VideoGuide: Improving Video Diffusion Models without Training Through a Teacher's Guide

Supplementary Material

The supplementary material is organized as follows:

- Section A: Importance of the low-pass filter.
- Section B: About classifier-free guidance.
- Section C: Pseudocodes for our algorithm.
- Section D: More experimental details.
- Section E: More quantitative results: user study.
- Section F: Comparison with orthogonal methods.
- Section G: Application on DiT-based models.
- Section H: Application on I2V tasks and cross-tasks.
- Section I: Limitations.
- Section J: More qualitative examples.

A. Importance of Low-Pass Filter

To evaluate the role of the low-pass filter in our methodology, we conduct experiments by varying the interpolation step I, both with and without the low-pass filter. These experiments are averaged across 800 prompts from the VBench categories for consistent evaluation. We apply the low-pass filter for the initial 5 timesteps, based on the observation that the mid-to-late timesteps in the diffusion process focus on generating mid- and high-frequency details. Replacing these frequencies with random components via the low-pass filter in the mid-to-late timesteps would result in degraded video quality, necessitating the early-timestep limitation. All corresponding results are presented in Fig. 5.

We measure the effects of the filter on Subject Consistency, Background Consistency, and Imaging Quality. Both Subject Consistency and Background Consistency steadily improves as the number of interpolation steps increases, demonstrating the effectiveness of our interpolation scheme in enhancing temporal coherence. Meanwhile, Imaging Quality is maintained up to approximately 10 interpolation steps without the low-pass filter. Beyond this point, a significant drop in quality is observed, indicating that excessive interpolation exacerbates the blurring effects caused by prolonged SDS optimization, as noted earlier in this work.

The improvement in consistency is significantly accelerated when using the low-pass filter. This acceleration is achieved while mitigating decline in imaging quality typically associated with increased interpolation steps. Furthermore, application of the filter reduces computational overhead during interpolation. Specifically, consistency achieved at $I = 4$ with the filter is comparable to consistency achieved at $I = 50$ without the filter, offering approximately a 7-fold reduction in inference time. Such results prove the effectiveness of the filter in balancing consistency improvement, imaging quality preservation, and computational cost.

B. Classifier-Free Guidance

Off-Manifold Behavior of CFG Recent study [7] demonstrates that employing a high Classifier- Free Guidance (CFG) [16] scale ($w > 1.0$) in the early timesteps of diffusion sampling leads to off-manifold behavior. This phenomenon results in denoised samples exhibiting problems such as color saturation and abrupt transitions, which negatively affect the interpolation between samples during these timesteps. We solve this by applying a lower guidance scale w during the early stages of sampling, ensuring smoother interpolation between the denoised samples. As illustrated in Fig. 6 (a), when using a high CFG scale ($w = 7.5$), the influence of the guiding diffusion model becomes minimal due to significant color saturation, making it difficult for the output of the guiding model to be reflected effectively. In contrast, as illustrated in Fig. 6 (b), a lower CFG scale ($w = 0.8$) facilitates smoother interpolation between the sampling diffusion model and the guiding diffusion model.

Configuration	SC (↑)	BC (↑)
Base		
+ CFG	0.9183	0.9437
+ CFG++	0.9176	0.9435
FreeInit		
+ CFG	0.9487	0.9604
+ CFG++	0.9473	0.9604
Ours		
+ CFG Interp.	0.9598	0.9635
+ CFG++ Interp.	0.9614	0.9664

Table 5. Comparison of consistency metrics (**SC**: Subject Consistency, **BC**: Background Consistency) across different configurations using CFG and CFG++ in AnimateDiff. Our approach with interpolated CFG++ achieves the best performance, significantly enhancing both subject and background consistency.

We provide quantitative analysis for using CFG and CFG++ across the Base Model, Base Model + FreeInit, and Base Model + VideoGuide (Ours) during the interpolation. As shown in Tab. 5, metrics for Base and FreeInit decrease when CFG++ is used, and metrics improve only when CFG++ is applied to our interpolation scheme. This implies the significant positive impact on consistency of CFG++ within the proposed interpolation scheme, especially compared to CFG. Also, this supports the idea that smooth interpolation of denoised samples positively impacts model performance, as discussed above.

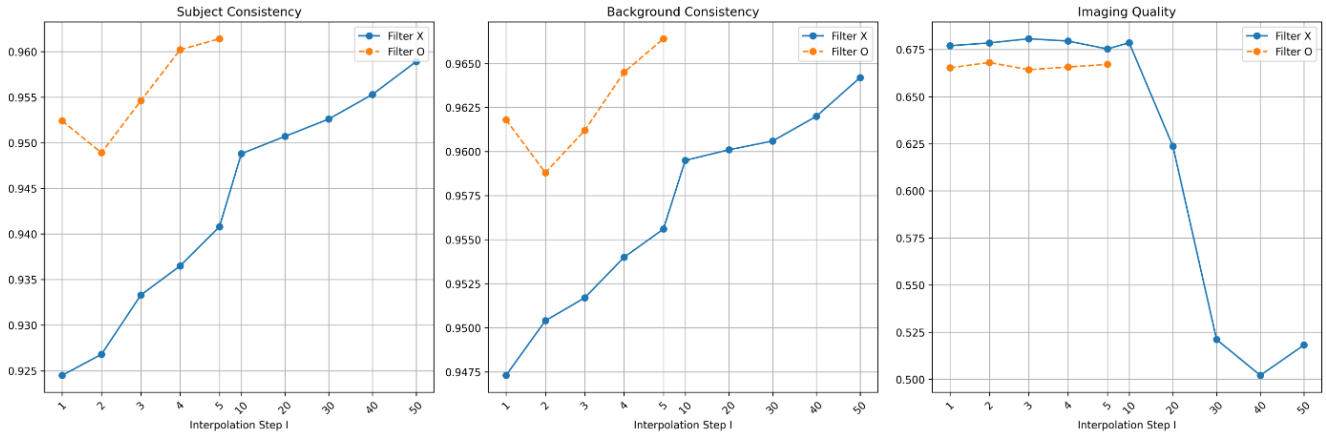


Figure 5. Comparison of Subject Consistency, Background Consistency, and Imaging Quality across interpolation steps (I) with and without the application of the low-frequency filter. Results indicate that the low-frequency filter accelerates convergence towards consistency while maintaining imaging quality.

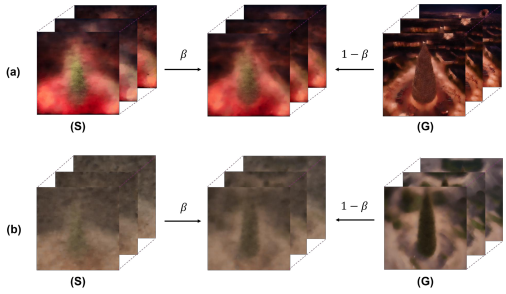


Figure 6. (a) The interpolation process between denoised samples from the sampling model (S) and the guiding model (G) for high guidance scale $w = 7.5$ is shown. (b) The interpolation process for low guidance scale $w = 0.8$ is shown. Both interpolations are performed at $T = 980$ and $\beta = 0.7$. Results indicate that with high guidance scale w, influence of the guiding diffusion model is significantly reduced due to color saturation.

C. Pseudocodes

Pseudocodes regarding our algorithm are provided. For clarity, the pseudo code describing our algorithm adopts the CFG++ reverse sampling framework for the entire process.

D. Experimental Details

D.1. Prompt Selection

In all experiments, we utilize 800 prompts from various categories in VBench [19] to evaluate the model's ability to generate across diverse categories.

D.2. Hyperparameter Selection

We employ a classifier-free guidance (CFG) scale of 7.5 during inference for both base models (AnimateDiff, LaVie) and FreeInit-applied cases. During interpolation of the denoised samples, we apply CFG++ reverse sampling with a guidance scale of $w = 0.8$ in DDIM 50-step sampling. After completing the interpolation step, we revert to CFG reverse sampling with a CFG scale of 7.5. In FreeInit, we use a Butterworth filter with a normalized frequency of 0.25, filter order $n = 4$, and perform 5 iterations, as recommended in prior work. The same filter is applied in our experiments with FreeInit. For AnimateDiff, we configure the guiding model with parameters $I = 5$, $\beta = 0.5$, and $\tau = 10$. In the case of LaVie, we set $I = 3$, $\beta = 0.5$, and $\tau = 10$ to optimize inference speed. Additionally, the τ intervals are not uniformly spaced as in the standard 50-step DDIM sampling. To better leverage temporally consistent samples, we divide the remaining interval into 25 steps for reverse sampling during guidance steps.

D.3. Figure Explanation

Base models used for Figure 3:
(a) AnimateDiff with pretrained T2I model RealisticVision.
(b) AnimateDiff with pretrained T2I model RealisticVision.
(c) LaVie.
(d) LaVie.
Base model used for Figure 4:
AnimateDiff with pretrained T2I model ToonYou.

E. User Study

We conduct a user study to evaluate generated video samples using three criteria: **Text Alignment**, **Overall Quality**, and **Smooth and Dynamic Motion**, with all metrics scored on a 1 to 5 scale. A total of 30 participants provided ratings for each metric, offering comprehensive feedback on the generated videos. Tab. 6 shows that our method surpasses the baseline and previous work in all evaluated aspects.
Text Alignment
- Measures how well the video corresponds to the prompt,

Algorithm 1 VideoGuide with Sampling Diffusion Model

Require: guidance scale $\lambda \in [0, 1]$, guiding steps I, interpolation scale β, extra step τ

Initialize $\boldsymbol{z}_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{z}_t, t) = \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \phi) + \lambda[\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, c) - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \phi)]$
  $\boldsymbol{z}_{0|t} = (\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{z}_t, t))/\sqrt{\bar{\alpha}_t}$
  $\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{0|t} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad where \quad \boldsymbol{\epsilon} \sim N(0, \mathbf{I})$
  if $T - t < I$ **then**
    for $j = 0, \ldots, \tau$ **do**
      $\boldsymbol{z}_{t-j-1} = \sqrt{\bar{\alpha}_{t-j-1}}\boldsymbol{z}_{0|t-j} + \sqrt{1-\bar{\alpha}_{t-j-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_{t-j}, t-j, \phi)$
    end for
    $\boldsymbol{z}'_{0|t} = \beta \cdot \boldsymbol{z}_{0|t} + (1-\beta) \cdot \boldsymbol{z}_{0|t-\tau}$
    $\boldsymbol{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{z}'_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \phi)$
    $\boldsymbol{z}_{t-1} = \mathit{LPF}_\gamma(\boldsymbol{z}_{t-1}) + \mathit{HPF}_\gamma(\boldsymbol{\epsilon}), \quad where \quad \boldsymbol{\epsilon} \sim N(0, \mathbf{I})$
  else
    $\boldsymbol{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{z}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \phi)$
  end if
end for
Output: Final video $\boldsymbol{z}_0$

Algorithm 2 VideoGuide with Guiding Diffusion Model

Require: guidance scale $\lambda \in [0, 1]$, guiding steps I, interpolation scale β, extra step τ, Guiding Model G parameterized by ψ, noise schedule $\bar{\alpha}^{(G)}$ of G

Initialize $\boldsymbol{z}_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{z}_t, t) = \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \phi) + \lambda[\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, c) - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \phi)]$
  $\boldsymbol{z}_{0|t} = (\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{z}_t, t))/\sqrt{\bar{\alpha}_t}$
  $\boldsymbol{z}_t^{(G)} = \sqrt{\bar{\alpha}_t^{(G)}}z_{0|t} + \sqrt{1-\bar{\alpha}_t^{(G)}}\boldsymbol{\epsilon}, \quad where \quad \boldsymbol{\epsilon} \sim N(0, \mathbf{I})$
  if $T - t < I$ **then**
    for $j = 0, \ldots, \tau$ **do**
      $\boldsymbol{z}_{0|t-j}^{(G)} = (\boldsymbol{z}_{t-j}^{(G)} - \sqrt{1-\bar{\alpha}_{t-j}^{(G)}}\hat{\boldsymbol{\epsilon}}_\psi(\boldsymbol{z}_{t-j}^{(G)}, t-j)/\sqrt{\bar{\alpha}_{t-j}^{(G)}}$
      $\boldsymbol{z}_{t-j-1}^{(G)} = \sqrt{\bar{\alpha}_{t-j-1}^{(G)}}\boldsymbol{z}_{0|t-j}^{(G)} + \sqrt{1-\bar{\alpha}_{t-j-1}^{(G)}}\boldsymbol{\epsilon}_\psi(\boldsymbol{z}_{t-j}^{(G)}, t-j, \phi)$
    end for
    $\boldsymbol{z}'_{0|t} = \beta \cdot \boldsymbol{z}_{0|t} + (1-\beta) \cdot \boldsymbol{z}_{0|t-\tau}^{(G)}$
    $\boldsymbol{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{z}'_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \phi)$
    $\boldsymbol{z}_{t-1} = \mathit{LPF}_\gamma(\boldsymbol{z}_{t-1}) + \mathit{HPF}_\gamma(\boldsymbol{\epsilon}), \quad where \quad \boldsymbol{\epsilon} \sim N(0, \mathbf{I})$
  else
    $\boldsymbol{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{z}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \phi)$
  end if
end for
Output: Final video $\boldsymbol{z}_0$

focusing on semantic coherence.

- Question: Do you think the videos reflect the given text condition well? (5: Strongly Agree / 4: Agree / 3: Neutral / 2: Disagree / 1: Strongly Disagree)

Overall Quality

- Assesses the video's visual consistency, image degradation, and aesthetic appeal.
- Question: Do you think the video's overall quality is good? (rich detail, unchanging objects) (5: Strongly Agree / 4: Agree / 3: Neutral / 2: Disagree / 1: Strongly Disagree)

Smooth and Dynamic Motion

- Evaluates the naturalness and fluidity of the motion in the video.
- Question: Do you think the video's overall motion is

smooth and dynamic? (5: Strongly Agree / 4: Agree / 3: Neutral / 2: Disagree / 1: Strongly Disagree)

Method	TA	OQ	SDM
Base	3.72	2.84	2.9
Base + FreeInit	<u>3.97</u>	<u>3.35</u>	<u>3.38</u>
Base + VideoGuide (Ours)	**4.36**	**4.37**	**4.36**

Table 6. User Study. Text Alignment (TA), Overall Quality (OQ), Smooth and Dynamic Motion (SDM) are evaluated among methods. **Bold**: best, <u>underline</u>: second best.

F. Comparison with Orthogonal Methods

A recent study, UniCtrl [4], attempts to improve semantic consistency and motion quality in an approach orthogonal to ours. In this section, we compare the performance of each technique and assess the feasibility of combining them. Following the recommendation in the paper, we use a motion injection degree of $c = 0.2$, while maintaining the same experimental configuration as described in Section D. As illustrated in Tab. 7, UniCtrl [4] improves temporal consistency but at the cost of a significant reduction in dynamic degree and imaging quality.

G. Application on DiT-based Models

We further evaluate the robustness of our methodology by applying it to different architectures and schedulers. Specifically, we present further evaluation on models that use Diffusion Transformer (DiT) [26] architecture: Open-Sora v1.0 [39] and Open-Sora v1.2 [39]. Each model employs a standard DDIM scheduler (50 steps) and a rectified flow [25] scheduler, respectively. In the rectified flow-based configuration, the objective for training can be formulated as follows:

$$\begin{aligned} \boldsymbol{z}_t &= (1-t)\boldsymbol{z}_0 + t\boldsymbol{\epsilon} \quad where \quad t \in [0,1] \\ \hat{\theta} &= \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}\left[||(\boldsymbol{z}_0 - \boldsymbol{\epsilon}) - v_\theta(\boldsymbol{z}_t, t)||_2^2\right] \end{aligned} \tag{17}$$

Using the objective above we can redefine our method as below:

$$\begin{aligned} \boldsymbol{z}_{0|t_i} &= \boldsymbol{z}_{t_i} + t_i \cdot v_\theta(\boldsymbol{z}_{t_i}, t_i) \\ \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_{t_i}, t_i) &= \boldsymbol{z}_{t_i} - (1-t_i) \cdot v_\theta(\boldsymbol{z}_{t_i}, t_i) \\ \boldsymbol{z}_{t_{i-1}} &= (1-t_{i-1}) f(\boldsymbol{z}_{0|t_i}, \beta, \tau) + t_{i-1}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_{t_i}, t_i) \end{aligned} \tag{18}$$

where $f(\boldsymbol{z}_{0|t_i}, \beta, \tau)$ is the interpolation function between $\boldsymbol{z}_{0|t_i}$ and $\boldsymbol{z}_{0|t_{i-\tau}}$ with scale β. The results in Tab. 8 demonstrate that our method improves temporal consistency for both baselines while preserving imaging quality and introducing only a minimal decrease in dynamic degree. These findings indicate that our method enhances performance regardless of the underlying architecture and scheduler.

H. Application on I2V Tasks and Cross-Tasks

Application to the I2V task by using VideoGuide self-guidance on DynamiCrafter [36] is shown in Tab. 9. Quantitative comparison is done on 355 image prompts, and the metric Video-Image Alignment (DINO feature similarity between the given image and generated video) is used in addition to the previous metrics.

Moreover, we explore cross-task functionality in student T2V and teacher I2V models. Fig. 10 shows a specific example in which the student T2V model is VideoCrafter2 and the teacher I2V model is DynamiCrafter with null text input. Our method guides VideoCrafter2 to creating samples that resemble that of the image input into DynamiCrafter *even though the image is never explicitly shown to VideoCrafter2*. Thus, VideoGuide can enable T2V generation to operate under auxiliary image awareness, which opens up new avenues such as usage of VideoGuide in image editing.

I. Limitations

While our approach significantly improves the performance of baseline models, it relies on sharing the same Variational Auto-Encoder (VAE) [23] space. In practice, many latent diffusion models utilize the same VAE, making this requirement generally feasible. However, if the VAE spaces differ, one potential solution is to decode, interpolate, and re-encode the features. This process, however, incurs additional computational overhead and risks losing fine details due to iterative encoding-decoding. Developing an effective method to address compatibility across different VAE spaces remains an avenue for future research.

J. More Qualitative Examples

Additional samples are provided in following pages:

- Supplemental examples of prior distillation.
- Qualitative comparison for various methods.
- Qualitative comparison for various base models.
- Usage of VideoGuide to solve sudden frame shifts in LaVie samples.

J.1. Prior Distillation

J.2. More Qualitative Comparison Results

J.3. More Qualitative Results

J.4. LaVie Sudden Shift

Method	Subject Consistency (↑)	Background Consistency (↑)	Imaging Quality (↑)	Motion Smoothness (↑)	Dynamic Degree (↑)
AnimateDiff [13]	0.9183	0.9437	0.6647	0.9547	26.67
AnimateDiff + UniCtrl [4]	0.9259	0.9413	0.6032	0.9584	14.96
AnimateDiff + Ours	0.9614	0.9664	0.6671	0.9772	16.78
AnimateDiff + UniCtrl + Ours	0.9639	0.9628	0.5883	0.9776	5.02

Table 7. Quantitative comparison with orthogonal methods.

Method	Subject Consistency (↑)	Background Consistency (↑)	Imaging Quality (↑)	Motion Smoothness (↑)	Dynamic Degree (↑)
OpenSora v1.0 [39] (DDIM [30])	0.9735	0.9689	0.6615	0.9678	4.97
OpenSora v1.0 + VideoGuide (self-guided)	0.9763	0.9689	0.6738	0.9754	3.88
OpenSora v1.2 [39] (Rectified Flow [25])	0.9725	0.9696	0.6582	0.9881	12.68
OpenSora v1.2 + VideoGuide (self-guided)	0.9808	0.9748	0.6689	0.9903	11.07

Table 8. Quantitative comparison of video generation in DiT-based architecture.

Method	Subject Consistency (↑)	Background Consistency (↑)	Imaging Quality (↑)	Motion Smoothness (↑)	Dynamic Degree (↑)	Video-Image Alignment (↑)
DynamiCrafter [36]	0.9663	0.9644	0.7042	0.9839	3.93	0.9535
DynamiCrafter + VideoGuide (self-guided)	0.9681	0.9654	0.7065	0.9840	3.94	0.9553

Table 9. Quantitative comparison for the self-guided I2V task.

Figure 7. Qualitative Comparison of UniCtrl and VideoGuide.

Figure 8. Qualitative Results of VideoGuide on Open-Sora v1.0.

Figure 9. Qualitative Results of VideoGuide on Open-Sora v1.2.

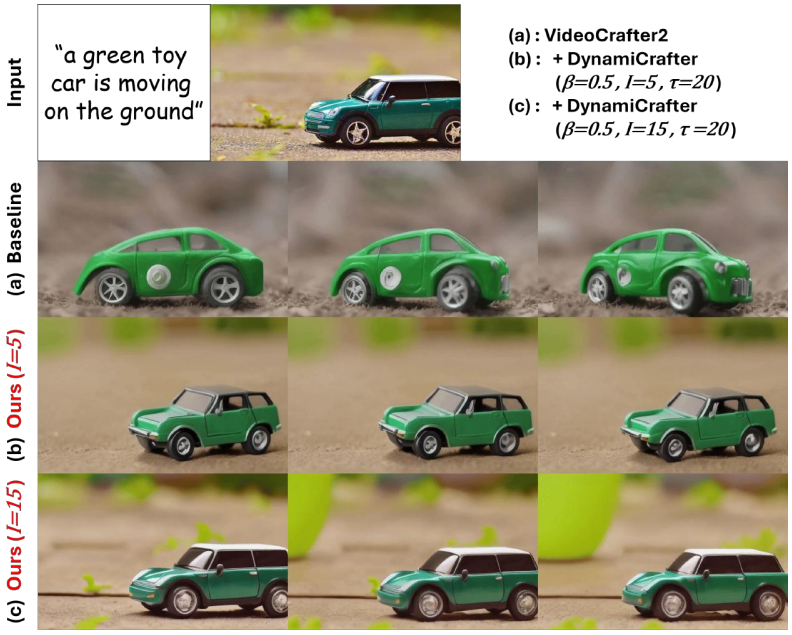


Figure 10. Cross-task functionality of student T2V and teacher I2V models.

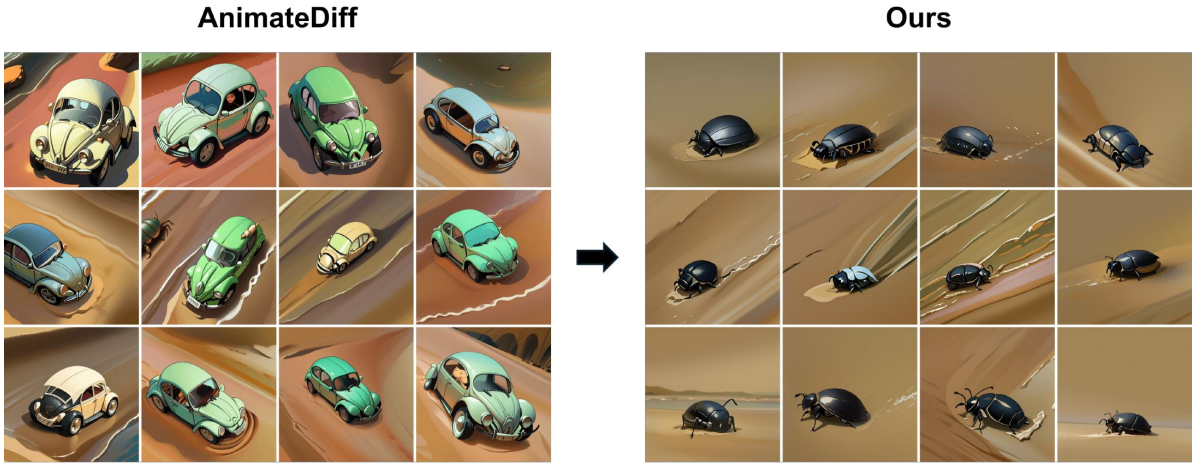
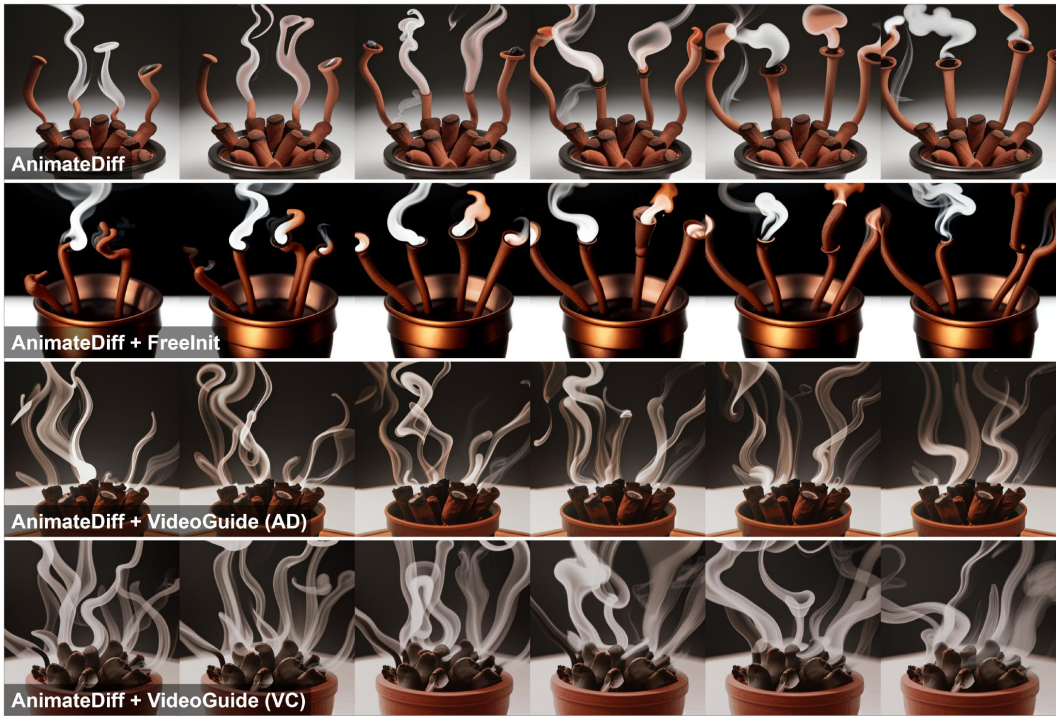


Figure 11. **Prior Distillation**. For each prompt, we share the same random seed for both methods.

"A car accelerating to gain speed"

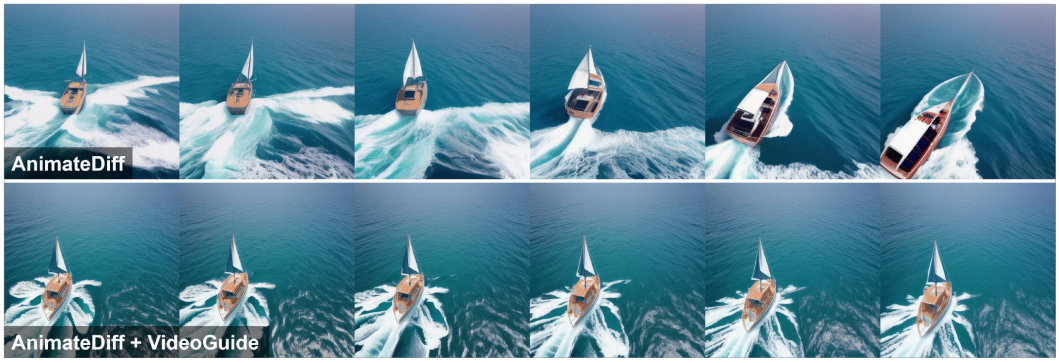


"Ashtray full of butts on table, smoke flowing on black background, close-up"

Figure 12. More Qualitative Comparison Results of VideoGuide. Top: AnimateDiff with ToonYou, Bottom: AnimateDiff with RCNZCartoon

"An airplane flying above the sea of clouds"

"Couple salsa dancing"



"Boat sailing in the middle of ocean"

"Curious cat sitting and looking around"

Figure 13. More Qualitative Results of VideoGuide on AnimateDiff (with RealisticVision).

"Slow motion footage of a racing car"

"A male vendor selling fruits"

"A dog drinking water"

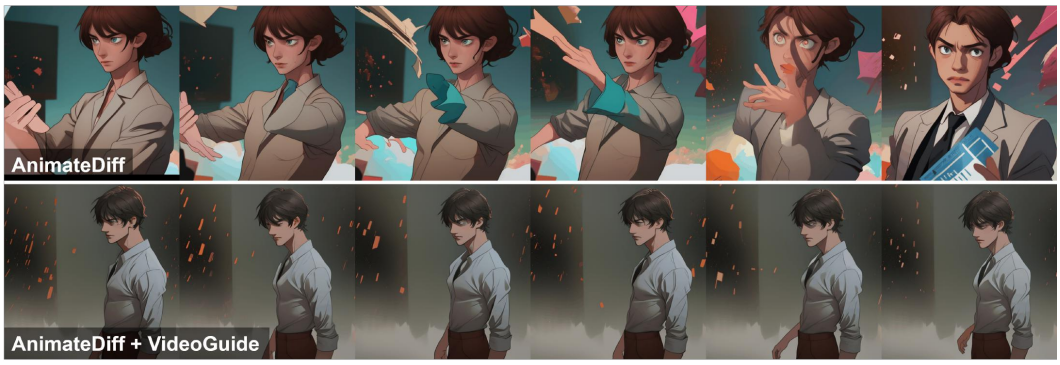
"A bear wearing red jersey"

Figure 14. More Qualitative Results of VideoGuide on AnimateDiff (with RealisticVision).

"Silhouette of the couple during sunset"

"Traffic in London street at night"

"A cute Pomeranian dog playing with a soccer ball"

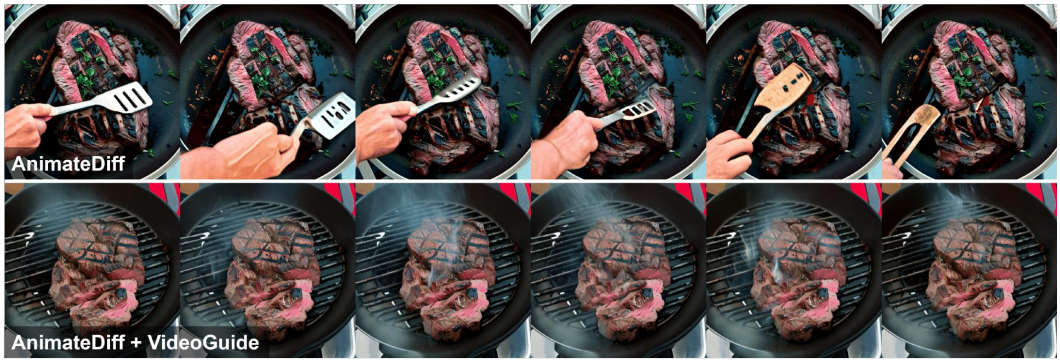


"A footage of actor movie scene"

Figure 15. More Qualitative Results of VideoGuide on AnimateDiff (with ToonYou).

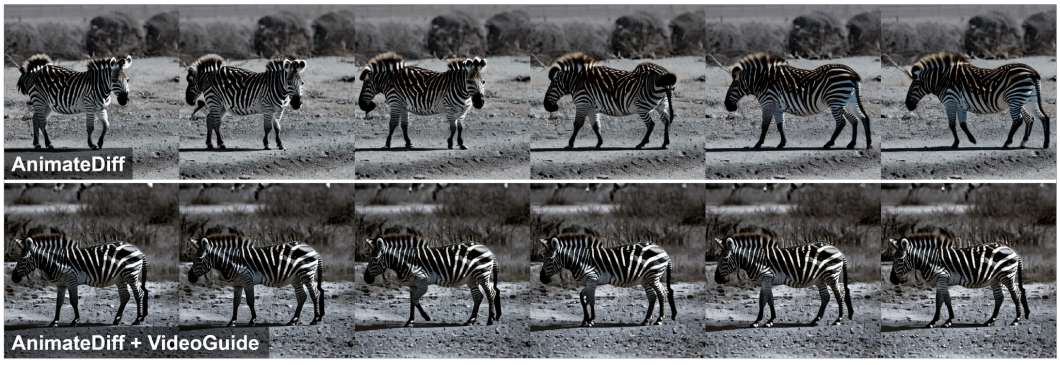
Figure 16. More Qualitative Results of VideoGuide on AnimateDiff (with RCNZCartoon).

"Dark clouds overshadowing the full moon"



"Grilling a steak on a pan grill"

"Fighter practice kicking"



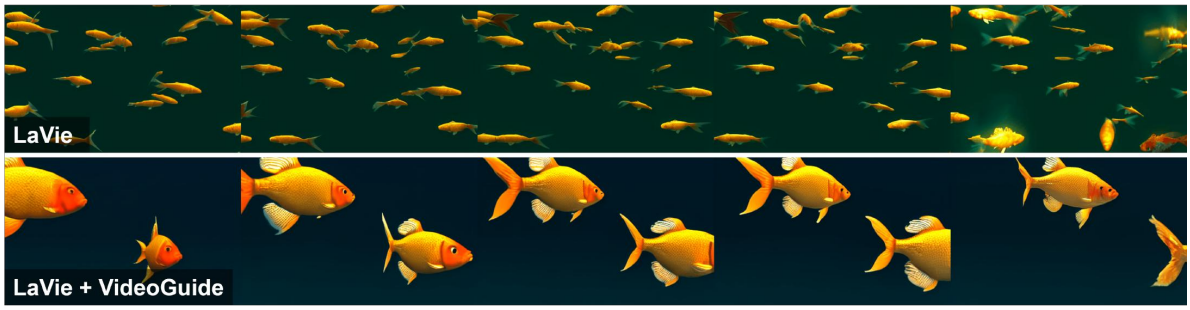
"A zebra taking a peaceful walk"

Figure 17. More Qualitative Results of VideoGuide on AnimateDiff (with FilmVelvia).

"Kid in a Halloween costume"

"A storm trooper vacuuming the beach"

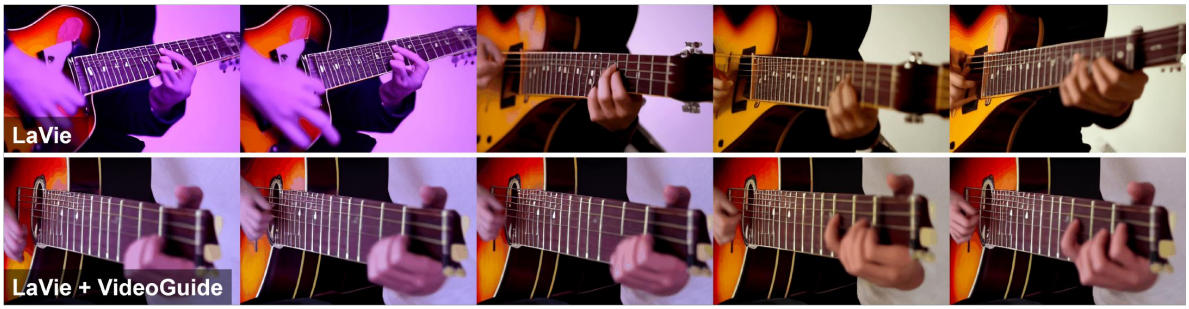
"Deer grazing in the field"



"Golden fish swimming in the ocean"

Figure 18. More Qualitative Results of VideoGuide on LaVie.

"A dog running happily"

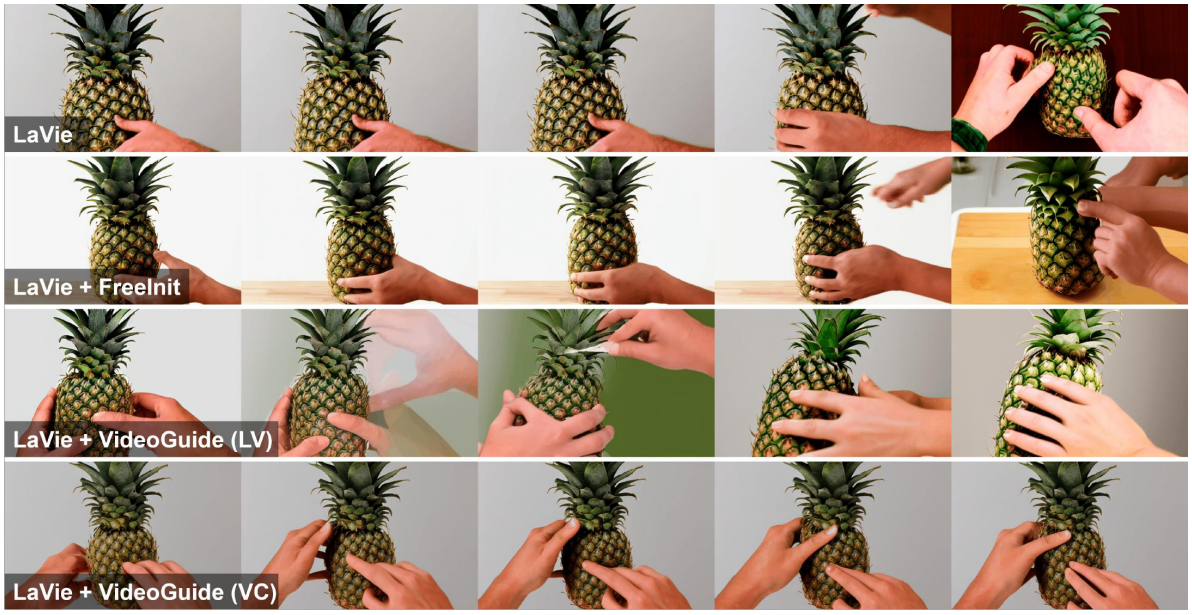


"A person playing guitar"

"Men loading Christmas tree on tow truck"

"A koala bear playing piano in the forest"

Figure 19. More Qualitative Results of VideoGuide on LaVie.



"Removing a pineapple leaf"

"Kids celebrating Halloween at home"

Figure 20. VideoGuide helps solve the issue of sudden frame shifts in LaVie samples. By integrating an external guiding model, VideoGuide provides smoother frame transitions to the base model. LV indicates that guidance model of LaVie is used (the self-guided case), and VC indicates that guidance model of VideoCrafter2 is used. Guidance given with the external model VideoCrafter2 solves sudden frame shift unsolvable by other methods.