernel Inception Distance (KID) to to assess fidelity of generated samples. For all experiments, we generate 256 images to evaluate the effectiveness of our proposed method.

Results. Table A1 shows the results of the three tasks. In this experiment, we use TITAN-Guide exclusively with sampled gradient guesses. Our observations indicate that TITAN-Guide outperforms previous methods in most tasks and metrics. Additionally, the image generation results demonstrate its effectiveness in image guidance tasks.

References

[1] A. G. Baydin, B. A. Pearlmutter, D. Syme, F. Wood, and P. Torr, "Gradients without backpropagation," *arXiv preprint arXiv:2202.08587*, 2022. 1

[2] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017. 1

[3] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *International Conference on Learning Representations*, 2024. 1

[3]Gender: https://huggingface.co/rizvandwiki/gender-classification

[4]Age: https://huggingface.co/londe33/hair_v0

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. 1, 2

[5] H. Wu, E. Zhang, L. Liao, C. Chen, J. H. Hou, A. Wang, W. S. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *International Conference on Computer Vision (ICCV)*, 2023. 1

[6] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2085–2094. 1

[7] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *CVPR*, 2023. 1

[8] H. Ye, H. Lin, J. Han, M. Xu, S. Liu, Y. Liang, J. Ma, J. Zou, and S. Ermon, "Tfg: Unified training-free guidance for diffusion models," 2024. 2, 3

[9] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 2

[10] B. Wallace, A. Gokul, S. Ermon, and N. Naik, "End-to-end diffusion latent optimization improves classifier guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7280–7290. 2

[11] Y. He, N. Murata, C.-H. Lai, Y. Takida, T. Uesaka, D. Kim, W.-H. Liao, Y. Mitsufuji, J. Z. Kolter, R. Salakhutdinov, and S. Ermon, "Manifold preserving guided diffusion," in *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[12] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang, "Freedom: Training-free energy-guided conditional diffusion model," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[13] Y. Xing, Y. He, Z. Tian, X. Wang, and Q. Chen, "Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners," in *CVPR*, 2024. 2

[14] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020. 2

[15] J. Elson, J. R. Douceur, J. Howell, and J. Saul, "Asirra: a captcha that exploits interest-aligned manual image categorization," in *Conference on Computer and Communications Security*, 2007. 3

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 3

Figure A3. Qualitative results of TITAN-Guide for audio-video alignment at 384×384 resolution. Top: Elephants in water accompanied by water surface sounds. Bottom: Chickens clucking in sync with the corresponding clucking sound.

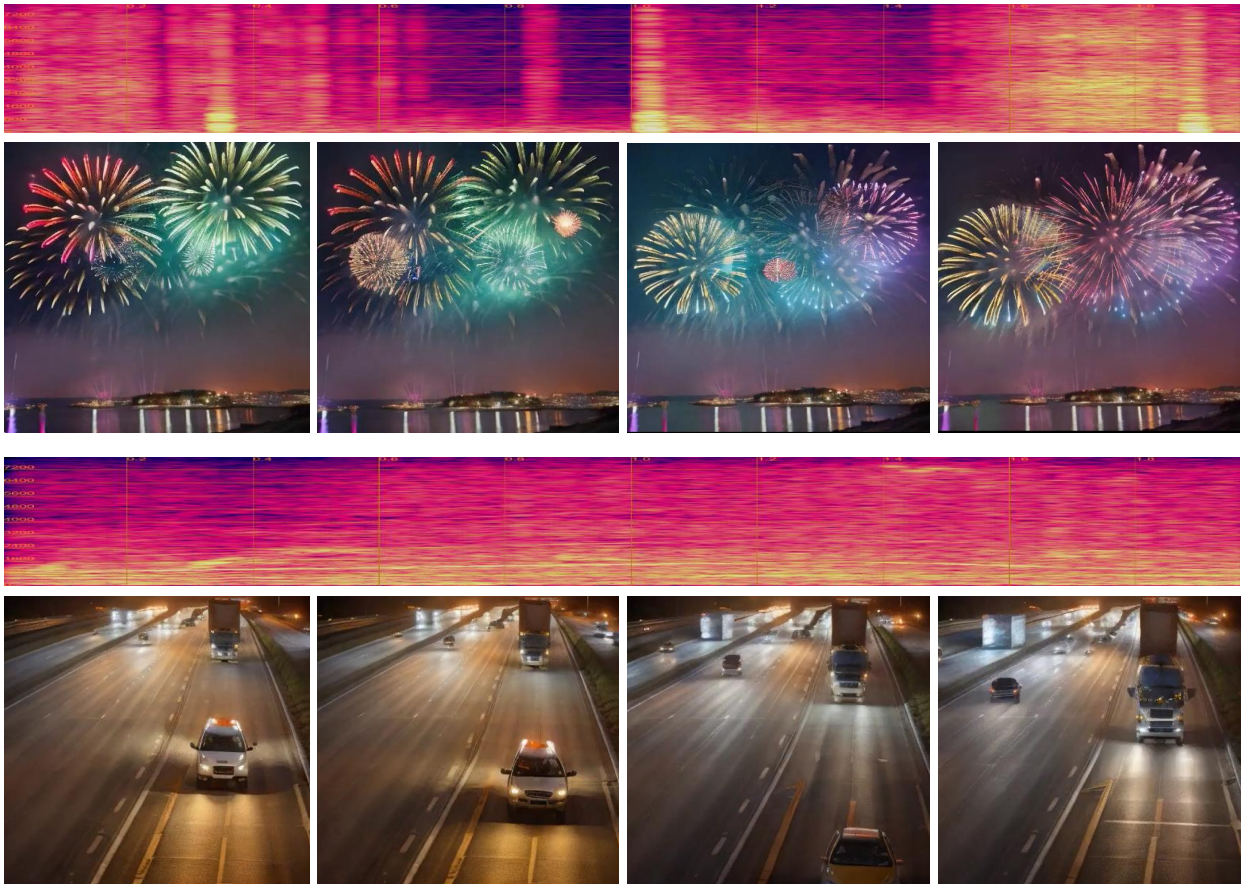

Figure A4. Qualitative results of TITAN-Guide for audio-video alignment at 384×384 resolution. Top: City fireworks synchronized with firework sounds. Bottom: Highway cars accompanied by accelerating engine sounds.