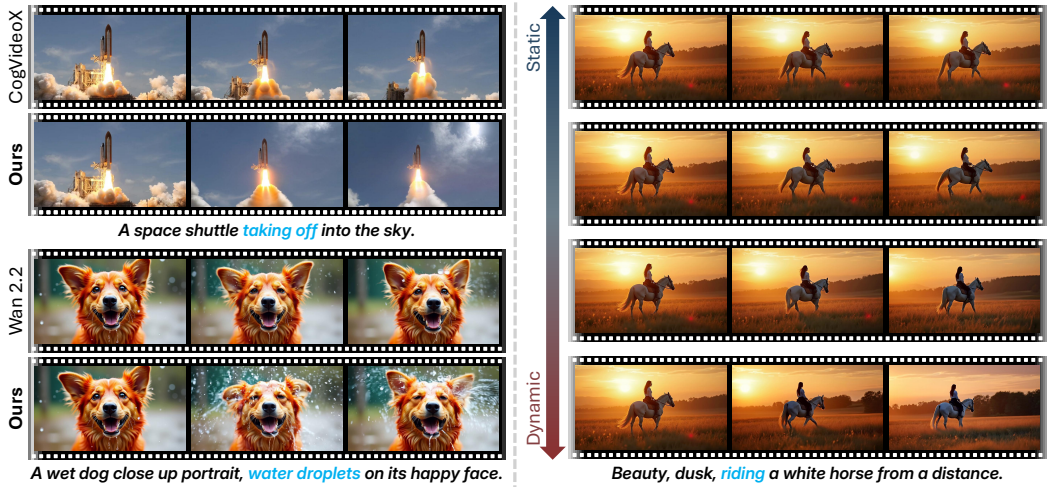
Rebalancing Reference Frame Dominance to Improve Motion in Image-to-Video Models

Wooseok Jeon[1]* **Seungho Park**[1]* **Seunghyun Shin**[2] **Sangeyl Lee**[1]
Hyeonho Jeong[3] **Hae-Gon Jeon**[1]†
[1]Yonsei University [2]GIST [3]Adobe Research



(a) Dynamic image-to-video generation (b) Continuous control over motion dynamics

Figure 1: **Example videos from our method.** We present DyMoS, a training-free and model-agnostic method for improving motion dynamics in image-to-video generation. (a) Comparison of generated videos from the same input image. DyMoS produces dynamic motion while preserving video quality. (b) Furthermore, DyMoS provides continuous control over motion dynamics.

Abstract

Image-to-video models often generate videos that remain overly static compared to text-to-video models. While prior approaches mitigate this issue by weakening or modifying the image-conditioning signal, they often require additional training or sacrifice fidelity to the reference image. In this work, we identify *reference-frame dominance* as a key mechanism behind motion suppression. We observe that non-reference frames in I2V models allocate excessive self-attention to reference-frame key tokens, causing reference information to be over-propagated across time and suppressing inter-frame dynamics. Based on this finding, we propose DyMoS (Dynamic Motion Slider), a training-free and model-agnostic method that rebalances the attention pathway from generated frames to the reference frame during initial denoising steps. DyMoS leaves both the input image and model weights unchanged and introduces a single scalar parameter for continuous control over motion strength. Experiments across multiple state-of-the-art I2V backbones demonstrate that DyMoS consistently improves motion dynamics while maintaining visual quality and fidelity to the reference image. Project page: https://i2v-dymos.github.io/

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Recent advances in generative models have been driven by the remarkable success of diffusion [1–4] and flow-matching [5–8] frameworks. Building on these paradigms, text-to-video (T2V) models [9–15] have made it possible to generate high-quality videos from text prompts. Image-to-video (I2V) models [9–12, 16–23] extend this capability by conditioning generation on reference images[1], enabling the creation of videos that follow text prompts while preserving the content and appearance of the input. These models are typically adapted from large-scale T2V backbones by introducing an additional image-conditioning pathway, which grounds the generation in specific visual input while retaining the motion priors learned from large-scale video data.

Despite this progress, I2V models often produce videos that remain overly static compared to their T2V counterparts, even when the prompt explicitly describes motion. Prior works address this static motion bias by either modifying the image-conditioning pathway or weakening the conditioning signal during generation [24–26]. While these approaches can encourage stronger motion, they often require additional training or involve direct modification of the input signal. In the latter case, improved motion dynamics are achieved at the cost of fidelity to the reference image.

In this work, we revisit the static motion bias by analyzing how reference image information is utilized within I2V models. Since I2V models process spatiotemporal latent tokens through self-attention, their attention maps allow us to analyze how information flows across frames during the iterative denoising process. By comparing paired T2V and I2V generations, we observe that query tokens in non-reference frames in I2V models allocate substantially more attention to the reference-frame key tokens than their paired T2V counterparts. We refer to this asymmetric attention pattern as *reference-frame dominance*.

This observation implies that the static nature of I2V models may stem from excessive propagation of reference-frame information through internal attention routing, rather than that from the input image itself. To test this hypothesis, we directly modulate the attention logits associated with reference-frame key tokens. Strengthening these logits makes generation more static, whereas weakening them restores motion and reduces the reference-frame dominance. This provides intervention-based evidence that the reference-frame dominance is a primary driver of motion suppression in I2V models.

Based on this insight, we present DyMoS (Dynamic Motion Slider), a simple, training-free method that mitigates the reference-frame dominance by rebalancing attention logits. DyMoS introduces a scalar bias shift to the attention logits corresponding to reference-frame key tokens during initial denoising steps, leaving the input image and model weights unchanged. This allows DyMoS to improve motion dynamics while preserving fidelity to the reference image. The magnitude of this bias serves as a single parameter for continuously controlling the generation, ranging from strict reference-image preservation to highly dynamic motion.

We validate the effectiveness of DyMoS in improving motion dynamics across state-of-the-art I2V models, including Wan 2.2 [10], Wan 2.1 [10], HunyuanVideo-1.5 [9] and CogVideoX-5B [11]. Across all backbones, DyMoS consistently improves motion dynamics while preserving reference fidelity, incurring negligible additional inference-time overhead (e.g., 1.2% on Wan 2.2). We further show that DyMoS provides continuous control over the static-to-dynamic spectrum, enabling users to adjust motion strength with a single scalar parameter. These results demonstrate that static motion bias in I2V generation can be effectively mitigated by intervening at the attention-logit level, without fine-tuning the model weights or degrading the input image.

2 Background

Text-to-video models. Modern flow-based text-to-video (T2V) models [10, 9] employ diffusion transformers (DiTs) [27, 28], operating in the latent space of a 3D variational autoencoder (VAE) [29], which consists of an encoder $\mathcal{E}(\cdot)$ and a decoder $\mathcal{D}(\cdot)$. An RGB video $\boldsymbol{x}$ is first compressed into a low-dimensional latent representation via $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x})$. The latent $\boldsymbol{z}_0$ is then perturbed along a flow-matching interpolation path $\boldsymbol{z}_t = (1-t)\boldsymbol{z}_0 + t\boldsymbol{\epsilon}$, where $t \in [0, 1]$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. A neural network $\boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})$ is trained to approximate a target velocity $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{z}_0$, where $\boldsymbol{c}$ denotes the text

[1] In this paper, a reference image serves as the initial frame of the generated video, rather than functioning solely as an appearance or identity reference.

condition embedded by a pretrained text encoder such as T5 [30]. Generation starts from a Gaussian latent $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and progressively denoises it into a clean latent $\boldsymbol{z}_0$ using the learned velocity field. The final latent is then decoded into pixel space as $\hat{\boldsymbol{x}} = \mathcal{D}(\boldsymbol{z}_0)$.

During the denoising process, classifier-free guidance (CFG) [31] is commonly applied in order to improve text-conditional generation by extrapolating between the unconditional and conditional predictions:

$$\tilde{\boldsymbol{v}}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}) = \boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \varnothing) + \omega \left(\boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}) - \boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \varnothing) \right), \tag{1}$$

where $\omega \geq 1$ is the guidance scale, and $\varnothing$ denotes the null text condition.

Self-attention mechanism. Within each DiT block, spatiotemporal latent tokens are processed by self-attention. Let $S = f \cdot h \cdot w$ be the number of latent tokens. For each attention head, the attention weights are computed as

$$\mathcal{A} = \mathrm{softmax}(\mathcal{L}), \qquad \mathcal{L} = \frac{QK^\top}{\sqrt{d_h}} \in \mathbb{R}^{S \times S}, \tag{2}$$

where $Q, K \in \mathbb{R}^{S \times d_h}$ are query and key projections, d_h is the per-head dimension, and $\mathcal{L}$ denotes the attention logits. Since the token sequence spans both spatial and temporal dimensions, this operation aggregates information within and across frames.

Image-to-video models. Image-to-video (I2V) models are commonly obtained by fine-tuning T2V models to additionally condition on a reference image $\boldsymbol{x}_{\mathrm{ref}}$, which corresponds to the first frame of the generated video. The reference image is first encoded as $\mathcal{E}(\boldsymbol{x}_{\mathrm{ref}})$ and temporally aligned with the noisy video latent. For simplicity, we denote the resulting reference-image latent by $\boldsymbol{z}_{\mathrm{ref}}$. The aligned reference-image latent is then concatenated with the noisy video latent along the channel dimension. This gives the denoising network direct access to reference-image features throughout sampling. During inference with an I2V model, CFG in Eq. 1 is extended by providing the reference-image latent to both predictions as

$$\tilde{\boldsymbol{v}}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{z}_{\mathrm{ref}}) = \boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \varnothing, \boldsymbol{z}_{\mathrm{ref}}) + \omega \left(\boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{z}_{\mathrm{ref}}) - \boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \varnothing, \boldsymbol{z}_{\mathrm{ref}}) \right). \tag{3}$$

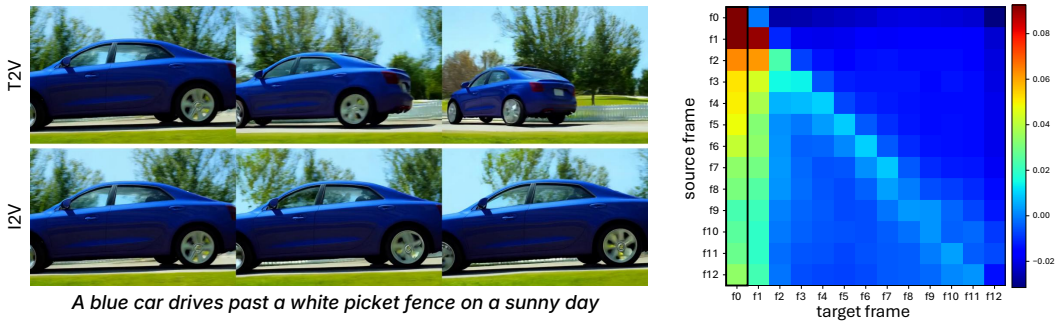
Some I2V models further preserve the reference frame by replacing the corresponding latent with the encoded reference latent after each sampling update.

Static motion bias in I2V generation. Although I2V models can faithfully preserve the appearance of the reference image, they often suffer from a stronger static motion bias than their T2V counterparts, where excessive reliance on the input image leads to videos with weak or limited motion. Several recent methods [24–26, 32] mitigate this bias by weakening or redesigning the image-conditioning pathway. Zhao et al. [24] attribute this phenomenon to conditional image leakage and mitigate it by perturbing the conditioning image with time-dependent noise, thereby reducing the model's reliance on the reference signal. FlashI2V [25] also targets the conditional image leakage by reformulating flow matching with Fourier-guided latent shifting, which injects the image condition implicitly rather than providing the reference latent through direct concatenation. Most closely related to our work, ALG [26] addresses over-conditioning at inference time by suppressing high-frequency components of the input image during early denoising steps.

However, these methods either require additional training or introduce a trade-off between motion enhancement and reference-image preservation. In contrast, our method leaves the image condition unchanged and directly intervenes in the self-attention mechanism during inference, enabling training-free and model-agnostic control over motion dynamics while preserving fidelity to the input image.

3 DyMoS: Dynamic Motion Slider

In this section, we first analyze frame-to-frame self-attention in I2V generation (§3.1) and examine whether modulating this attention pattern controls motion dynamics (§3.2). Finally, we present DyMoS, a training-free and model-agnostic self-attention modulation method (§3.3).



(a) (b)

Figure 2: **Reference-frame dominance in I2V self-attention.** (a) Qualitative comparison between paired T2V and I2V generations. (b) Frame-to-frame self-attention difference map $\mathcal{A}^{\mathrm{I2V}} - \mathcal{A}^{\mathrm{T2V}}$, averaged over the first 10% of inference steps.

3.1 Reference-frame dominance in I2V self-attention

To understand how image conditioning changes information routing within the model, we compare frame-to-frame self-attention distributions between an I2V model and its corresponding T2V counterpart. We first generate 50 videos using a T2V model, and then use the first frames as reference images for an I2V model, using the same text prompts sourced from T2V-CompBench [33]. Since this setup ensures the same prompt and first frame for both models, we can analyze the differences in their respective attention distributions. The detailed procedure is provided in Appendix B.

In this paired setting, the two models generate videos with distinct motion dynamics and self-attention patterns. As shown in Fig. 2(a), the I2V model tends to preserve the reference image with limited inter-frame movement, whereas the T2V counterpart exhibits more dynamic motion given the same prompt. Fig. 2(b) shows the frame-to-frame attention difference map $\mathcal{A}^{\mathrm{I2V}} - \mathcal{A}^{\mathrm{T2V}}$. Positive values indicate that I2V allocates significantly more attention to the reference key tokens than its T2V counterpart as highlighted in black. We refer to this structural asymmetry as *reference-frame dominance*. Notably, this dominance is most pronounced during the initial denoising stage, which is consistent with the coarse-to-fine nature of the denoising process. Since early steps largely determine global motion of the generated video [26, 34, 35], we target our subsequent intervention at this stage.

3.2 Correlation between reference-frame dominance and motion dynamics

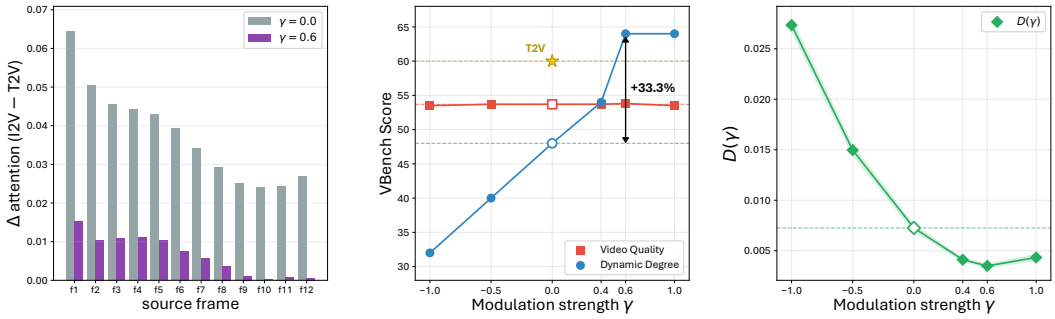
Based on this observation, we hypothesize that reference-frame dominance is a primary factor underlying static I2V generation, and we demonstrate that directly modulating this dominance impacts the motion dynamics of the generated videos. Since the reference frame contains the input image information, excessive attention to reference-frame key tokens propagates the reference appearance along the temporal axis, thereby weakening frame-to-frame interactions necessary for motion dynamics. Accordingly, we argue that attenuating reference-frame dominance should affect both the generated motion and the frame-to-frame attention structure.

To test this, we directly intervene on the attention logits during the first 10% of denoising steps. At every self-attention layer, we add a scalar bias to the attention logits from non-reference-frame query tokens to reference-frame key tokens before the softmax operation:

$$\tilde{\mathcal{L}}[i,j] = \mathcal{L}[i,j] - \gamma \cdot \mathbb{1}[j \in \mathcal{I}_{f_0}] \cdot \mathbb{1}[i \notin \mathcal{I}_{f_0}] \tag{4}$$

where $\mathcal{I}_{f_0}$ denotes the set of token indices belonging to the reference frame, and γ represents the strength of the intervention. A negative value of γ increases the logits assigned to reference-frame keys and thereby strengthens reference-frame dominance, whereas a positive γ suppresses these logits and weakens the dominance. Fig. 3(a) shows the resulting change in reference-frame attention for $\gamma = 0.6$, confirming that the intervention effectively attenuates the intended attention pathway.

We next examine whether this intervention alters generation behavior with VBench [36] metrics: Dynamic Degree and Video Quality. As shown in Fig. 3(b), increasing γ consistently increases Dynamic Degree, moving the I2V model from near-static outputs toward more dynamic generation.



(a) (b) (c)

Figure 3: **Modulating reference-frame dominance controls motion dynamics.** **(a)** Absolute difference in reference-frame attention between the vanilla I2V model and the modulated I2V model with $\gamma = 0.6$, measured over non-reference query frames. **(b)** Dynamic Degree and Video Quality as γ varies. The yellow star denotes the Dynamic Degree of the paired T2V generation. **(c)** T2V–I2V attention distance $D(\gamma)$ measured by Jensen–Shannon divergence.

Notably, Video Quality remains stable over a broad range of γ values. This shows that weakening reference-frame dominance can improve motion dynamics without degrading visual quality. Conversely, negative γ strengthens reference-frame dominance and results in more static videos.

Finally, we quantify how the modulation affects the frame-to-frame attention structure relative to the paired T2V generation. Let $\mathcal{A}_\gamma^{\mathrm{I2V}}$ denote the frame-to-frame attention matrix of the I2V model under a modulation strength of γ. We define the T2V–I2V attention distance as

$$D(\gamma) = \frac{1}{F-1} \sum_{a=1}^{F-1} \mathrm{JSD}\big(\mathcal{A}_\gamma^{\mathrm{I2V}}[a,:] \,\|\, \mathcal{A}^{\mathrm{T2V}}[a,:]\big)\,, \tag{5}$$

where JSD denotes the Jensen–Shannon divergence [37], and the average is taken over non-reference query frames. A smaller $D(\gamma)$ indicates that the I2V attention pattern is closer to that of the paired T2V model. Fig. 3(c) shows an inverse relationship between the modulation strength and the T2V–I2V attention distance. In particular, $D(\gamma)$ reaches its minimum value around $\gamma = 0.6$, where the Dynamic Degree is also closest to that of the paired T2V generation. This consistency confirms that attenuating reference-frame dominance moves the I2V model toward a more T2V-like regime in both attention structure and motion dynamics. To further validate the specificity of this intervention, we provide complementary experimental results in Appendix B.2.

3.3 Proposed method: DyMoS

Our analysis provides the key insight that static generation in I2V models can be mitigated by attenuating reference-frame dominance in self-attention. Moreover, varying the attenuation strength enables continuous control over motion dynamics. To this end, we introduce DyMoS (Dynamic Motion Slider), a simple yet effective inference-time method that subtracts a positive scalar γ from the logits associated with reference-frame key tokens by generalizing Eq. 4:

$$\tilde{\mathcal{L}}[i,j] = \mathcal{L}[i,j] - \gamma \cdot \phi(f(i)) \cdot \mathbb{1}[j \in \mathcal{I}_{f_0}] \cdot \mathbb{1}[i \notin \mathcal{I}_{f_0}], \tag{6}$$

where $f(i)$ denotes the latent-frame index of query token i, and $\phi(\cdot)$ is a frame-wise modulation schedule that controls how the modulation strength varies across query frames.

This design explicitly targets the pathway identified in our analysis: attention from generated frames to reference frame. Since the modulation is applied in logit space prior to the softmax operation, DyMoS does not directly discard reference information. Instead, it reduces the excessive propagation of reference-frame information to non-reference frames while preserving fidelity to the input image.

DyMoS incorporates two key design choices. First, we apply the modulation only during the first $\lambda \in [0,1]$ fraction of sampling steps and revert to the original attention computation thereafter. Here, λ determines the transition point between modulated and standard attention. This approach is motivated by the observation that reference-frame dominance is most pronounced during early denoising steps,

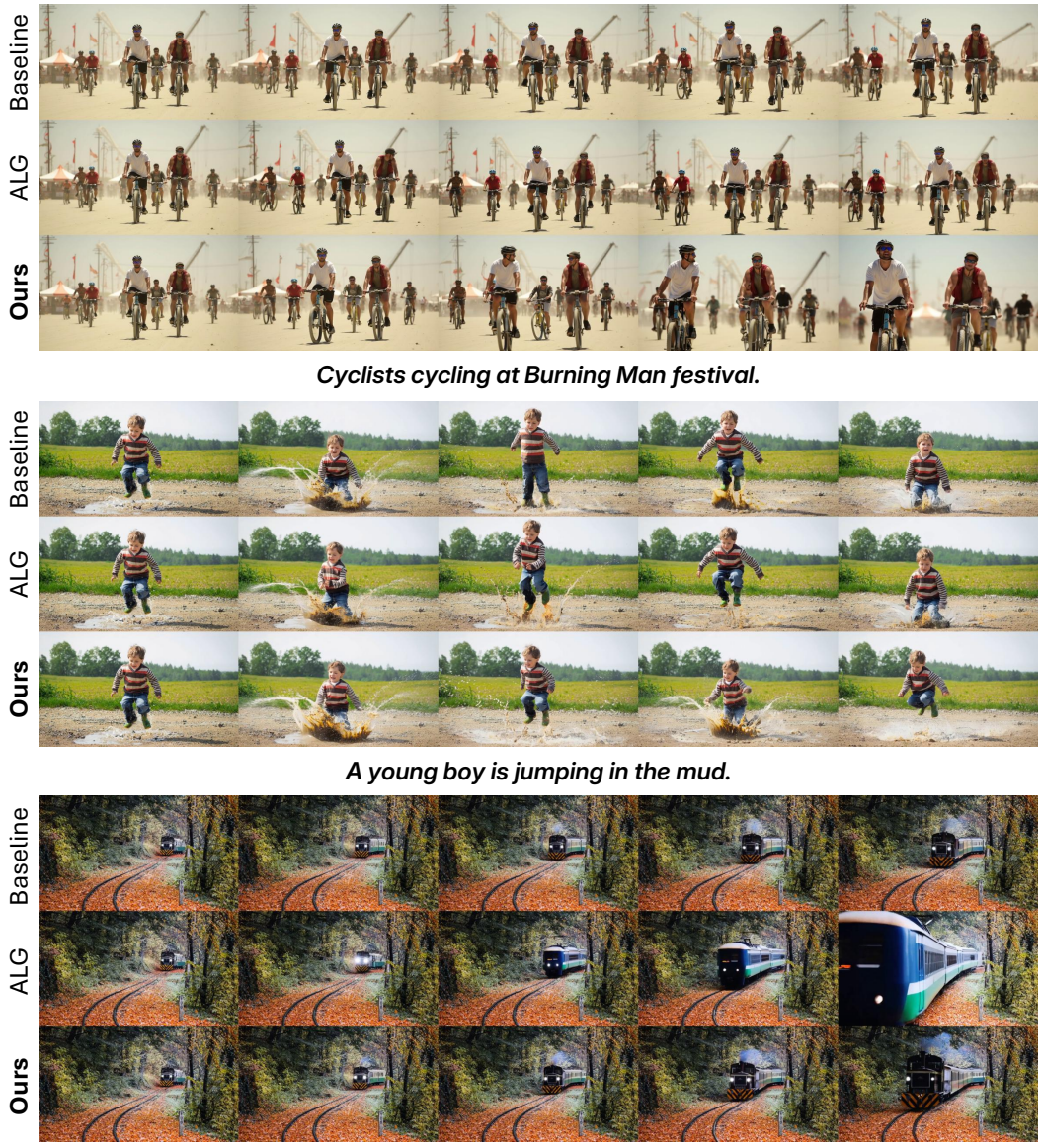


Figure 4: **Qualitative comparison with vanilla I2V baseline and ALG.** The leftmost images are the reference images. Our method (DyMoS) produces substantially more dynamic motion than the vanilla baseline and ALG while preserving fidelity to the reference image. In contrast, ALG introduces motion at the cost of visible degradation. We highly encourage readers to visit our project page for a better visualization of our results.

as shown in Fig. 2(b). Applying the modulation specifically in this early stage enhances motion, while preserving appearance consistency and temporal coherence in later refinement steps.

Second, under CFG in Eq. 3, we apply DyMoS only to the text-conditional branch as

$$\tilde{\boldsymbol{v}}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{z}_{\text{ref}}) = \boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \varnothing, \boldsymbol{z}_{\text{ref}}) + \omega(\boldsymbol{v}_\theta^{\tilde{\mathcal{L}}}(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{z}_{\text{ref}}) - \boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \varnothing, \boldsymbol{z}_{\text{ref}})), \tag{7}$$

where $\boldsymbol{v}_\theta^{\tilde{\mathcal{L}}}(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{z}_{\text{ref}})$ denotes the text-conditional prediction computed with modulated attention logits $\tilde{\mathcal{L}}$, and $\boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \varnothing, \boldsymbol{z}_{\text{ref}})$ is the standard prediction under the null text condition. This design steers the guided update away from reference-frame dominance while keeping the null-text branch unchanged. The full sampling algorithm for DyMoS is provided in Appendix A.

Table 1: Quantitative results across various I2V backbones on VBench-I2V test set.

Model	Method	Dynamic Degree		Video Quality	ViCLIP	Vision Reward
		VBench	VideoScore			
Wan 2.2-14B	Baseline	51.7	2.83	**58.0**	0.2614	0.143
	ALG	55.3	2.83	57.7	0.2616	0.140
	Ours	**64.8**	**2.84**	57.7	**0.2621**	**0.143**
Wan 2.1-14B	Baseline	48.7	3.21	**57.7**	0.2601	**0.132**
	ALG	52.5	3.18	57.2	0.2615	0.127
	Ours	**54.5**	**3.25**	57.4	**0.2618**	0.126
HunyuanVideo-1.5	Baseline	48.0	2.87	**58.2**	0.2624	0.140
	ALG	46.3	2.87	58.2	0.2629	0.141
	Ours	**55.3**	**2.88**	58.1	**0.2653**	**0.141**
CogVideoX-5B	Baseline	30.9	2.90	**57.8**	0.2612	**0.140**
	ALG	52.0	**2.98**	57.1	0.2647	0.127
	Ours	**52.9**	2.92	57.1	**0.2652**	0.131

4 Experiments

4.1 Experimental setup

I2V backbones and comparison methods. To verify the robustness of our method, we conduct experiments on several state-of-the-art I2V models, including Wan 2.2 [10], Wan 2.1 [10], HunyuanVideo-1.5 [9] and CogVideoX-5B [11]. Wan 2.2 and Wan 2.1 adopt a separate self-attention and cross-attention design, while HunyuanVideo-1.5 and CogVideoX-5B adopt MM-DiT architectures. Evaluating on both designs allows us to verify that our findings and method generalize across different architectures. We compare DyMoS against the vanilla I2V backbones (baseline) and ALG [26] using the official default configurations. For Wan 2.2, we set $\gamma = 0.6$ and $\lambda = 0.2$. We select γ and λ separately for each backbone to account for differences in attention scale and denoising schedule, and keep them fixed across all prompts. Per-backbone configurations are provided in Appendix C.1.

Evaluation metrics. We evaluate the generated videos in various aspects: motion dynamics, video quality, text-video alignment, and human-aligned preference. Since DyMoS mainly targets enhancing motion dynamics in I2V generation, we adopt *Dynamic Degree* from both VBench [36] and VideoScore [38] as our primary metric. For video quality, we report *Video Quality*, an aggregate score computed over selected VBench quality-related dimensions. For text-video alignment, we use ViCLIP [39] to compute the similarity between the input prompt and the generated video. Finally, we include VisionReward [40], a VLM-based human-aligned reward model for image and video generation. Further details on evaluation metrics are provided in Appendix C.2.

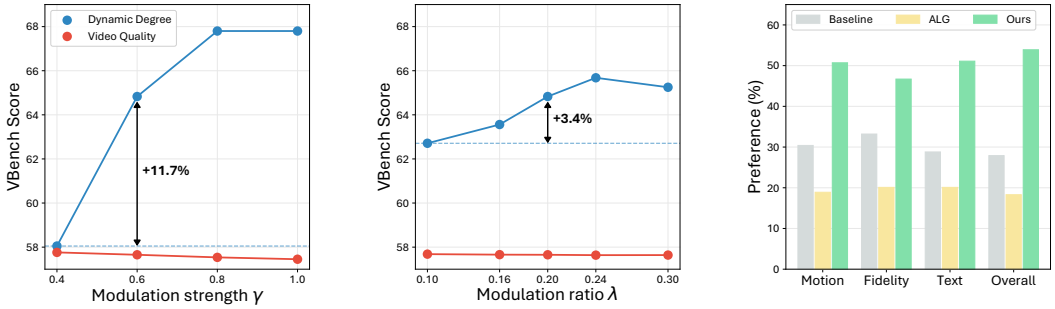
Datasets. We assess our method on three datasets. For our primary evaluation, we use 246 image-caption pairs from the VBench-I2V test set [36], which consists of diverse scenes and motion types. We additionally evaluate on 200 image-caption pairs sampled from PE Video Dataset (PVD) [41], a large-scale video-caption dataset containing both synthetic and human-annotated captions, from which we use the first frame of each video as the conditioning image. Finally, we include VidProM [42], a large-scale text-to-video prompt dataset, to assess generalization to diverse text prompts. We randomly sample 500 prompts from VidProM and generate corresponding reference images using Flux.1-dev [43]. Further details on each dataset are provided in Appendix C.3.

4.2 Main results

Qualitative evaluation. As illustrated in Fig. 4, DyMoS achieves a better balance between motion dynamics and reference-image fidelity than both the vanilla I2V baseline and ALG. The vanilla I2V baseline often produces nearly static videos, where the generated frames remain overly close to the reference image even when the prompt describes object motion. ALG improves motion dynamics over the baseline, but it can introduce visible degradation due to its input-level attenuation. For example, in the train sequence, ALG induces motion but also changes the appearance of the train as

Table 2: Quantitative results with Wan 2.2-14B across various benchmark datasets.

Dataset	Method	Dynamic Degree		Video Quality	ViCLIP	Vision Reward
		VBench	VideoScore			
VBench-I2V	Baseline	51.7	2.83	**58.0**	0.2614	0.143
	ALG	55.3	2.83	57.7	0.2616	0.140
	Ours	**64.8**	**2.84**	57.7	**0.2621**	**0.143**
PVD	Baseline	79.0	2.80	55.0	**0.2147**	**0.087**
	ALG	77.5	2.80	55.1	0.2141	0.084
	Ours	**85.0**	**2.80**	**55.1**	0.2143	0.086
VidProM	Baseline	54.2	2.78	57.3	0.2427	**0.064**
	ALG	59.0	2.78	58.2	**0.2433**	0.062
	Ours	**67.0**	**2.80**	**60.1**	0.2432	0.062



(a) (b) (c)

Figure 5: **Hyperparameter analysis and user study.** (a) Effect of modulation strength γ. (b) Effect of modulation step ratio λ. (c) User study results. The final preference score for each method is computed as the fraction of times it is selected under each criterion.

the video progresses, indicating a loss of reference fidelity. In contrast, DyMoS produces substantially more dynamic videos while faithfully preserving the reference image, as shown in the first two examples. In particular, in the boy example, DyMoS sustains the jumping motion over more frames, resulting in an additional jump cycle compared to the other methods. Additional qualitative examples are provided in Appendix D.5.

Quantitative evaluation. Tab. 1 reports quantitative results across four different I2V backbones on VBench-I2V. Our method consistently achieves higher Dynamic Degree than both the baseline and ALG, demonstrating its effectiveness in improving motion dynamics. Crucially, DyMoS maintains Video Quality scores comparable to ALG while achieving substantially higher Dynamic Degree. This indicates that our attention-level intervention improves motion dynamics more effectively than input-level attenuation, while preserving visual quality. Our method also achieves the highest ViCLIP scores among the compared methods, suggesting that reducing reference-frame dominance allows the model to better follow motion described in text prompts. In addition, VisionReward scores remain stable or improve, indicating that enhanced motion dynamics do not degrade overall perceptual quality and human preference. Tab. 2 further validates the effectiveness of our method across various benchmark datasets, supporting its robust performance across diverse reference images and text prompts. We further provide a comparison of computation costs in Appendix D.2.

User study. We conduct a user study with 30 participants to complement the automated evaluation. Each participant evaluates 25 video triplets, each comprising videos generated from the same reference image and text prompt using the vanilla baseline, ALG, and DyMoS, with method names hidden and presentation order randomized. Participants select the best video according to four criteria: *Motion Dynamicness*, *Reference-Frame Fidelity*, *Text Alignment*, and *Overall Preference*. As shown in Fig.5(c), DyMoS is preferred most frequently across all four criteria, demonstrating its perceptual advantages. Further details are provided in the Appendix C.4.



Figure 6: **Continuous control over static-to-dynamic generation with DyMoS.** Rows from top to bottom correspond to $\gamma \in \{-2, -1, 0, 0.6, 1.0\}$, with $\gamma = 0$ denoting the baseline.

4.3 Ablation studies

We conduct ablation studies on two key hyperparameters of DyMoS, the modulation strength γ and the modulation step ratio λ, to investigate how the strength and duration of attention modulation affect motion dynamics and video quality.

Effect of modulation strength γ. To assess the impact of the modulation strength γ, we vary $\gamma \in \{0.4, 0.6, 0.8, 1.0\}$ and measure Dynamic Degree and Video Quality for each setting. As shown in Fig. 5(a), Dynamic Degree consistently increases as γ increases, which is consistent with our analysis in §3.2. For example, increasing γ from 0.4 to 0.6 yields an 11.7% relative gain in Dynamic Degree. In contrast, Video Quality remains nearly stable across the same range. These results show that γ can control the strength of motion dynamics without substantially degrading video quality.

Effect of modulation step ratio λ. Additionally, we investigate how the modulation step ratio λ affects motion dynamics and video quality. As shown in Fig. 5(b), increasing λ initially improves Dynamic Degree while Video Quality remains stable. The improvement saturates beyond a certain point, as increasing λ from 0.24 to 0.30 reduces Dynamic Degree. This demonstrates that applying DyMoS to an appropriate number of initial denoising steps is sufficient to enhance motion, while switching back to the original attention computation helps maintain generation quality.

4.4 Application: Continuous control over motion dynamics

A key practical advantage of DyMoS is that it provides a simple control knob for continuously adjusting the balance between reference preservation and motion strength. Under our convention, $\gamma = 0$ recovers the vanilla I2V baseline, $\gamma > 0$ weakens reference-frame dominance to encourage stronger motion, and $\gamma < 0$ strengthens reference-frame dominance to further suppress motion.

Fig. 6 illustrates this behavior across different values of γ. At $\gamma = 0$, the model strongly preserves the reference image and produces limited motion. As γ becomes negative, the generated videos become even more static, with the reference image effectively propagated along the temporal axis. For example, at $\gamma = -2$, both the car and the horses remain nearly static. As γ increases above zero, the generated videos exhibit richer motion, such as the car moving along the road with a following camera. Moderate positive values of γ improve motion while preserving reference fidelity, whereas overly large values suppress reference-frame dominance beyond the level observed in T2V models, yielding much more dynamic motion at the cost of appearance consistency. These results demonstrate that DyMoS enables continuous control over the static-to-dynamic generation spectrum.

5 Conclusion

We present DyMoS (Dynamic Motion Slider), a training-free and model-agnostic method for improving motion dynamics in image-to-video generation. Our analysis identifies that static motion bias in I2V models is primarily caused by reference-frame dominance. DyMoS mitigates this issue by subtracting a log-space bias from reference-frame key logits during initial denoising steps, thereby reducing excessive reference propagation without modifying the input image or model weights. With a single parameter γ, DyMoS offers continuous control over the balance between reference preservation and motion strength. Experiments across recent I2V backbones show that DyMoS significantly improves motion dynamics while maintaining reference fidelity and visual quality.

Limitation and future work. Although DyMoS improves motion dynamics across diverse I2V backbones, overly large modulation strengths can reduce appearance consistency and temporal stability by excessively suppressing reference-frame information. In addition, since our method currently uses fixed hyperparameters per backbone, an automatic adapation of the modulation strength to different prompts or motion types remains for future work.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean Government (MSIT) (RS-2020-II201361 and RS-2025-25441838), and by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2024-00338439).

References

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

[2] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

[3] 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.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[5] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[6] 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`.

[7] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=li7qeBbCR1t`.

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[9] Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. URL `https://arxiv.org/abs/2511.18870`.

[10] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[11] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=LQzN6TRFg9`.

[12] 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.

[13] 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.

[14] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Fx2SbBgcte`.

[17] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[18] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18444–18455. IEEE, 2023.

[19] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[20] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

[21] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024.

[22] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[23] Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7276–7287, 2025.

[24] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=o9Lkiv1qpc.

[25] Yunyang Ge, Xinhua Cheng, Chengshu Zhao, Xianyi He, Shenghai Yuan, Bin Lin, Bin Zhu, and Li Yuan. Flashi2v: Fourier-guided latent shifting prevents conditional image leakage in image-to-video generation. *arXiv preprint arXiv:2509.25187*, 2025.

[26] June Suk Choi, Kyungmin Lee, Sihyun Yu, Yisol Choi, Jinwoo Shin, and Kimin Lee. Enhancing motion dynamics of image-to-video models via adaptive low-pass guidance. *arXiv preprint arXiv:2506.08456*, 2025.

[27] 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.

[28] 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. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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

[32] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=j8Vr3E3vhy.

[33] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416, 2025.

[34] Byungjun Kim, Soobin Um, and Jong Chul Ye. Motioncfg: Boosting motion dynamics via stochastic concept perturbation. *arXiv preprint arXiv:2603.14073*, 2026.

[35] Wooseok Jeon, Seunghyun Shin, Dongmin Shin, and Hae-Gon Jeon. Motion prior distillation in time reversal sampling for generative inbetweening. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GRElsj9W2t.

[36] 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.

[37] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 2002.

[38] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, 2024.

[39] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MLBdiWu4Fw.

[40] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11269–11277, 2026.

[41] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=INqBOmwIpG`.

[42] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

[43] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[45] Dohun Lee, Chun-Hao Paul Huang, Xuelin Chen, Jong Chul Ye, Duygu Ceylan, and Hyeonho Jeong. Memory-v2v: Augmenting video-to-video diffusion models with memory. *arXiv preprint arXiv:2601.16296*, 2026.

[46] Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13-14, 2012. Proceedings*, 2012.

[47] Seunghyun Shin, Dongmin Shin, Jisu Shin, Hae-Gon Jeon, and Joon-Young Lee. Video color grading via look-up table generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19141–19152, 2025.

[48] Seunghyun Shin, Jisu Shin, Jihwan Bae, Inwook Shim, and Hae-Gon Jeon. Close imitation of expert retouching for black-and-white photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25037–25046, June 2024.

[49] Seonmi Park, Inhwan Bae, Seunghyun Shin, and Hae-Gon Jeon. Kinetic typography diffusion model. In *European Conference on Computer Vision*, pages 166–185. Springer, 2024.

[50] Chanhui Lee, Seunghyun Shin, Donggyu Choi, Hae-gon Jeon, and Jeany Son. Universal image immunization against diffusion-based image editing via semantic injection. *arXiv preprint arXiv:2602.14679*, 2026.

[51] Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-a-video: 4d video generation as video-to-video translation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11164–11175. IEEE, 2025.

A Full algorithm of DyMoS

Algorithm 1 DyMoS

Input: Reference image I_{ref}, text prompt $\boldsymbol{c}$, total inference steps N, guidance scale ω, modulation strength γ, modulation step ratio λ, frame-wise schedule $\phi(\cdot)$

$\boldsymbol{z}_{\text{ref}} \leftarrow \mathcal{E}(I_{\text{ref}})$
$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
$\mathcal{I}_{f_0} \leftarrow$ token indices of the reference frame
for $i = 1$ **to** N **do**
  $t \leftarrow \frac{i}{N}$
  if $t < \lambda$ **then**
    $\tilde{\mathcal{L}}[p, q] \leftarrow \mathcal{L}[p, q] - \gamma \cdot \phi(f(p)) \cdot \mathbb{1}[q \in \mathcal{I}_{f_0}] \cdot \mathbb{1}[p \notin \mathcal{I}_{f_0}]$ ▷ Eq. 6
  else
    $\tilde{\mathcal{L}} \leftarrow \mathcal{L}$
  end if
  $\tilde{\boldsymbol{v}}_\theta(\boldsymbol{z}, t, \boldsymbol{c}, \boldsymbol{z}_{\text{ref}}) \leftarrow \boldsymbol{v}_\theta(\boldsymbol{z}, t, \varnothing, \boldsymbol{z}_{\text{ref}}) + \omega\Big(\boldsymbol{v}_\theta^{\tilde{\mathcal{L}}}(\boldsymbol{z}, t, \boldsymbol{c}, \boldsymbol{z}_{\text{ref}}) - \boldsymbol{v}_\theta(\boldsymbol{z}, t, \varnothing, \boldsymbol{z}_{\text{ref}})\Big)$ ▷ Eq. 7
  $\boldsymbol{z} \leftarrow \text{SchedulerStep}(\boldsymbol{z}, \tilde{\boldsymbol{v}}_\theta, t)$
end for
return $\hat{\boldsymbol{x}} \leftarrow \mathcal{D}(\boldsymbol{z})$

B Additional details on analysis

B.1 Experimental setting

We provide the experimental procedure used to obtain the frame-to-frame self-attention comparison reported in §3.1 and Fig. 2.

Paired T2V/I2V generation. We construct paired T2V and I2V generations using CogVideoX-5B. We first generate a video $\boldsymbol{v}^{\text{T2V}}$ with CogVideoX-5B-T2V using a text prompt $\boldsymbol{c}$ sampled from T2V-CompBench and a fixed random seed. We then extract the first frame $\boldsymbol{x}_{\text{ref}} = \boldsymbol{v}^{\text{T2V}}[0]$ and feed it together with the same prompt $\boldsymbol{c}$ and the same seed to CogVideoX-5B-I2V to obtain $\boldsymbol{v}^{\text{I2V}}$. We keep the sampling hyperparameters matched between the two runs, including the number of denoising steps, classifier-free guidance scale, resolution, and number of latent frames. We repeat the procedure above for $N = 50$ prompts sampled from T2V-CompBench, covering a range of motion types, including camera motion, object motion, and human action, as well as diverse scene categories. All reported attention statistics are averaged over this prompt set.

Attention extraction. At each denoising step $s \in \{0, \ldots, N-1\}$, self-attention layer ℓ, and attention head h, we record the post-softmax attention matrix $\mathcal{A}^{(s,\ell,h)} \in \mathbb{R}^{L \times L}$, where L is the total number of latent tokens. We convert this token-level attention into a frame-to-frame attention matrix by summing attention over key tokens within each frame and averaging over query tokens within each query frame:

$$\bar{\mathcal{A}}^{(s,\ell,h)}[a, b] = \frac{1}{|\mathcal{I}_a|} \sum_{i \in \mathcal{I}_a} \sum_{j \in \mathcal{I}_b} \mathcal{A}^{(s,\ell,h)}[i, j], \tag{8}$$

where $\mathcal{I}_a$ and $\mathcal{I}_b$ denote the token indices belonging to frames a and b, respectively. This quantity represents the average fraction of attention mass that query tokens in frame a assign to key tokens in frame b. For Fig. 2, we compute $\mathcal{A}^{\text{T2V}}$ and $\mathcal{A}^{\text{I2V}}$ by averaging $\bar{\mathcal{A}}^{(s,\ell,h)}$ over the initial 10% of denoising steps, attention heads, layers, and prompts.

B.2 Further validation of reference-frame dominance

We provide further validation of reference-frame dominance in §3.2.

Ablation study on suppression target. We examine whether motion enhancement is specific to suppressing reference-frame attention or can also arise from suppressing other attention pathways. Using CogVideoX-5B and 50 prompts from T2V-CompBench, we compare three suppression targets: reference-frame keys, keys from a randomly selected non-reference frame, and a random token subset containing the same number of keys as the reference frame. We use the same modulation strength and denoising interval across the three interventions. As shown in Tab. 3, suppressing reference-frame keys increases Dynamic Degree from 48.0 to 64.0, while the two control interventions yield scores of 50.0 and 48.0. Video Quality remains comparable to the baseline under reference-frame suppression. Together with the bidirectional modulation results in the main paper, these controls support a specific contribution of the reference-frame pathway to motion suppression. They do not establish that reference-frame dominance is the sole source of static generation.

Table 3: **Experiments on varying suppression target.**

Suppression target	DD ↑	VQ ↑
None (baseline)	48.0	53.7
Random non-reference frame	50.0	53.2
Random token subset	48.0	53.3
Reference frame (DyMoS)	64.0	53.8

Generalization across backbones. We extend the paired T2V/I2V analysis to Wan 2.2, Wan 2.1, and HunyuanVideo-1.5 using 50 prompts from T2V-CompBench. Following the procedure in Appendix B.1, we measure the attention mass assigned by each non-reference query frame to reference-frame keys and average over the first 10% of denoising steps. Tab. 4 reports the relative increase in this mass for I2V compared with the corresponding T2V model at selected latent frames. All three backbones exhibit increased reference attention across multiple non-reference frames. The pattern varies across architectures and temporal positions. In particular, the final reported frame has a negative difference for Wan 2.2 and Wan 2.1. These results extend the evidence beyond CogVideoX-5B without implying a uniform increase at every frame.

Table 4: **Reference attention across additional backbones.** Relative change in reference-frame attention mass from T2V to I2V, in percent. The reference frame is f_0.

Backbone	f_1	f_2	f_3	f_4	f_5	$\cdots$	f_{18}	f_{19}	f_{20}
Wan 2.2	$+35.9$	$+53.3$	$+31.2$	$+45.9$	$+35.9$	$\cdots$	$+25.3$	$+11.1$	-13.1
Wan 2.1	$+41.5$	$+45.6$	$+28.2$	$+27.7$	$+16.6$	$\cdots$	$+14.6$	$+0.7$	-4.7
HunyuanVideo-1.5	$+55.9$	$+61.0$	$+79.0$	$+74.7$	$+81.3$	$\cdots$	$+151.5$	$+146.9$	$+114.8$

Evolution over denoising steps. We evaluate the T2V–I2V attention distance in Eq. 5 at five denoising steps using CogVideoX-5B. For this analysis, attention statistics are aggregated separately at each reported step. As shown in Tab. 5, the distance is largest at step 5 among the measured steps and decreases to 0.0012 at the final step. This indicates that the frame-to-frame attention distributions differ more strongly during early denoising than during later refinement.

Table 5: **T2V–I2V Attention distance over denoising steps.** Step 50 is the final sampling step.

Denoising step	1	5	10	25	50
Attention distance	0.0079	0.0109	0.0082	0.0030	0.0012

C Additional details on the experimental setup

C.1 Implementation and setup

Implementation details. We implement DyMoS on top of the official open-source releases of Wan 2.2, Wan 2.1, HunyuanVideo-1.5 and CogVideoX-5B. The model checkpoints we used in our experiments are provided in Tab. 6. DyMoS is applied identically across all backbones. During the

Table 6: Detailed configurations for each I2V backbone.

	Models	Wan 2.2-14B	Wan 2.1-14B	HunyuanVideo-1.5	CogVideoX-5B
Base	Resolution	81×832×480	81×832×480	81×832×480	49×720×480
	FPS	16	16	16	8
	Denoising steps	40	40	50	50
	CFG scale	3.5	5.0	6.0	6.0
	Scheduler	UniPC	UniPC	FlowMatch-Euler	DDIM
DyMoS	γ	0.6	1.0	1.0	0.8
	λ	0.2	0.25	0.2	0.24
	$\phi(\cdot)$	1	1	1	1

Table 7: Model checkpoints used in our main experiments.

Model	**Type**	**Source**
Wan 2.2	I2V	`https://huggingface.co/Wan-AI/Wan2.2-I2V-A14B-Diffusers`
Wan 2.1	I2V	`https://huggingface.co/Wan-AI/Wan2.1-I2V-14B-480P-Diffusers`
HunyuanVideo-1.5	I2V	`https://huggingface.co/tencent/HunyuanVideo-1.5`
CogVideoX-5B	I2V	`https://huggingface.co/zai-org/CogVideoX-5b-I2V`

first λN denoising steps, we subtract a bias γ to the pre-softmax self-attention logits for query-key pairs where the key is the frame-0 video token and the query is a non-frame-0 video token. The frame-0→frame-0 attention block is exempted to preserve the reference frame.

Configuration. Tab. 6 summarizes the per-backbone sampling configurations used in all experiments, including resolution, FPS, denoising steps, CFG scale, scheduler, and our hyperparameters $(\gamma, \lambda, \phi_f)$. All experiments are conducted on either a single NVIDIA H100 GPU or a single NVIDIA A100 GPU. For each backbone, all methods are run on the same GPU type, and no multi-GPU inference is used. Random seeds are fixed across methods within each pair to ensure that comparisons better isolate the effect of the method itself. For ALG [26], we use the official implementation [2] and the recommended hyperparameters reported in the paper.

C.2 Evaluation metrics

We provide details of the evaluation metrics used throughout our experiments, including VBench [36], VideoScore [38], ViCLIP [39], and VisionReward [40]. Our evaluation covers four aspects: motion dynamics, video quality, text-video alignment, and human-aligned preference.

Motion dynamics. We evaluate motion dynamics using *Dynamic Degree* from both VBench and VideoScore. VBench Dynamic Degree is designed to measure whether a generated video contains sufficiently large motion by computing the highest 5% optical flows across frames using RAFT [44]. VideoScore Dynamic Degree is predicted by a VLM-based video evaluator trained on human feedback. We report both metrics to verify that DyMoS improves motion dynamics under both a conventional metric and a learned human-aligned evaluator.

Video quality. We report *Video Quality* as an aggregate score over selected VBench quality-related dimensions as done in previous works [11, 23, 25, 26, 45]. Specifically, we compute

$$\text{VideoQuality} = 0.1 \cdot \text{TF} + 0.25 \cdot \text{MS} + 0.1 \cdot \text{AQ} + 0.25 \cdot \text{IQ},$$

where TF, MS, AQ, and IQ denote Temporal Flickering, Motion Smoothness, Aesthetic Quality, and Imaging Quality, respectively. These dimensions capture complementary aspects of perceptual video quality: TF penalizes frame-to-frame flickering artifacts, MS measures the temporal coherence of motion, and AQ/IQ assess the visual appeal and low-level imaging quality of individual frames. The weights follow the official VBench aggregation; we omit Dynamic Degree, Subject Consistency, and Background Consistency to isolate video quality from motion strength and reference preservation.

[2] https://github.com/choi403/ALG

Text-video alignment. We use ViCLIP to evaluate text-video semantic alignment. ViCLIP is a ViT-L-based video-text representation model trained on InternVid with contrastive video-language learning, mapping videos and text prompts into a shared embedding space. We compute the similarity between the input prompt and the generated video representation as a global measure of text-video semantic matching.

Human-aligned preference. We additionally evaluate generated videos with VisionReward, a VLM-based reward model for image and video generation. VisionReward decomposes human visual preferences into multiple fine-grained assessment questions (e.g. prompt alignment, fidelity, motion realism, stability) and combines the corresponding judgments through interpretable linear weighting to produce a human-aligned reward score. It is trained to capture multi-dimensional human preferences over both images and videos, including dynamic aspects that are important for video evaluation. We use VisionReward to complement benchmark-specific metrics and assess whether DyMoS's outputs are also favored by a learned human-aligned evaluator.

C.3 Evaluation datasets

We evaluate DyMoS on three datasets covering complementary scenarios: a standard I2V benchmark, a recently released video-caption dataset, and a large-scale prompt collection paired with synthesized images. Note that we adopt the same dataset suite as ALG to ensure consistent comparison.

VBench-I2V. *VBench-I2V* is the official I2V benchmark of VBench, consisting of image-caption pairs for evaluating image-to-video generation. Since our study focuses on video motion dynamics, we use the 246 image-caption pairs obtained after excluding subsets designed for background-quality evaluation and camera-motion instruction following. The provided reference images are used directly as conditioning inputs. We do not report the I2V-specific subject/background consistency dimensions, as our evaluation focuses on motion and video quality.

PE Video Dataset (PVD). *PE Video Dataset (PVD)* is a large-scale video-caption dataset released alongside the Perception Encoder, containing both synthetic and human-annotated video captions. We randomly sample 200 videos from PVD and use the first frame of each video as the conditioning image, paired with the corresponding caption as the text prompt. PVD allows us to evaluate DyMoS on real-world videos with naturally diverse motion patterns.

VidProM. *VidProM* is a large-scale text-to-video prompt dataset originally collected to study real user prompts in T2V generation. We randomly sample 500 prompts from VidProM and generate the corresponding conditioning images using a text-to-image model Flux.1-dev. This setup probes whether DyMoS generalizes to diverse text prompts beyond curated I2V benchmarks.

C.4 User study protocol

To further evaluate beyond automated metrics, we conduct a user study via MTurk [46] following the previous works [47–51]. For each trial, participants are shown three videos generated from the same reference image and text prompt by the vanilla I2V baseline, ALG, and DyMoS. Method names are hidden, and the video order is randomized.

Participants answer four questions:

1. **Motion:** Which video has the most dynamic and realistic motion? Examples include water ripples, cloth movement, human action, and camera motion.
2. **Fidelity:** Which video best preserves the appearance of the reference image throughout the sequence? Examples include the subject, background, and colors.
3. **Text alignment:** Which video most faithfully reflects the content described in the text prompt?
4. **Overall preference:** Overall, which video do you prefer?

We collect 30 responses for each question over 25 randomly sampled image–prompt pairs from VBench-I2V. This yields 750 participant–triplet evaluations and 3,000 criterion-level judgments. The final preference score for each method is computed as the fraction of times it is selected under each criterion.

Table 8: Full quantitative results across various I2V backbones on the VBench-I2V test set.

Backbone	Method	Dynamic Degree		VBench							CLIP-I	ViCLIP	Vision Reward
		VBench	VideoScore	Video Quality	SC	BC	TF	MS	AQ	IQ			
Wan 2.2-14B	Baseline	51.7	2.83	**58.0**	**94.7**	**94.9**	**96.4**	**98.0**	**64.3**	**69.6**	**94.4**	0.2614	0.143
	ALG	55.3	2.83	57.7	94.1	94.5	95.9	97.7	63.8	69.6	93.9	0.2616	0.140
	DyMoS	**64.8**	**2.84**	57.7	93.5	94.1	95.9	97.7	63.8	69.1	94.2	**0.2621**	**0.143**
Wan 2.1-14B	Baseline	48.7	3.21	**57.7**	**94.5**	**95.6**	**97.0**	**98.2**	**60.9**	**69.4**	**90.6**	0.2601	**0.132**
	ALG	52.5	3.18	57.2	92.6	94.9	96.8	98.1	60.5	67.7	89.3	0.2615	0.127
	DyMoS	**54.5**	**3.25**	57.4	93.2	94.9	96.5	98.0	60.5	68.9	89.4	**0.2618**	0.126
HunyuanVideo-1.5	Baseline	48.0	2.87	**58.2**	**95.0**	**94.9**	**97.7**	**99.1**	**61.2**	**70.2**	**90.5**	0.2624	0.140
	ALG	46.3	2.87	58.2	94.7	94.8	97.6	99.1	61.2	70.0	89.9	0.2629	0.141
	DyMoS	**55.3**	**2.88**	58.1	93.1	94.1	97.6	99.1	61.0	70.0	89.9	**0.2653**	**0.141**
CogVideoX-5B	Baseline	30.9	2.90	**57.8**	**95.3**	**95.8**	**97.2**	**98.6**	**60.5**	**69.6**	91.6	0.2612	**0.140**
	ALG	52.0	**2.98**	57.1	92.0	94.0	96.3	98.0	59.2	68.0	88.6	0.2647	0.127
	DyMoS	**52.9**	2.92	57.1	88.3	92.4	96.7	98.0	59.2	68.0	89.4	**0.2652**	0.131

D Additional experimental results

D.1 Full quantitative results

We report the full quantitative results for all four backbones evaluated on the VBench-I2V dataset in Tab. 8. Specifically, we report each scores of VBench [36] including Subject Consistency (SC), Background Consistency (BC), Temporal Flickering (TF), Motion Smoothness (MS), Aesthetic Quality (AQ) and Imaging Quality (IQ) in addition to the aggregate score. Since SC and BC can decrease as the generated motion becomes more dynamic, we additionally assess reference fidelity using per-frame CLIP-I similarity score. Despite producing stronger motion, DyMoS achieves comparable or higher CLIP-I scores than ALG across all backbones, suggesting a better balance between motion dynamics and reference fidelity.

D.2 Computational cost

Both ALG and DyMoS are training-free methods, but their runtime overhead comes from different sources. During its active interval, ALG requires an additional diffusion transformer forward pass for low-pass conditioned branch. In contrast, DyMoS keeps the original sampling loop unchanged. We use the official denoising steps in the official configuration of each backbone and the standard batched unconditional/conditional forward pass for classifier-free guidance.

The only extra computation in DyMoS is the self-attention score modification applied during the first $\lambda_{\text{Ours}}N$ denoising steps. This modification is fused into the attention kernel via PyTorch FlexAttention, and therefore adds only a lightweight element-wise operation on top of the existing attention computation, without any additional model forward pass.

Tab. 9 reports per-video inference time measured on a single NVIDIA A100 80GB GPU. Although DyMoS uses a larger active interval than ALG for most backbones, its amortized runtime overhead remains between 1.2% and 4.3%. By contrast, ALG incurs a larger overhead due to its additional forward pass, reaching up to 10.3% on Wan 2.1.

Table 9: Computational cost comparison of our method with ALG across I2V backbones.

	Wan 2.2	Wan 2.1	HV-1.5	CogVideoX
λ_{ALG}	0.10	0.20	0.04	0.04
λ_{Ours}	0.20	0.25	0.20	0.24
Baseline (s)	831	735	486	232
+ALG (s)	870 (+4.7%)	811 (+10.3%)	515 (+5.9%)	235 (+1.3%)
+Ours (s)	841 (+1.2%)	749 (+1.9%)	503 (+3.5%)	242 (+4.3%)

D.3 Ablation study on frame-wise modulation schedule

We further analyze the frame-wise modulation schedule $\phi(\cdot)$, which controls how the modulation strength varies across query frames in Eq. 6. Let $f(i) \in \{0, \ldots, F-1\}$ denote the latent-frame index of query token i. We compare three simple schedules:

$$\text{Uniform:} \quad \phi(f(i)) = 1, \tag{9}$$

$$\text{Linear:} \quad \phi(f(i)) = \frac{f(i)}{F-1}, \tag{10}$$

$$\text{Log:} \quad \phi(f(i)) = \frac{\log(1+f(i))}{\log(F)}. \tag{11}$$

The Uniform schedule applies the same modulation strength to all non-reference query frames. The Linear and Log schedules apply weaker modulation to frames closer to the reference frame and stronger modulation to later frames, with the Log schedule increasing more rapidly for early frames and saturating for later ones.

Table 10: Ablation study on frame-wise modulation schedule ϕ_f on Wan 2.2-14B.

ϕ_f	Dynamic Degree		Video Quality	ViCLIP	Vision Reward
	VBench	VideoScore			
Uniform	64.8	2.84	57.7	0.2621	0.143
Linear	56.1	2.82	57.9	0.2598	0.143
Log	60.2	2.80	57.7	0.2597	0.141

Tab. 10 reports the ablation results on Wan 2.2 under the same γ and λ settings as the main experiments. Among the evaluated schedules, the Uniform schedule achieves the highest Dynamic Degree while maintaining comparable Video Quality and text-video alignment. We therefore use the Uniform schedule as the default setting in the main experiments.

D.4 Motion evaluation across temporal segments

Early denoising steps and early video frames represent different axes of generation. Although DyMoS operates only during early denoising steps, its uniform frame schedule modulates queries from all non-reference frames. To examine whether the resulting motion enhancement is localized to specific temporal regions, we divide each generated video into four consecutive segments and measure the average top-5% optical-flow magnitude within each segment. As shown in Tab. 11, DyMoS consistently increases motion magnitude across all four temporal segments on Wan 2.2-14B and VBench-I2V. This indicates that the motion enhancement induced during early denoising steps persists throughout the generated video rather than being concentrated in the early frames.

Table 11: **Motion magnitude across temporal segments.** Average top-5% optical-flow magnitude on Wan 2.2-14B and VBench-I2V. Relative gains of DyMoS over the baseline are shown in parentheses.

Method	0–25%	25–50%	50–75%	75–100%
Baseline	12.74	13.78	14.55	15.77
DyMoS	16.30 (+28.0%)	18.15 (+31.7%)	18.90 (+29.9%)	20.42 (+29.4%)

D.5 Additional qualitative results

We provide additional qualitative comparisons and continuous-control examples.



Figure 7: **Additional comparison results with the vanilla baseline and ALG.** The leftmost images are the reference images. Our method qualitatively outperforms the vanilla baseline and ALG across various cases, demonstrating superior motion dynamics and visual fidelity. (Top) The vanilla baseline and ALG exhibit static scenes of a man riding a mountain bike. In contrast, our method generates fluid and natural riding movements. (Middle and Bottom) Moreover, our method successfully captures complex dynamic interactions, such as cars kicking up dust, while maintaining high visual quality.

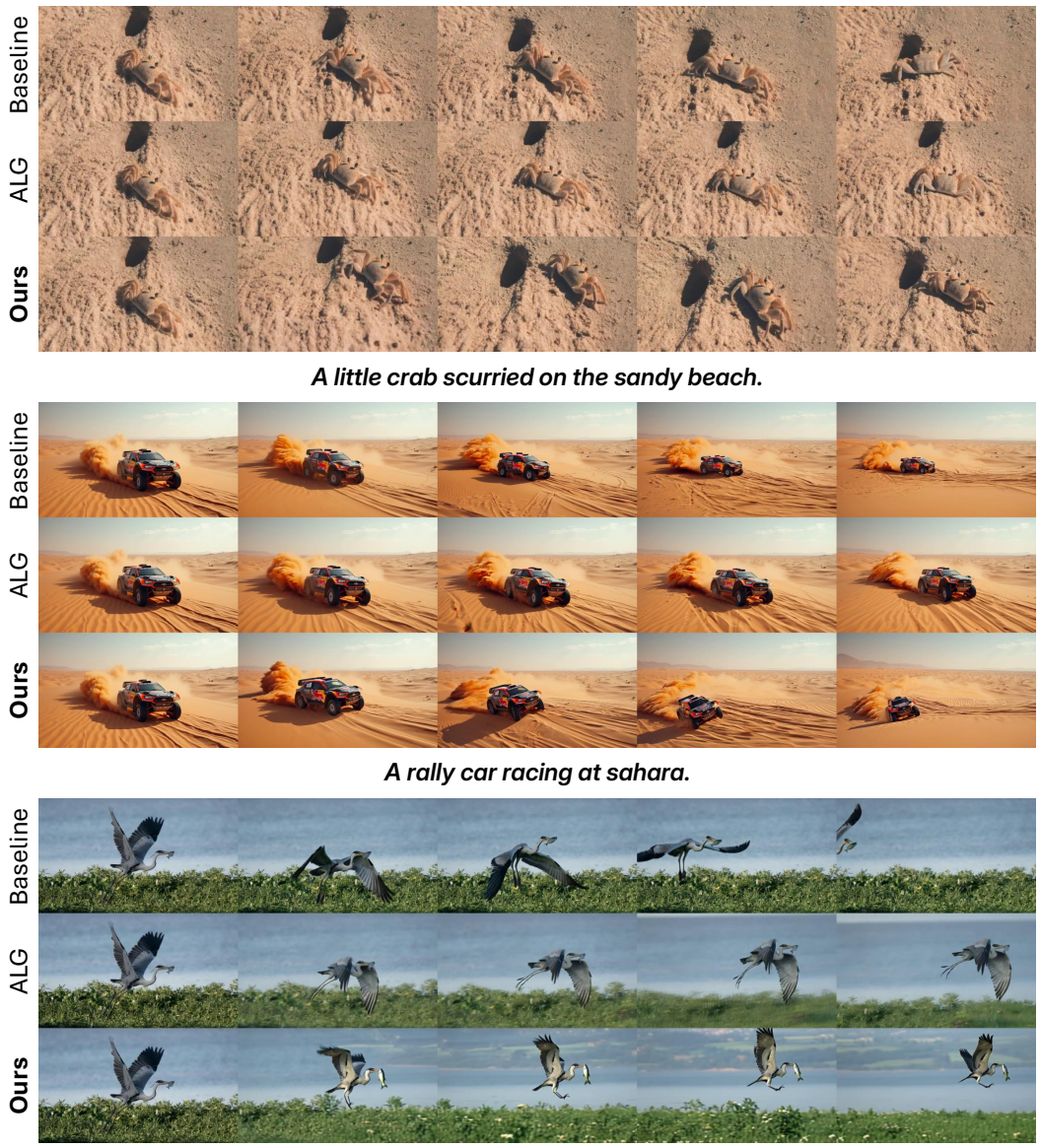


Figure 8: **Additional comparison results with the vanilla baseline and ALG.** The leftmost images are the reference images. (Top) ALG produces highly static scenes where the crab barely moves. In contrast, our method successfully synthesizes vivid and realistic motions. (Bottom) The vanilla baseline produces physically weird motions, such as the bird flying backwards. While ALG generates physically plausible movements, it fails to maintain visual fidelity, exhibiting noticeable distortions in the background.

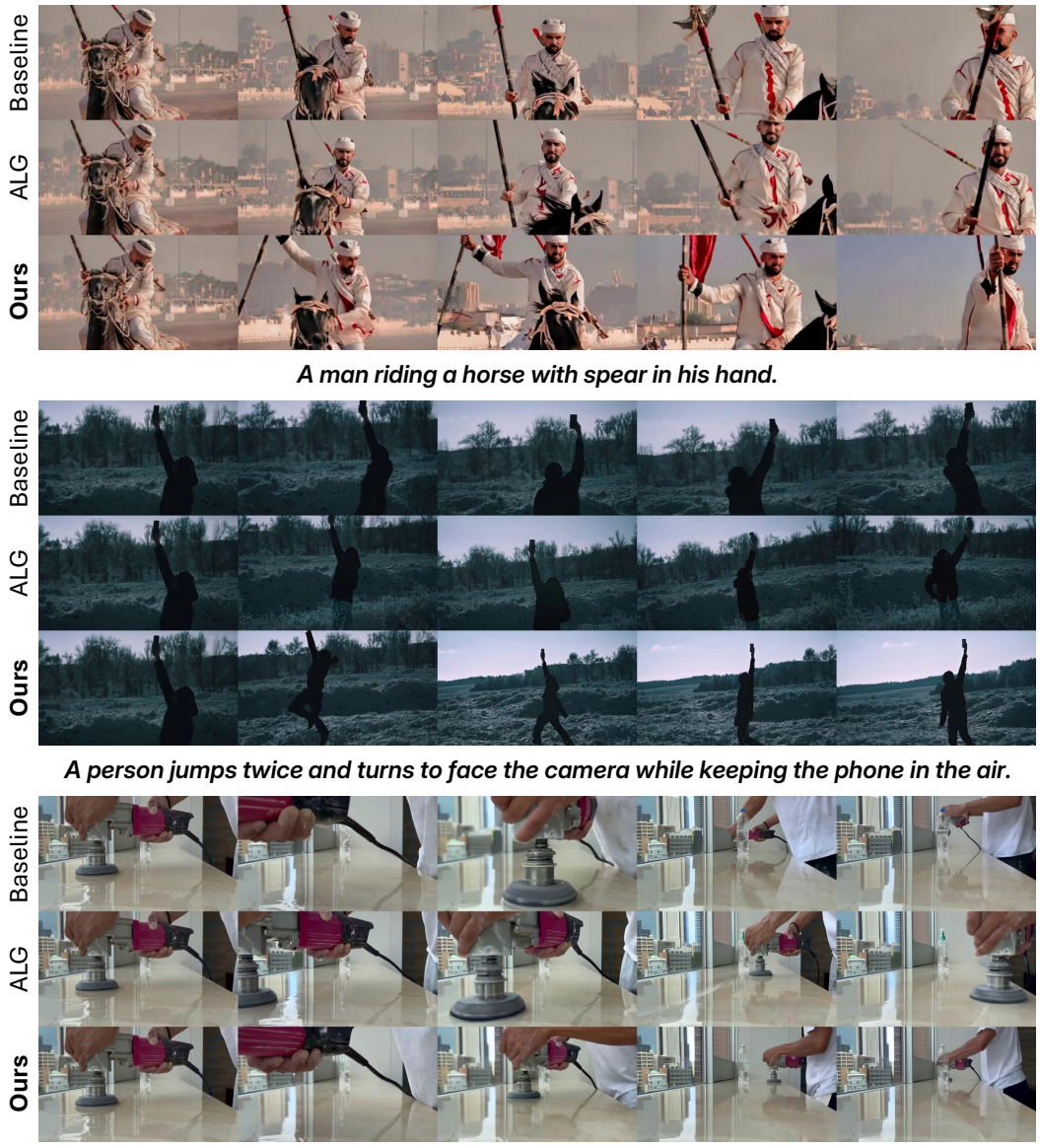


Figure 9: **Additional comparison results with the vanilla baseline and ALG.** The leftmost images are the reference images. (Top) Both the vanilla baseline and ALG struggle to synthesize dynamic motions, resulting in rigid and unnatural movements of the man. In contrast, our method generates natural motions of the man hoisting a spear. (Middle) Unlike the other methods, DyMoS successfully generates a semantically aligned video in which the person jumps twice. (Bottom) The vanilla baseline produces a physically implausible video where the person's arms penetrate the water bottle, whereas ALG simply fails to generate dynamic motions.

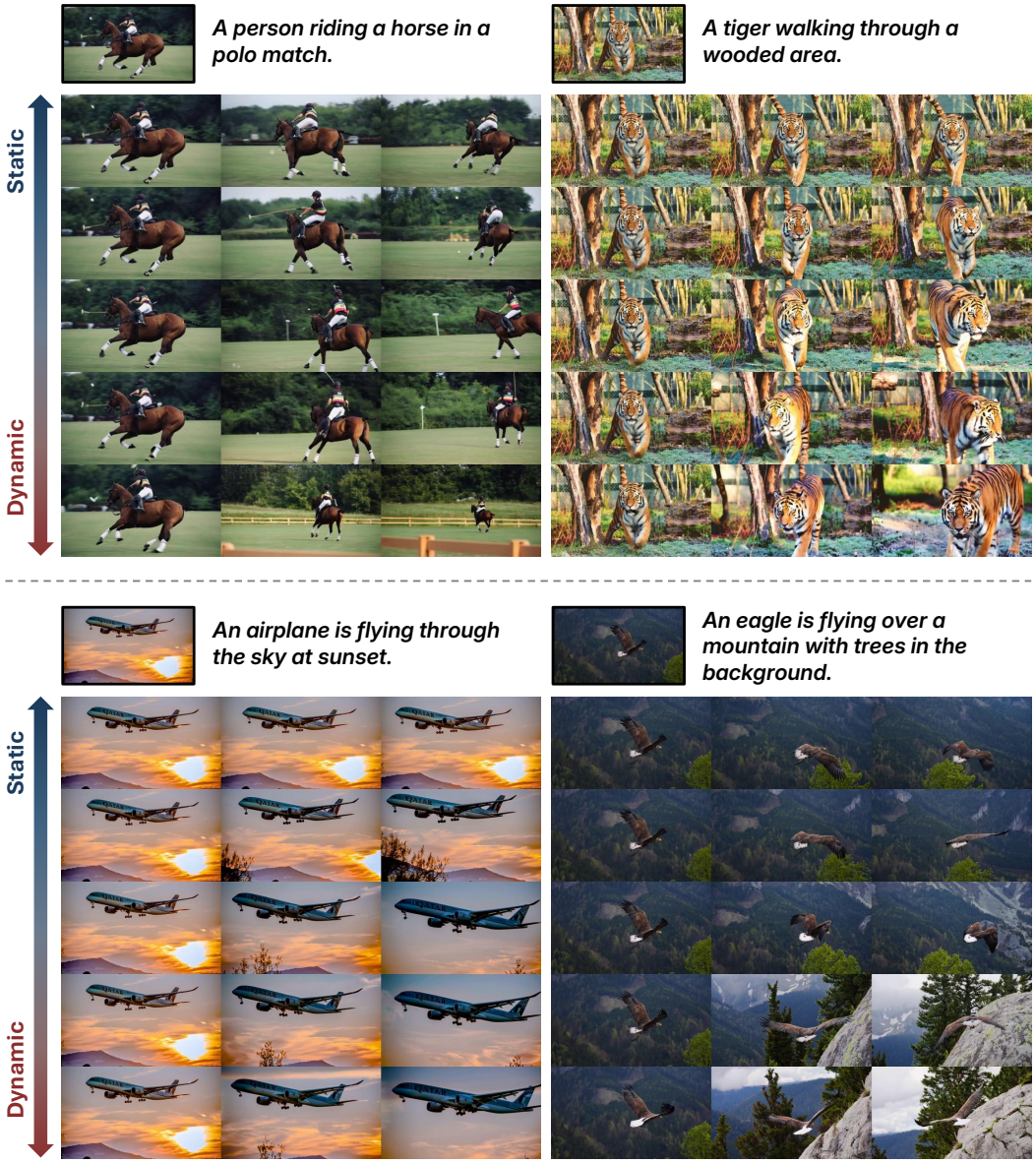


Figure 10: **Additional examples of continuous control over motion dynamics with DyMoS.** Rows from top to bottom correspond to $\gamma \in \{-2, -1, 0, 0.6, 1.0\}$, with $\gamma = 0$ denoting the baseline. (Top Left) The movement of the person riding a horse transitions naturally across the static-to-dynamic spectrum as the parameter increases. (Top Right) Our guidance smoothly scales the tiger's walking speed from slow to fast. Similarly, our method allows for dynamic control over the speeds of the airplane (Bottom Left) and the eagle (Bottom Right).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the scope of DyMoS, and we detail all claims throughout the main paper and the Appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses practical limitations of DyMoS, including the trade-off between motion dynamics and appearance consistency when the modulation strength is set too high.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Based on the observation presented in Sec. 3.1, we detail our hypothesis and provide detailed validation in Sec. 3.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the complete DyMoS sampling algorithm together with detailed experimental configurations in the Appendix. We release the code to reproduce the results upon completion of the blind review period.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the pseudo-code of DyMos in Appendix. A for reproducibility. Additionally, we publicly release our codes and instructions as stated above.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include all evaluation methods in the main paper and Appendix. We will include them with the code release as stated above.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports results across multiple backbones and datasets, but does not include error bars, confidence intervals, or statistical significance tests for the main quantitative results due to high computational cost on multi-seed generation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, please refer to Appendix. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our work currently focuses primarily on the motion dynamics of the videos, which is not directly associated with negative societal impacts. However, if reviewers find it necessary, we will include a discussion for potential harms with generative models, particularly regarding the controllable generation of motion dynamics.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new pretrained generative model, model checkpoint, or scraped dataset. DyMoS is an inference-time method applied to existing I2V backbones, and any code release will follow the licenses and usage restrictions of those underlying models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, the paper cites the original paper that produced dataset and models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper includes MTurk user study. The main paper describes the blinded and randomized comparison protocol, the evaluated criteria, and the number of image–prompt pairs and responses, while Appendix C.4 provides the participant-facing questions and detailed evaluation protocol. Method names are hidden, video order is randomized, and results are reported only as aggregate preference fractions.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The paper includes an MTurk preference study, but no IRB approval, exemption, or equivalent institutional review was obtained. The study involves low-risk video preference judgments and reports only aggregate results.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The paper does not involve LLMs as part of proposed method, experiments, or core contributions.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.