SELF-REFLECTING VIDEO DIFFUSION WITH BOOSTED MOTION DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern video diffusion models have demonstrated strong performance, enabling the generation of high-fidelity and temporally coherent videos. A key component behind this success is temporal attention, which enables tokens to exchange information across frames to maintain consistency over time. However, we argue that this very mechanism also contributes to *limited motion dynamics*, a persistent failure mode where generated videos remain overly static. Specifically, we find that weak motion is closely related to tokens over-attending to other frames, a pattern that emerges as early as the initial denoising steps. Building on this observation, we propose *Self-Reflecting Optimization (SRO)*, a training-free framework that optimizes the initial noise to encourage tokens to attend to their own frame, while keeping it close to the original sample and consistent with the Gaussian prior. Since SRO modifies neither the model nor the conditioning signals, it applies to both text-to-video (T2V) and image-to-video (I2V) generation, in stark contrast to existing approaches tied to specific conditioning types. Extensive experiments on both T2V and I2V backbones demonstrate that SRO substantially boosts motion dynamics with only marginal degradation in text-video alignment. **Our project page, including additional demo videos, is available at** `https://sro-psi.vercel.app/`.

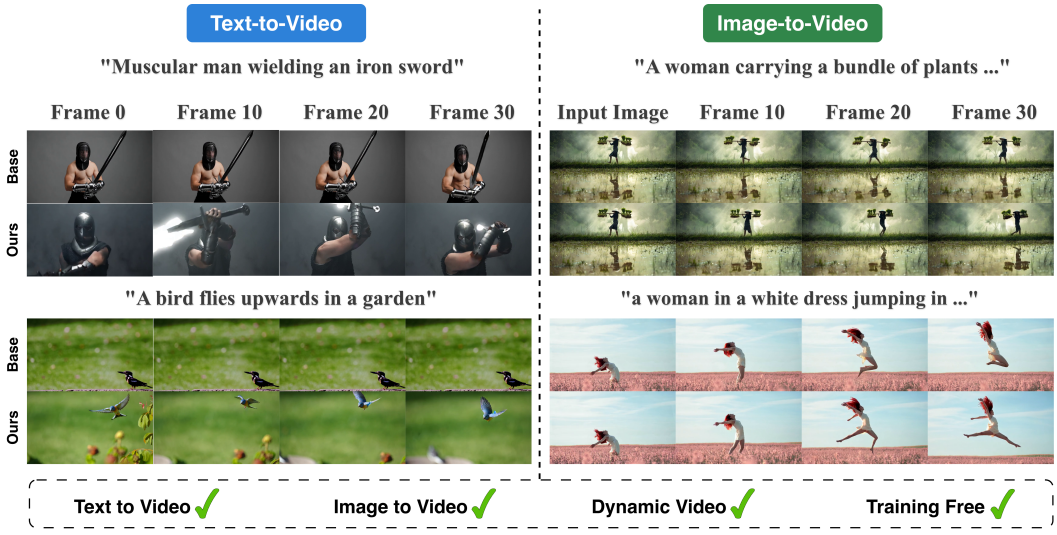


Figure 1: **Example results of Self-Reflecting Optimization (SRO).** Video diffusion models often suffer from limited motion dynamics, yielding visually static outputs ("Base") across both text-to-video (T2V) and image-to-video (I2V) models. In contrast, SRO addresses this issue in a training-free manner and applies to both T2V and I2V backbones, producing videos with noticeably richer motion while preserving the given conditions ("Ours").

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Diffusion-based video generators (Ho et al., 2022; Wan et al., 2025) have demonstrated remarkable performance, producing high-fidelity synthesis that faithfully respects user-provided conditions such as text prompts or reference images. Beyond high-quality individual frames, they also achieve strong temporal consistency by preserving object identity, appearance, and scene structure over time. A key ingredient behind this coherence is *temporal attention* (Ho et al., 2022), which enables video tokens at different timesteps to exchange information, thereby suppressing abrupt appearance changes and structural inconsistencies across frames. Most video diffusion frameworks rely on temporal attention for time coherency (Chen et al., 2023; Yang et al., 2025; Wan et al., 2025), highlighting its central role in video generation.

Despite the remarkable progress of video diffusion models, *limited motion dynamics* remains a fundamental bottleneck. As illustrated in Fig. 1, even under strongly dynamic conditions (e.g., "A bird flies upwards in a garden."), generated videos often exhibit only subtle frame-to-frame variation, resulting in visually static outputs (Ruan et al., 2024; Shaulov et al., 2026). This failure mode is broadly observed across both text-to-video (T2V) (Kim et al., 2026b) and image-to-video (I2V) (Choi et al., 2026) models. Existing motion-enhancement methods partially alleviate this problem, but introduce practical constraints: they are often coupled to a specific modality (i.e., applicable to only one of T2V or I2V (Choi et al., 2026; Kim et al., 2026b)) and some require additional machinery, such as external LLMs (Choi et al., 2026; Kim et al., 2026b).

In this work, we develop a simple and unified motion-boosting approach that operates independently of the conditioning modality. Our framework begins with a key observation: temporal attention, while essential for maintaining temporal coherence, can suppress motion when tokens over-attend to other frames. To quantify this behavior, we introduce the *self-reflecting score*, which measures the fraction of attention that tokens assign to their own frame. We find that a lower self-reflecting score is correlated with more limited motion in generated videos; see Fig. 3 for illustration. Moreover, this pattern emerges already at the onset of denoising: the self-reflecting score measured at the first timestep is strongly associated with the motion dynamics of the final video.

Based on this observation, we develop *Self-Reflecting Optimization (SRO)*, which optimizes a perturbation of the initial noise to increase the self-reflecting score at an early Transformer block, thereby boosting motion dynamics. To keep the optimized noise a valid initialization, SRO confines the perturbation to a low-frequency trust region around the original noise and re-standardizes the result to match the Gaussian prior. The optimized noise is then used to initialize the original sampler, without modifying the pretrained model or conditioning signals.

To validate the effectiveness of SRO, we conduct extensive experiments on various conditional video diffusion frameworks, including Wan2.1 (Wan et al., 2025) and CogVideoX (Yang et al., 2025). Across both T2V and I2V settings, SRO substantially boosts motion dynamics at only a marginal cost in text-video alignment, achieving state-of-the-art performance.

Our contributions are summarized as follows:

- We identify that weak motion is closely related to tokens over-attending to other frames in temporal attention. To quantify this, we introduce the *self-reflecting score* and show that a higher score is consistently associated with richer motion across multiple motion metrics, even when measured at the onset of denoising.
- Based on this finding, we propose *Self-Reflecting Optimization (SRO)*, a simple training-free method that optimizes the initial noise to increase the self-reflecting score at an early Transformer block, without modifying model parameters or conditioning signals.
- Extensive experiments on both T2V and I2V backbones demonstrate that SRO substantially boosts motion dynamics at only a marginal cost in text-video alignment, achieving state-of-the-art performance.

2 RELATED WORK

Video Diffusion Models. Diffusion models have been extended from image synthesis to video generation by explicitly modeling temporal dependencies across frames (Ho et al., 2020; 2022). Early

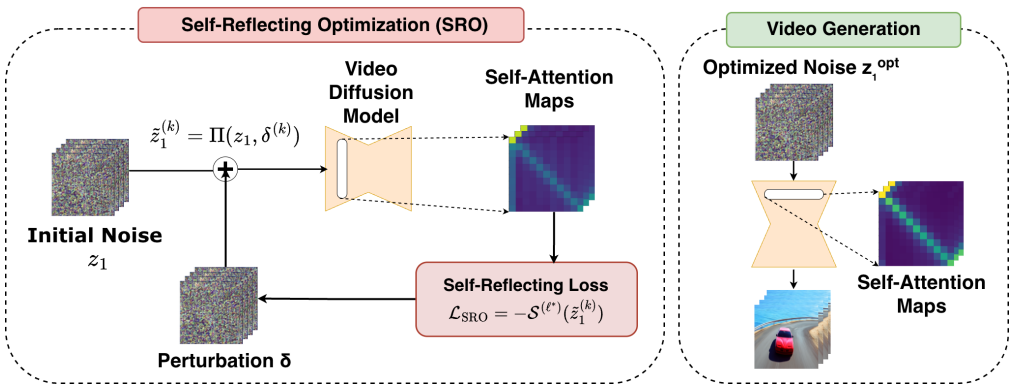


Figure 2: **Overview of SRO.** (Left) At each step k, the perturbation $\delta^{(k)}$ is combined with the initial noise z_1 through the constraint operator Π, and the resulting candidate $\tilde{z}_1^{(k)}$ is fed to the video diffusion model at the first denoising step. From the self-attention maps at block ℓ^*, we compute the self-reflecting score $\mathcal{S}^{(\ell^*)}$, i.e., the fraction of attention that tokens assign to their own frame, whose negative serves as the loss $\mathcal{L}_{\text{SRO}}$ for updating $\delta^{(k)}$. (Right) After K steps, the optimized noise $z_1^{\text{opt}} := \Pi(z_1, \delta^{(K)})$ initializes the original sampler. The resulting self-attention concentrates more strongly within each frame, as shown by the more pronounced diagonal, leading to richer motion without modifying the model parameters, conditioning signals, or sampling procedure.

approaches augment image diffusion architectures with temporal convolution or attention, while recent models increasingly adopt Transformer-based architectures that jointly model spatial and temporal tokens. Large-scale models such as CogVideoX (Yang et al., 2025) and Wan (Wan et al., 2025) demonstrate strong generation quality under both text and image conditioning. A key component of these models is cross-frame self-attention, which promotes temporal consistency by exchanging information across frames. However, excessive temporal dependency may also suppress frame-wise variation and motion. In this work, we investigate how temporal attention behavior at the beginning of generation relates to the motion dynamics of the final video.

Enhancing Motion Dynamics in Video Generation. A line of work has investigated inference-time strategies for enhancing motion without retraining video diffusion models. For instance, ALG (Choi et al., 2026) weakens image conditioning during early denoising through adaptive low-pass filtering, allowing larger temporal variations in image-to-video generation. MotionCFG (Kim et al., 2026b) instead perturbs motion-related text embeddings and modifies classifier-free guidance to encourage dynamic generation in text-to-video models. While effective, these methods intervene in modality-specific conditioning pathways. In contrast, our method operates directly on temporal attention, providing a conditioning-agnostic mechanism applicable to both text-to-video and image-to-video.

Noise Optimization for Diffusion Models. Adjusting the initial noise of diffusion models has been recently explored as a lightweight alternative to fine-tuning, since the initial noise largely determines the final generation outcome (Qi et al., 2024; Xu et al., 2025). One instance is InitNO (Guo et al., 2024), which optimizes the initial latent using cross-attention and self-attention-based objectives to improve semantic fidelity. On the other hand, FreeInit (Wu et al., 2024) investigates initial noise in video diffusion models from a frequency-domain perspective and iteratively refines its low-frequency components to bridge the training-inference initialization gap, thereby improving temporal consistency. More recently, CNO (Kim et al., 2026a) optimizes initial noise in the Tweedie prediction space with a contrastive objective to enhance sample diversity while preserving fidelity. However, our work targets *motion dynamics* in video generation by optimizing the initial noise according to temporal self-attention behavior observed at the beginning of generation, thereby encouraging attention patterns associated with more dynamic videos.

3 Method

3.1 Background and Preliminaries

Video Diffusion Models. Recent video diffusion models are commonly formulated with latent flow matching (Lipman et al., 2023; Wan et al., 2025), where generation follows a continuous transport

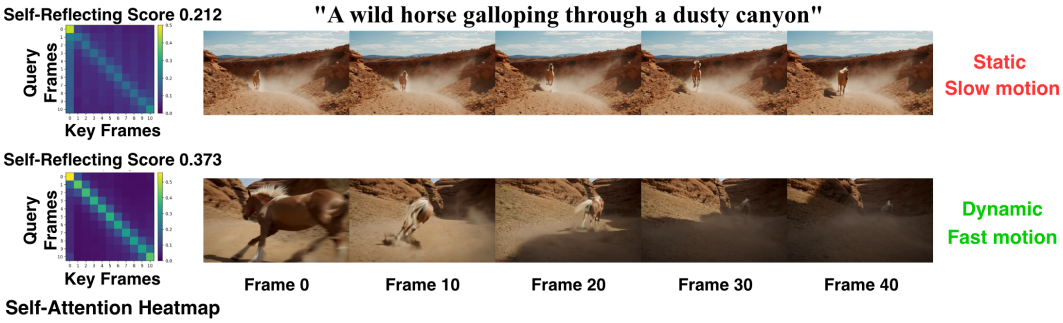


Figure 3: **Self-reflecting score and motion dynamics.** Frame-level attention maps at the first denoising step (left) and the corresponding generated videos (right) for the same prompt. Each map shows the attention mass between query and key frames, where the diagonal corresponds to within-frame attention. The static video (top) exhibits a lower self-reflecting score than the dynamic video (bottom).

path between the data distribution and Gaussian noise in the latent space of a VAE (Kingma & Welling, 2014). Let $x_0 \sim p_{\text{data}}$ denote a clean video sample and $z_0 := E(x_0)$ its latent representation obtained via a VAE encoder $E(\cdot)$. Given Gaussian noise $z_1 \sim \mathcal{N}(0, I)$, a linear interpolation path between z_0 and z_1 is defined as

$$z_t = (1-t)z_0 + tz_1, \qquad t \in [0,1]. \tag{1}$$

Along this path, the target velocity is $\mathrm{d}z_t/\mathrm{d}t = z_1 - z_0$. A velocity model $v_\theta(z_t, t, c)$, conditioned on c (e.g., a text prompt or a reference image), is trained to regress this target:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\, z_0,\, z_1}\left[\|v_\theta(z_t,t,c) - (z_1 - z_0)\|_2^2\right]. \tag{2}$$

At inference, generation starts from Gaussian noise $z_1 \sim \mathcal{N}(0, I)$ and numerically solves the ODE $\mathrm{d}z_t/\mathrm{d}t = v_\theta(z_t, t, c)$ from $t = 1$ to $t = 0$. The resulting latent is then decoded into a video by the VAE decoder, i.e., $\hat{x}_0 = D(z_0)$. We denote this entire sampling procedure by $\hat{x}_0 = \mathcal{G}_\theta(z_1, c)$.

Temporal Self-Attention. Transformer-based video diffusion models (Yang et al., 2025; Wan et al., 2025) represent a video latent as N spatio-temporal tokens spanning F latent frames, and model their interactions through self-attention (Vaswani et al., 2017). For head h of a Transformer block ℓ, the attention matrix is

$$A^{(\ell,h)} = \text{Softmax}\left(\frac{Q^{(\ell,h)}{K^{(\ell,h)}}^\top}{\sqrt{d_h}}\right), \tag{3}$$

where $A_{ij}^{(\ell,h)}$ denotes the attention assigned by query token i to key token j. Let $f(i) \in \{1, \ldots, F\}$ denote the frame index of token i. Attention between tokens with $f(i) \neq f(j)$ allows information to flow across frames, which is the key mechanism for preserving object identity, appearance, and scene structure over time.

Although effective in encouraging temporal consistency, such cross-frame interaction may also tie each frame too closely to its neighbors, potentially leading to side effects such as limited motion dynamics. We investigate this point in depth in Sec. 3.2.

3.2 Connecting Cross-Frame Attention to Limited Motion

Here we analyze how the balance between within-frame and cross-frame attention relates to the motion dynamics of generated videos. To this end, we first introduce a metric that quantifies this balance, and then present our empirical observations.

Self-Reflecting Score. To quantify how much each frame relies on itself rather than on other frames, we measure the attention mass that each token assigns to tokens within its own frame. Aggregating over all tokens and heads of block ℓ yields the *self-reflecting score*:

$$\mathcal{S}^{(\ell)} := \frac{1}{HN}\sum_{h=1}^{H}\sum_{i=1}^{N}\sum_{j:\, f(j)=f(i)} A_{ij}^{(\ell,h)}. \tag{4}$$

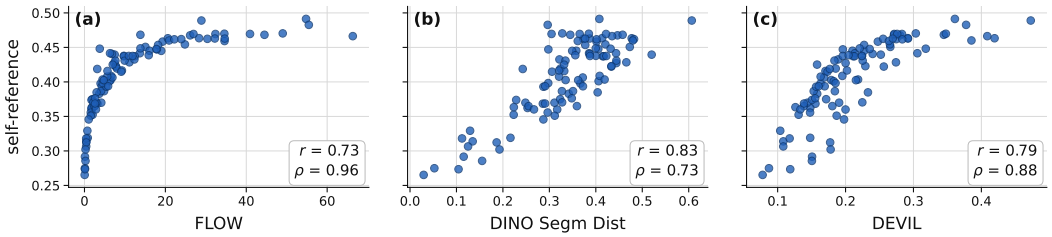


Figure 4: **Self-reflecting score of the initial noise correlates with the motion of the generated video.** Using Wan2.1-1.3B (Wan et al., 2025), we generate 100 videos from a single prompt by varying only the seed, and plot the self-reflecting score at block 2 against (a) optical-flow magnitude (FLOW), (b) DINO segment distance, and (c) the DEVIL dynamics score. All three metrics show a strong positive relationship. See Sec. A.4 for analyses on other backbones (Fig. 7).

A larger $\mathcal{S}^{(\ell)}$ indicates stronger within-frame attention, whereas a smaller value represents greater dependence on other frames. For reference, uniform attention over all tokens gives $\mathcal{S}^{(\ell)} = 1/F$.

Illustrative Example. Given the self-reflecting score, we examine how it relates to the motion dynamics of generated videos. We use Wan2.1 (Wan et al., 2025) as the backbone and compute $\mathcal{S}^{(\ell)}$ at the first denoising step ($t = 1$), i.e., directly from the initial noise z_1 and the condition c. Unless otherwise stated, we use an early Transformer block $\ell^* = 2$. Fig. 3 compares two videos generated from the same dynamic prompt with different seeds. The static video exhibits strong cross-frame attention with a self-reflecting score of 0.212, while the dynamic video concentrates attention within each frame, yielding a score of 0.373. We highlight that this difference is already visible at the first denoising step, before visual content has fully formed.

Quantitative Analysis. To examine this relationship more systematically, we employ Wan2.1-1.3B (Wan et al., 2025) to generate 100 videos from a single prompt by varying only the random seed. We then compare $\mathcal{S}^{(\ell^*)}$ at $t = 1$ with the motion of each final video. We measure motion using three complementary metrics: optical-flow magnitude (FLOW), DINO segment distance, and the DEVIL dynamics score (Liao et al., 2024). As shown in Fig. 4, the self-reflecting score is strongly and positively correlated with all three metrics (Pearson $r \geq 0.73$, Spearman $\rho \geq 0.72$). Since the model input at $t = 1$ is determined solely by z_1 given c, these results suggest that the initial noise already encodes much of the motion of the final video, and that motion can be promoted by steering z_1 toward a higher self-reflecting score. We observe a similar trend in other video backbones; see Sec. A.4 for details (Fig. 7).

3.3 Self-Reflecting Optimization: Steering Initial Noise toward Dynamic Motion

Constrained Noise Optimization. Based on our observation, we propose *Self-Reflecting Optimization (SRO)*, which adjusts the sampled initial noise $z_1 \sim \mathcal{N}(0, I)$ so that the early attention of the model becomes more self-reflecting. One natural question is how to formulate this adjustment. A naive approach is to directly maximize the self-reflecting score with respect to the noise:

$$z_1^* = \arg\max_z \mathcal{S}^{(\ell^*)}(z), \tag{5}$$

where $\mathcal{S}^{(\ell^*)}(z)$ denotes the score at block ℓ^* when z is fed to the model at $t = 1$ with condition c, and the optimization is initialized at z_1. However, this unconstrained formulation has two issues. First, nothing prevents the optimized noise from drifting away from the Gaussian prior $\mathcal{N}(0, I)$. Since the pretrained model is trained only on inputs drawn from this prior, such out-of-distribution noise can degrade the quality of the generated video (Guo et al., 2024). Second, the attention score can be increased through fine-grained patterns that have little to do with actual motion, e.g., high-frequency components that the model treats as frame-specific details; see Table 4 for evidence for these points.

To address these issues, we resort to optimizing a perturbation δ around z_1 under explicit constraints. Specifically at each optimization step, we map the perturbed noise to a feasible candidate $\tilde{z}_1 = \Pi(z_1, \delta)$, where Π is a constraint operator that restricts the perturbation to a low-frequency trust

region around z_1 and re-standardizes the result to match the Gaussian prior (detailed below). We then compute $\mathcal{S}^{(\ell^*)}(\tilde{z}_1)$ via Equation 4 and update δ by gradient descent on

$$\mathcal{L}_{\mathrm{SRO}}(\delta) \coloneqq -\mathcal{S}^{(\ell^*)}(\Pi(z_1, \delta)) \,. \tag{6}$$

Since Π is applied at every step, the gradient flows through the constraints, and the optimization only explores perturbations that remain valid initializations. As a result, SRO does not search for an entirely new noise sample, but finds a constrained perturbation around z_1, shifting its attention behavior toward patterns associated with stronger motion while largely preserving the overall structure of the original initialization.

Constraint Operator. The operator Π is designed to address the two issues above, and consists of three operations applied in sequence. To prevent the optimization from exploiting fine-grained patterns (the second issue), we first restrict the perturbation to spatial low-frequency components of each latent frame:

$$\delta_{\mathrm{lp}} \coloneqq \mathcal{F}_{2D}^{-1}\left(M_\kappa \odot \mathcal{F}_{2D}(\delta)\right), \tag{7}$$

where $\mathcal{F}_{2D}$ denotes the 2D Fourier transform and M_κ is a binary mask retaining a fraction κ of the lowest spatial frequencies. To keep the candidate near the original sample (the first issue), we then bound the perturbation magnitude with an ℓ_2 trust region:

$$\delta_{\mathrm{tr}} \coloneqq \delta_{\mathrm{lp}} \cdot \min\left(1, \frac{\varepsilon\sqrt{N_\delta}}{\|\delta_{\mathrm{lp}}\|_2}\right), \tag{8}$$

where N_δ is the number of elements of δ, so that ε bounds the root-mean-square deviation per element. Finally, since even a bounded perturbation can shift the statistics of the noise, we re-standardize the perturbed noise and clip extreme values:

$$\Pi(z_1, \delta) \coloneqq \mathrm{clip}\left(\frac{z_1 + \delta_{\mathrm{tr}} - \mu}{\sigma}, -3, 3\right), \tag{9}$$

where μ and σ are the mean and standard deviation of $z_1 + \delta_{\mathrm{tr}}$. All three operations are differentiable almost everywhere, allowing gradients of Equation 6 to propagate back to δ.

Efficient Optimization and Generation. Since $\mathcal{S}^{(\ell^*)}$ depends only on the attention at block ℓ^*, both forward and backward passes are truncated at ℓ^* rather than running the entire Transformer, which keeps the optimization cost low. After K steps, we obtain the optimized noise $z_1^{\mathrm{opt}} \coloneqq \Pi(z_1, \delta^{(K)})$ and generate the final video with the original sampler, i.e., $\hat{x}_0 = \mathcal{G}_\theta(z_1^{\mathrm{opt}}, c)$. No gradient computation, attention modification, or guidance is applied during denoising. SRO therefore leaves the pretrained model, the conditioning signals, and the sampling procedure unchanged, which makes it directly applicable to both T2V and I2V generation. We leave pseudocode of SRO in Sec. A.1; see Alg. 1.

4 Experiments

Models and Datasets. To validate SRO across conditioning modalities, we evaluate it on both text-to-video (T2V) and image-to-video (I2V) generation. For T2V, we use Wan2.1-1.3B, Wan2.1-14B (Wan et al., 2025), and CogVideoX-5B (Yang et al., 2025); for I2V, we use Wan2.1-I2V (Wan et al., 2025). For T2V evaluation, we use general-purpose prompts from T2V-CompBench (Sun et al., 2025). In addition, following Kim et al. (2026b), we use motion-intensive prompts generated by Gemini, each describing explicit motion (see Sec. A.3 for details). For I2V evaluation, we use VBench-I2V (Huang et al., 2024).

Metrics. We evaluate generated videos in terms of visual quality, text alignment, and motion dynamics. For visual quality and text alignment, we report CLIPScore (Hessel et al., 2021) and PickScore (Kirstain et al., 2023), computed on individual frames and averaged over each video, together with X-CLIP (Ma et al., 2022), which measures the alignment between the text prompt and the entire video sequence. For motion dynamics, we use RAFT Flow (Teed & Deng, 2020), DINO-Segm (Mumuni & Mumuni, 2024), and DEVIL (Liao et al., 2024), which capture pixel-level motion, semantic variation, and overall temporal dynamics, respectively. We additionally report DINO similarity (Caron et al., 2021) and SSIM (Wang et al., 2004) between adjacent frames, where lower values indicate larger semantic and structural variation across frames, respectively. Since automatic

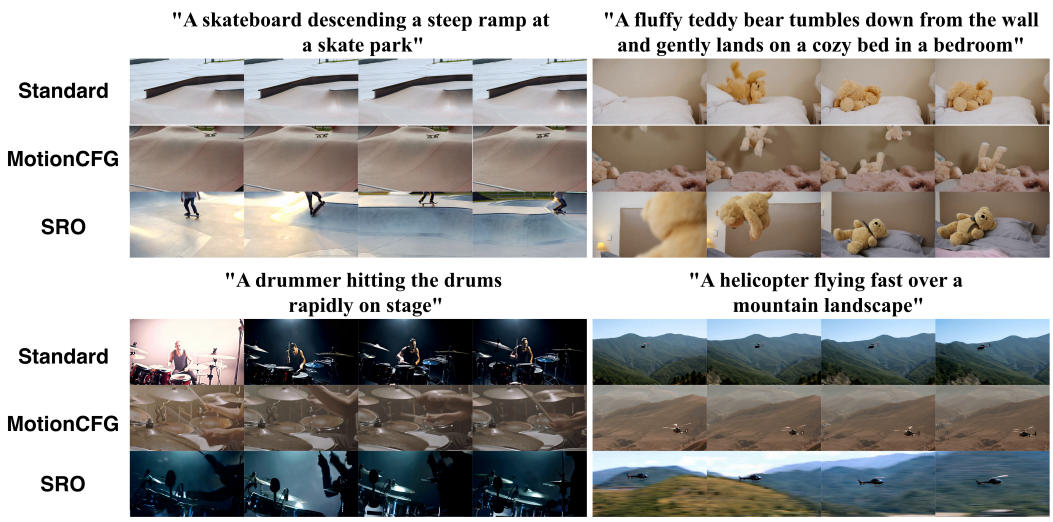


Figure 5: **Qualitative comparison on text-to-video generation.** Videos generated by the standard sampler, MotionCFG (Kim et al., 2026b), and SRO from the same prompts and initial noise. SRO produces noticeably richer motion while preserving the visual quality and text alignment of the base model. Additional video samples are available on our project page.



Figure 6: **Qualitative comparison on image-to-video generation.** Videos generated by the base model, ALG (Choi et al., 2026), and SRO on Wan2.1-I2V, conditioned on the same input image (first frame) and text prompt. The standard baseline often produces limited motion, whereas SRO yields noticeably larger and more natural movements while remaining consistent with the input image. Additional video samples are available on our project page.

metrics may not fully reflect perceived quality, we further conduct a user study to validate our results with human judgments; see Sec. A.6 for details.

For I2V evaluation, we additionally report six metrics from VBench (Huang et al., 2024): Dynamic Degree (Teed & Deng, 2020), I2V Subject (Oquab et al., 2023), MUSIQ (Ke et al., 2021), Aesthetic Score (Schuhmann et al., 2022), Subject Consistency (Oquab et al., 2023), and Temporal Flicker. Dynamic Degree measures the magnitude of generated motion, while I2V Subject evaluates preservation of the subject specified by the conditioning image. MUSIQ and Aesthetic Score assess perceptual image quality, whereas Subject Consistency and Temporal Flicker measure temporal identity preservation and frame-to-frame stability, respectively.

Table 1: **Quantitative comparison on T2V-CompBench prompts (Sun et al., 2025).** We compare SRO with the standard sampler, FlowMo (Shaulov et al., 2026) and MotionCFG (Kim et al., 2026b), the current state-of-the-art. SRO improves motion dynamics over the baselines while maintaining competitive text alignment and visual quality. **Bold**: best; underline: second-best within each model.

Model	Method	Dynamics / Temporal					Framewise		Full Video
		RAFT Flow ↑	DEVIL ↑	DINO-Segm ↑	DINO ↓	SSIM ↓	CLIPScore ↑	PickScore ↑	X-CLIP ↑
Wan2.1-1.3B	Standard	3.47	0.146	0.155	0.966	0.974	**27.72**	20.93	22.86
	FlowMo	1.97	0.132	0.138	0.969	0.966	27.61	**20.95**	22.86
	MotionCFG	5.45	0.165	0.188	0.957	0.964	27.17	20.88	23.09
	SRO (Ours)	**9.09**	**0.225**	**0.320**	**0.941**	**0.910**	27.21	20.91	**23.78**
Wan2.1-14B	Standard	3.63	0.147	0.158	0.962	0.960	**28.52**	**21.27**	23.20
	MotionCFG	7.54	0.179	0.205	0.951	0.958	28.00	21.14	23.08
	SRO (Ours)	**14.54**	**0.249**	**0.352**	**0.933**	**0.893**	27.94	21.17	**23.85**
CogVideoX-5B	Standard	6.32	0.196	0.278	0.948	0.830	**27.82**	**21.17**	24.42
	MotionCFG	6.91	0.226	0.424	0.936	0.685	26.78	20.90	24.86
	SRO (Ours)	**10.73**	**0.252**	**0.479**	**0.928**	**0.548**	27.61	20.96	**25.09**

Table 2: **Quantitative comparison on motion-intensive prompts.** Using motion-intensive prompts generated by Gemini (see Sec. A.3 for details), we compare SRO with the standard sampler, FlowMo (Shaulov et al., 2026) and MotionCFG (Kim et al., 2026b), the current state-of-the-art. Even on these motion-demanding prompts, SRO demonstrates consistent improvements in dynamics across all three backbones **Bold**: best; underline: second-best within each model.

Model	Method	Dynamics / Temporal					Framewise		Full Video
		RAFT Flow ↑	DEVIL ↑	DINO-Segm ↑	DINO ↓	SSIM ↓	CLIPScore ↑	PickScore ↑	X-CLIP ↑
Wan2.1-1.3B	Standard	12.86	0.232	0.262	0.937	0.923	26.27	20.67	21.62
	FlowMo	3.21	0.172	0.188	0.954	0.950	**26.63**	**20.89**	21.80
	MotionCFG	24.84	0.271	0.320	0.923	0.918	25.96	20.74	**22.41**
	SRO (Ours)	**50.84**	**0.468**	**0.548**	**0.847**	**0.796**	25.86	20.33	22.08
Wan2.1-14B	Standard	17.97	0.231	0.278	0.928	0.887	**27.46**	**21.19**	22.48
	FlowMo	8.98	0.212	0.234	0.938	0.920	27.41	21.06	22.07
	MotionCFG	22.41	0.262	0.323	0.929	0.918	26.89	21.09	22.87
	SRO (Ours)	**65.16**	**0.475**	**0.555**	**0.857**	**0.781**	26.47	20.62	**22.99**
CogVideoX-5B	Standard	15.49	0.276	0.354	0.929	0.772	27.19	**21.06**	22.74
	FlowMo	6.93	0.216	0.288	0.944	0.855	25.92	20.49	22.62
	MotionCFG	8.92	0.251	**0.446**	0.934	**0.677**	26.11	20.78	**23.30**
	SRO (Ours)	**15.88**	**0.288**	0.372	**0.925**	0.756	**27.21**	20.96	22.75

4.1 Qualitative Results

Text-to-Video Generation. Fig. 5 compares SRO with the baseline and MotionCFG across diverse text-to-video examples. The default baseline often produces nearly static outputs despite strongly dynamic prompts, while MotionCFG yields moderate improvements in temporal variation. In contrast, SRO generates substantially larger object displacement, pose changes, and scene-level motion while preserving the semantic content and overall appearance of the generated video. See our project page for additional video samples.

Image-to-Video Generation. Fig. 6 demonstrates that the same behavior extends to image-conditioned generation. The default I2V model tends to remain close to the reference image, resulting in limited motion, while ALG partially alleviates this effect by weakening early image conditioning. SRO instead enhances motion directly through initial-noise optimization, without modifying the conditioning image. As a result, SRO produces richer temporal evolution while largely preserving subject identity and scene appearance. Additional video samples are available on our project page.

4.2 Quantitative Results

Text-to-Video Generation. Table 1 evaluates SRO on general-purpose prompts from T2V-CompBench (Sun et al., 2025). We compare SRO with MotionCFG (Kim et al., 2026b), the current state-of-the-art, and FlowMo (Shaulov et al., 2026), another training-free motion enhancement method. SRO consistently improves motion dynamics over the baselines across all backbones while largely preserving frame-level quality.

Table 2 further reports results on motion-intensive prompts generated by Gemini. SRO again improves motion dynamics while maintaining comparable text alignment and visual quality, suggesting that it remains effective even when the prompt demands substantial temporal variation. We also

Table 3: **Quantitative comparison on image-to-video generation.** We compare SRO with the base sampler and ALG (Choi et al., 2026), the current state-of-the-art for I2V motion enhancement, on Wan2.1-I2V. SRO improves motion dynamics over both methods while maintaining comparable visual quality and temporal consistency. **Bold**: best; underline: second-best.

Method	Dynamic Degree ↑	RAFT Flow ↑	Motion ↑	I2V Subj. ↑	MUSIQ ↑	Aesthetic ↑	Subj. Cons. ↑	Temp. Flicker ↓
Standard	22.0	2.19	0.0287	**0.9430**	**68.26**	**5.301**	**0.9756**	**0.0062**
ALG	30.0	3.30	0.0349	0.9331	67.83	5.276	0.9711	0.0083
SRO	**38.0**	**4.01**	**0.0389**	0.9290	67.26	5.265	0.9693	0.0091

Table 4: **Ablation of SRO components.** Each component is removed from the full configuration. **Bold**: best; underline: second-best.

Variant	RAFT Flow ↑	DEVIL ↑	DINO ↓	SSIM ↓	CLIPScore ↑	PickScore ↑
SRO (Full)	**84.21**	0.363	0.884	**0.785**	27.30	**21.05**
w/o Low-pass	78.07	0.358	0.884	0.791	27.34	**21.05**
w/o Clipping	73.62	0.347	0.887	0.788	27.32	20.93
w/o Trust Region	54.57	**0.547**	**0.830**	0.938	14.68	17.97
w/o Renormalization	78.14	0.354	0.879	0.798	**27.37**	21.00

Table 5: **Inference cost.** Average generation time per video. SRO adds only modest overhead without any auxiliary model.

Method	Wan-1.3B	Wan-14B	CogX-5B
Standard	**74.2s**	**410.0s**	**389.8s**
FlowMo	243.0s	1318.0s	849.0s
MotionCFG	157.6s	481.2s	456.8s
SRO (K=25)	82.8s	441.3s	405.9s

provide a human evaluation, which confirms that these improvements are consistent with human preference; see Sec. A.6 for details (Table 7).

Image-to-Video Generation. We further evaluate SRO on I2V generation to examine its generalization beyond text-only conditioning. As shown in Table 3, SRO consistently improves motion dynamics over both the default I2V baseline and ALG (Choi et al., 2026) (i.e., the state-of-the-art) across all motion-related metrics. Despite the stronger motion, visual quality and subject consistency remain largely preserved, e.g., with minor changes in MUSIQ and aesthetic scores. These results show that SRO remains effective even under strong image conditioning, suggesting that its motion-enhancement capability is not limited to a specific conditioning modality.

4.3 Ablation and Efficiency Analysis

Component Ablation. Table 4 removes each component of Π from the full configuration. Without the trust region, generation collapses: CLIPScore drops from 27.30 to 14.68, and its high DEVIL reflects incoherent content rather than meaningful motion. Removing any other component preserves quality but weakens motion, e.g., RAFT Flow drops from 84.21 to 73.62 without clipping. Thus, the trust region keeps the optimized noise a valid initialization, while the remaining components further strengthen motion. In addition, we provide trade-off analysis in A.5; see Fig. 8 therein.

Computational Cost. Table 5 compares the inference cost of SRO with key baselines. FlowMo incurs by far the largest overhead, requiring multiple times the inference time of the standard sampler across all backbones. While MotionCFG is less costly, loading and executing the auxiliary LLM introduces non-negligible overhead, which grows noticeably with the LLM size. In contrast, SRO introduces only a modest overhead over the standard sampler without relying on any external model, making it comparable to or cheaper than MotionCFG across all settings.

5 Conclusion

We introduced Self-Reflecting Optimization (SRO), a training-free method that enhances motion dynamics in video diffusion models by optimizing the initial noise. SRO increases temporal self-reflection of the initial noise while constraining the optimization with low-frequency filtering, trust-region projection, and distribution preserving normalization. The optimized noise is then used with the original sampler without modifying model parameters or conditioning signals. Across text-to-video and image-to-video settings, SRO improves motion dynamics while largely preserving visual quality and semantic fidelity, suggesting that temporal self-reflection can serve as a unified signal for controlling video dynamics across various conditioning types.

Limitations and future work. Although effective, SRO primarily amplifies the overall magnitude of motion and does not provide explicit control over motion direction or object-specific dynamics. In addition, SRO relies on temporal attention, so extending it to future video architectures that do not employ such attention is not straightforward. Promising directions could be incorporating directional or region-level motion control into the optimization objective, and deriving motion signals that do not depend on a specific attention design.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code implementation and manuscript editing for readability and clarity. We did not use generative AI tools to formulate mathematical claims or proofs. All AI-assisted code and text were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including any text, code, or other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the algorithmic details of Self-Reflecting Optimization (SRO) in Sec. A.1. Full implementation details, including the self-reflecting score measurement, query strides, optimization hyperparameters, low-pass frequency ratio, trust-region radius, and noise normalization settings, are provided in Sec. A.2. The models, datasets, evaluation metrics, and generation protocols used in our experiments are described in Sec. 4. Examples of the LLM-generated motion-intensive prompt set are provided in Sec. A.3. These details are intended to enable reproduction of the proposed method and the reported experimental results.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

June Suk Choi, Kyungmin Lee, Sihyun Yu, Yisol Choi, Jinwoo Shin, and Kimin Lee. Improving motion in image-to-video models via adaptive low-pass guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40364–40375, 2026.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9380–9389. IEEE, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

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

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.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Byungjun Kim, Soobin Um, and Jong Chul YE. Diverse text-to-image generation via contrastive noise optimization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 29642–29669, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/327115e4cf461dce189938342e1d49cd-Paper-Conference.pdf.

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

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024.

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.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 638–647, 2022.

Fuseini Mumuni and Alhassan Mumuni. Segment anything model for automated image data annotation: empirical studies using text prompts from grounding dino. *arXiv preprint arXiv:2406.19057*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024.

Penghui Ruan, Pichao Wang, Divya Saxena, Jiannong Cao, and Yuhui Shi. Enhancing motion in text-to-video generation with decomposed encoding and conditioning. *Advances in Neural Information Processing Systems*, 37:70101–70129, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. Flowmo: Variance-based flow guidance for coherent motion in video generation. *Advances in Neural Information Processing Systems*, 38: 82725–82754, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416. IEEE, 2025.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European conference on computer vision*, pp. 378–394. Springer, 2024.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3024–3034. IEEE, 2025.

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, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

A APPENDIX

A.1 PSEUDOCODE

Algorithm 1 Self-Reflecting Optimization (SRO)

Require: Pretrained video diffusion model $\mathcal{G}_\theta$, condition c, target block ℓ^*, frequency ratio κ, trust-region radius ε, optimization steps K, learning rate η
Ensure: Generated video $\hat{x}_0$

```
Sample z_1 ~ N(0, I)
Initialize δ^(0) ← 0
for k = 0, ..., K − 1 do
    z̃_1^(k) ← Π(z_1, δ^(k))
    Forward z̃_1^(k) up to Transformer block ℓ*
    Extract self-attention maps A^(ℓ*,h)
    Compute self-reflecting score S^(ℓ*)(z̃_1^(k))
    L_SRO ← −S^(ℓ*)(z̃_1^(k))
    δ^(k+1) ← δ^(k) − η∇_δ L_SRO
end for
z_1^opt ← Π(z_1, δ^(K))
x̂_0 ← G_θ(z_1^opt, c)
return x̂_0
```

Algorithm 1 summarizes the overall optimization procedure of SRO. Starting from an initial Gaussian noise z_1, we iteratively optimize a perturbation δ to increase the self-reflecting score measured at the selected Transformer block ℓ^*. At each iteration, the constraint operator Π maps the perturbed noise to a feasible initialization before computing the self-reflecting objective. After K optimization steps, the resulting noise z_1^{opt} is directly used by the original video diffusion sampler. Importantly, no model parameters are updated during either optimization or generation.

A.2 IMPLEMENTATION DETAILS

Self-Reflecting measurement. We measure self-reflecting scores at the self-attention layer of Transformer block $\ell^* = 2$ to reduce memory consumption, the score is evaluated on a strided subset of query tokens while attending to all keys. We use query strides of 16, 48, and 32 for Wan2.1-1.3B, Wan2.1-14B and CogVideoX-5B respectively. For CogVideoX, which jointly attends over text and video tokens, we compute self-reference using only video keys and normalize by the total attention mass assigned to video tokens.

Optimization settings. We optimize the initial-noise perturbation with Adam for 25 iterations using a learning rate of 0.05. The low-pass frequency ratio is fixed to $\kappa = 0.25$, and noise values are clipped to ± 3 after normalization. We set $\varepsilon = 0.003$ for T2V experiments and $\varepsilon = 0.002$ for I2V experiments. Since the objective is evaluated at block ℓ^*, we terminate the forward pass immediately after this block during optimization, so gradients are propagated only through the preceding blocks.

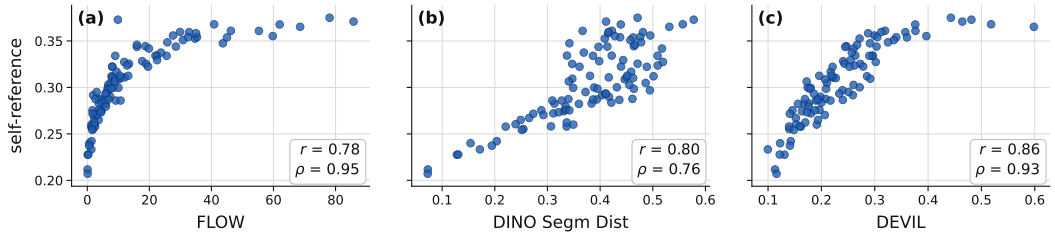
A.3 DETAILS OF GEMINI-GENERATED DYNAMIC PROMPTS

To further evaluate the ability of each method to generate large and diverse motion, we construct an additional motion-intensive prompt set using Gemini. The prompts cover a wide range of scenarios, including animals, humans, vehicles, sports, natural phenomena, objects, machines, fantasy scenes, urban environments, and food-related motion. Each prompt contains at least one explicit motion-related action

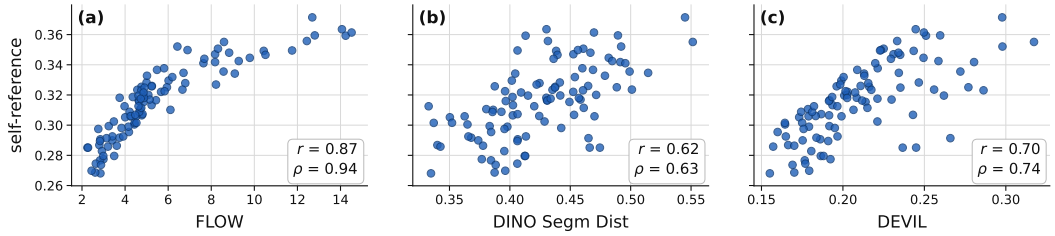
Table 6 lists the prompts used in this evaluation. We use the same prompt set and generation settings for all compared method, allowing us to directly compare their ability to translate motion-intensive text descriptions into dynamic video content.

Table 6: Representative examples from the Gemini-generated motion-intensive prompt set.

Category	Representative Prompts
Animal	A wild horse galloping through a dusty canyon. A cheetah sprinting across the African savanna.
Human	A martial artist delivering a fast spinning kick. A parkour athlete vaulting over a concrete wall.
Vehicle	A race car speeding around a sharp track corner. A rally car drifting around a muddy corner.
Sports	A surfer riding a huge crashing ocean wave. A gymnast tumbling across the floor mat.
Nature	A tornado swirling across an open flat plain. A river rushing over rapids in a deep gorge.
Object	A spinning top wobbling across a wooden table. A ball bouncing repeatedly down a stone staircase.
Machine	Turbine blades spinning fast inside a power plant. An excavator digging and dumping piles of dirt.
Fantasy	A dragon soaring and swooping over a burning castle. A giant robot stomping through a collapsing city.
City	Cars zooming past neon signs on a rainy night. A busy crosswalk with crowds surging in every direction.
Food	Popcorn kernels popping rapidly in a hot pan. Coffee pouring in a swirling stream into a white cup.



(a) Wan2.1-14B



(b) CogVideoX-5B

Figure 7: **Self-reflecting score of the initial noise correlates with motion across backbones.** Following Fig. 4, we generate 100 videos from a single prompt by varying only the seed, and plot the self-reflecting score at block 2 at the first denoising step against RAFT Flow, DINO-Segm, and DEVIL on (a) Wan2.1-14B and (b) CogVideoX-5B.

A.4 Additional Analysis of Temporal Self-Reference

In Fig. 4, we showed on Wan2.1-1.3B that the self-reflecting score is strongly correlated with the motion dynamics of the final video. To examine whether this relationship depends on model scale or architecture, we repeat the same seed-level analysis on Wan2.1-14B and CogVideoX-5B. As shown in Fig. 7, the self-reflecting score remains positively correlated with all three motion metrics on both

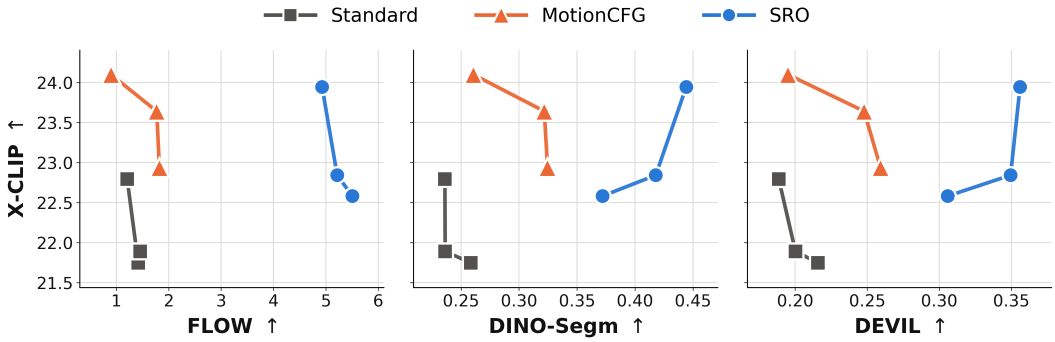


Figure 8: **Trade-off between text fidelity and motion dynamics.** Each curve sweeps the hyperparameter of a method: guidance scale for Standard, motion-token noise σ_c for MotionCFG (Kim et al., 2026b), and trust-region radius ε for SRO. Upper right is better.

Table 7: **User study results.** We conduct a user study comparing SRO against other methods, including the standard baseline, MotionCFG. Participants select the preferred video in terms of motion dynamics, visual quality, and text alignment.

Comparison	Method	Dynamics	Visual Quality	Text Alignment
vs. Standard	Standard	4.2%	15.8%	16.7%
	SRO (Ours)	**95.8%**	**84.2%**	**83.3%**
vs. MotionCFG	MotionCFG	27.2%	33.9%	23.3%
	SRO (Ours)	**72.8%**	**66.1%**	**76.7%**

backbones. These results suggest that the relationship between the self-reflecting score and motion consistently emerges across different model scales and architectures.

A.5 Trade-off Analysis

Fig. 8 visualizes the trade-off between text fidelity and motion dynamics by sweeping the key hyperparameter of each method. More specifically, we swept the guidance scale for Standard, the motion-token noise σ_c for MotionCFG (Kim et al., 2026b), and the trust-region radius ε for SRO. We observe that the standard baseline offers little control over motion dynamics, and MotionCFG increases dynamics only at the cost of text fidelity. In contrast, SRO consistently occupies the upper-right region, achieving higher motion dynamics while maintaining text fidelity, i.e., a more favorable trade-off than the baselines.

A.6 User Study

We conducted a user study to compare SRO against the standard sampler and MotionCFG in terms of human preference. In each trial, 30 participants were shown a pair of videos, one from SRO and one from a baseline, presented in random order. They were then asked to choose the preferred one w.r.t. three criteria: motion dynamics, visual quality, and text alignment. As shown in Table 7, SRO was preferred in 81.7% of comparisons for motion dynamics, 70.9% for visual quality, and 77.8% for text alignment, indicating that human observers favor SRO not only for its stronger motion but also for its visual quality and text alignment.