TOWARDS SUBJECT CONSISTENCY OVER DYNAMIC SUBJECT SETS IN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We argue that as video generation extends to longer durations, subject consistency should be evaluated over *dynamic subject sets*. We therefore introduce **DynSC-Eval**, an evaluation framework that dynamically tracks eligible subjects throughout their visible lifespans and measures local continuity and global identity preservation using six complementary object-level metrics, with explicit detection of inconsistency events. To validate its effectiveness, we design synthetic experiments that actively inject inconsistency events, demonstrating both the sensitivity of DynSC-Eval and the limitations of existing metrics. Evaluations of diverse models on 5s, 15s, and 60s video generation further reveal substantial subject consistency differences that are obscured by conventional metrics. Beyond evaluation, we construct rewards from DynSC-Eval and apply DiffusionNFT post-training in an autonomous-driving testbed. On 5s generation, our approach reduces the six inconsistency metrics by an average of 13.82% for Wan-2.1-1.3B and 5.66% for SANA-2B, with improvements also observed on the I2V model ReSim. Qualitative comparisons further demonstrate the effectiveness of our method. We then extend generation to 10s and 30s through curriculum learning and show that consistency optimization remains effective while largely preserving other capabilities.



Figure 1: **Examples of subject inconsistency in long video generation.** The corresponding prompts are provided in Appendix B.7. Additional failure cases are presented in the supporting material.

1 INTRODUCTION

Video models are increasingly moving towards long-video generation. Many commercial models now natively support generation of high-quality videos more than 10 seconds (OpenAI, 2025; Wiedemer et al., 2025; Team, 2025; Seedance et al., 2026; MiniMax, 2026), while some streaming models can even achieve minute-level video generation (Chen et al., 2025a; Yang et al., 2025c; Yuan et al., 2026; Gao et al., 2026; Zhu et al., 2026). Over longer temporal horizons, subjects may enter, leave, and interact, allowing both the composition of the visible subject set and the relationships among its members to evolve. Even within a continuous single shot, maintaining temporal consistency across these changing subjects remains challenging. Fig. 1 demonstrates failures in long-video generation.

The frequency and severity of these failures can reveal differences in models' capability. An ideal video generator should maintain each subject's temporal consistency throughout its visible lifespan.

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.

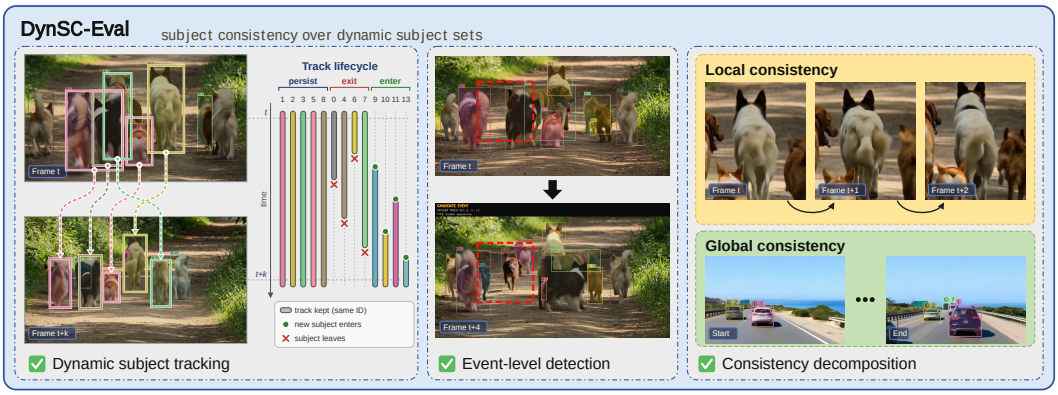

Figure 2: **Overview of DynSC-Eval.** The framework dynamically tracks subjects and their lifecycle, detects inconsistency events (e.g., sudden appearance/disappearance), and decomposes consistency into local and global consistency, enabling fine-grained evaluation of dynamic subject sets.

Thus, *evaluating subject consistency over dynamic subject sets* is a more appropriate assessment way. Here, *dynamic subject sets* means both the membership and cardinality of the visible subject set can change over time. Evaluation of subject consistency should cover the union of dynamic subject sets throughout the video. However, existing consistency metrics primarily rely on holistic frame-level comparisons (Huang et al., 2023; Liu et al., 2023b; Huang et al., 2024; Liu et al., 2026), sparsely sampled frames (Sun et al., 2024), or reference-based assessment of selected subjects (Zhou et al., 2026; Zhang et al., 2026; Ying et al., 2026; Li et al., 2026). These approaches either lack sensitivity in subject detection or overlook inconsistencies involving subjects that enter later in the video.

To address this gap, we introduce **DynSC-Eval**, an evaluation framework that dynamically updates the subject set to track each eligible subject throughout its visible lifespan. It decomposes temporal consistency into local and global components, measured using six complementary metrics. Local consistency captures short-term temporal continuity and incorporates explicit detection and counting of inconsistency events, while global consistency measures long-term identity preservation.

We design a controlled synthetic experiment that intentionally injects local and global inconsistencies. In this setting, evaluating only subjects visible in the initial frames cannot fully capture subject consistency throughout the video. Meanwhile, compared with holistic frame-level evaluation like VBench, our metrics show substantially greater sensitivity and clearly reflect changes in the prevalence of injected inconsistencies. We further validate DynSC-Eval across on a diverse set of commercial and open-source models at 5s, commercial models at 15s, and streaming models at 60s. DynSC-Eval distinguishes models with similar VBench consistency scores and reveals complementary failure patterns in local continuity and global identity preservation, demonstrating its effectiveness.

Furthermore, we design rewards based on DynSC-Eval to guide model optimization and improve subject consistency. We adopt autonomous driving as a representative testbed, where high-speed motion with multiple objects tend to expose temporal inconsistencies, which is practically critical for this domain. We first fine-tune text-to-video (T2V) models Wan-2.1-1.3B (Wang et al., 2025) and SANA-2B (Chen et al., 2025b) on 100 hours of real-world videos to establish 5s baselines. We then design consistency rewards based on DynSC-Eval metrics and perform post-training with DiffusionNFT (Zheng et al., 2025). On an independent held-out evaluation set of 100 videos, our approach yields average relative reductions of **13.82%** and **5.66%** across all six inconsistency metrics for Wan-2.1-1.3B and SANA-2B, respectively, while other capabilities, such as visual quality and motion dynamics, are preserved. Qualitative results also align with these numerical findings. We further apply the same optimization framework to the image-to-video (I2V) model ReSim (Yang et al., 2025b) and improve all consistency metrics of DynSC-Eval. These results demonstrate the effectiveness of our consistency rewards across both T2V and I2V generation.

Building on the improvements achieved for 5s generation, we extend our study to longer temporal horizons. Through curriculum learning, we progressively adapt the model to 10s and then 30s generation using an additional 100 and 200 hours of OpenDV data, respectively. We then apply

Table 1: Summary of DynSC-Eval metrics.

Dimension	Metric	Measured Property
Local	Spike Rate (SR)	Fraction of subjects with inconsistency events
Local	Events/100 Track-s (ER)	Frequency of inconsistency events
Local	Mean Adjacent Drift (MAD)	Track-averaged mean adjacent appearance distance
Global	Endpoint Drift (ED)	Start-to-end identity shift
Global	Mean Reference Drift (MRD)	Average long-term identity drift
Global	Worst-frame Drift (WFD)	Maximum identity deviation

DiffusionNFT with our consistency reward to both models, improving DynSC at both durations while largely preserving other capabilities. Our contributions are summarized as follows:

(i) We highlight the importance that subject consistency evaluation should account for dynamic subject sets as video generation extends to longer durations.

(ii) We introduce **DynSC-Eval**, an evaluation framework that combines dynamic subject tracking, event-level detection, and consistency decomposition. We validate its sensitivity through synthetic experiments and demonstrate its effectiveness across diverse models and video durations.

(iii) We design rewards based on DynSC-Eval and adopt autonomous driving as a testbed to post-train 5s models, achieving substantial subject consistency gains across both T2V and I2V models while largely preserving other capabilities, with qualitative results further supporting these improvements.

(iv) We then extend video generation from 5s to 30s via curriculum learning, demonstrating that our reward optimization also yields significant consistency gains in long-video generation.

2 Related Works

In this work, we focus on two primary aspects: evaluating subject consistency over dynamic subject sets and improving it through post-training. Due to space constraints, we provide a comprehensive review of temporal consistency evaluation and post-training for temporal consistency in Appendix A.

3 DynSC-Eval

In this section, we present our evaluation framework. Section 3.1 describes the evaluation pipeline and defines six metrics covering local and global consistency. Section 3.2 presents controlled synthetic experiments with deliberately injected inconsistency events, validating the effectiveness of our metrics. Section 3.3 evaluates the reliability and discriminative power of the metrics across diverse video models, demonstrating advantages over conventional metrics such as VBench (Huang et al., 2023).

3.1 Measurement

3.1.1 Trajectory Extraction

We evaluate temporal consistency over *dynamic subject sets*, accounting for all valid subjects throughout the video and construct an explicit trajectory for each target subject. Formally, given a video $\mathbb{V} = \{I_t\}_{t=1}^{T}$, we detect target subjects and use SAM2 (Ravi et al., 2024) to segment and track their visible regions. Newly observed objects are assigned new trajectories. After validity filtering, we obtain a trajectory set $\mathbb{T} = \{\mathcal{T}_i\}_{i=1}^{N}$, where N is the total number of valid object trajectories across the video. Each trajectory is represented as $\mathcal{T}_i = \{(t, x_{i,t}) \mid t \in \mathcal{I}_i\}$, where $\mathcal{I}_i \subseteq \{1, \dots, T\}$ contains the frames with valid visible observations of object i, and $x_{i,t}$ denotes its segmented object crop. The valid subject set at frame t is therefore defined as:

$$\mathcal{S}_t = \{i \mid t \in \mathcal{I}_i\}, \tag{1}$$

which evolves as subjects enter, leave, or become occluded. Evaluation covers trajectories across $\bigcup_{t=1}^{T} \mathcal{S}_t$. Each trajectory serves as the basic unit for the subsequent consistency metrics.

For each observation $x_{i,t}$, we extract its visual representation using a pre-trained DINOv3 (Siméoni et al., 2025) encoder $\phi(\cdot)$ and the representation $\boldsymbol{f}_{i,t} = \phi(x_{i,t})$. All embeddings are ℓ_2 normalized.

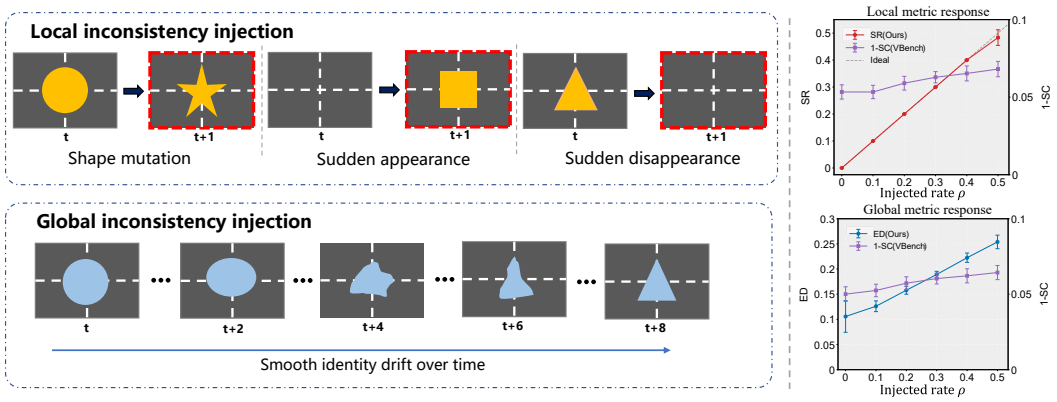

Figure 3: **Controlled synthetic experiments with actively injected inconsistencies.** We inject local inconsistencies (shape mutation, sudden appearance, sudden disappearance) and global inconsistencies (smooth identity drift) to evaluate metric sensitivity. As the corruption rate increases, our metrics (SR and ED) show a clear response, while VBench (1-SC) largely fails to capture these failures.

For two observations of the same subject at frames $t, s \in \mathcal{I}_i$, we define their appearance distance as

$$d(\boldsymbol{f}_{i,t}, \boldsymbol{f}_{i,s}) = 1 - \boldsymbol{f}_{i,t}^{\top} \boldsymbol{f}_{i,s}. \tag{2}$$

To improve evaluation robustness, we introduce two stabilization strategies: *temporal smoothing* and *trajectory boundary truncation*. Details are provided in Appendix B.2.

3.1.2 Local Consistency

Local consistency measures both **discrete inconsistency events** and **continuous short-term appearance variation**. Inconsistency events can be grouped into three categories: (i) *Sudden appearance*: a trajectory begins away from the temporal and spatial boundaries of the video, while the corresponding region in the preceding frame contains no compatible object; (ii) *Sudden disappearance*: a trajectory terminates within the valid region, and no compatible object remains in the corresponding region of the subsequent frame; (iii) *Severe appearance mutation*: an existing object undergoes an abrupt deformation or identity change. The first two event types are identified from detected trajectory onsets and terminations, subject to the above spatial and temporal boundary and compatibility constraints. For the third type, a transition is classified as a severe appearance mutation if the appearance distance is larger than τ, where $\tau \in (0, 1)$ is a conservative threshold for capturing large and visually salient changes. In the following, we describe the three local consistency metrics and formal definitions are provided in the Appendix B.3.1

Spike Rate (SR): The fraction of valid trajectories containing at least one inconsistency event.

Events per 100 Track-seconds (ER): The number of inconsistency events per 100 seconds of valid trajectory duration, capturing repeated failures.

Mean Adjacent Drift (MAD): The appearance distance between consecutive temporally smoothed observations, averaged within each valid trajectory and then across trajectories, capturing short-term variation even below the event threshold.

Together, these metrics characterize local consistency through failure prevalence, event frequency, and short-term appearance drift.

3.1.3 Global Consistency

Global consistency evaluates whether an object preserves its initially established identity throughout its trajectory. Since global consistency requires a sufficiently long temporal span to establish a reliable reference appearance, we only compute global consistency metrics for trajectories containing at least 16 frames. Trajectories shorter than 16 frames are evaluated only using the local consistency metrics. For each trajectory, we construct a robust identity reference by averaging and normalizing the

Table 2: **Subject consistency evaluation across diverse models.** DynSC-Eval metrics are reported to four significant figures (lower is better). VBench scores are reported on a 0–100 scale (higher is better): SC denotes Subject Consistency, Dyn denotes Dynamic Degree, QS denotes the Quality Score aggregated over seven quality dimensions. Best results within each group are in bold.

	DynSC-Eval Metrics						*VBench Metrics*		
	Local Consistency			*Global Consistency*					
Model	**SR↓**	**ER↓**	**MAD↓**	**ED↓**	**MRD↓**	**WFD↓**	**SC↑**	**Dyn↑**	**QS↑**
Open-source Models (5s)									
LTX-Video-2B (HaCohen et al., 2024)	0.3628	28.39	0.2629	0.3665	0.3475	0.5474	84.58	96.00	73.82
Wan-2.1-1.3B (Wang et al., 2025)	0.2530	10.96	0.1757	0.3384	0.2717	0.5041	88.09	96.00	77.09
SANA-2B (Chen et al., 2025b)	**0.1932**	**7.214**	**0.1224**	**0.3035**	**0.2081**	**0.4108**	**94.10**	99.00	**82.75**
Wan-2.1-14B (Wang et al., 2025)	0.1959	8.342	0.1510	0.3065	0.2409	0.4564	92.63	94.00	79.44
HunyuanVideo-13B (Kong et al., 2024)	0.2583	11.95	0.1582	0.3320	0.2550	0.4713	91.14	92.00	82.27
Commercial Models (5s)									
MiniMax H3 (MiniMax, 2026)	0.1378	**4.427**	**0.0960**	0.2665	0.1929	0.4042	**94.65**	93.00	**84.54**
Seedance 2.0 (Seedance et al., 2026)	**0.1320**	4.865	0.1012	0.2517	**0.1841**	**0.4009**	93.92	92.00	83.56
Kling 3.0 (Kling AI, 2026)	0.1414	4.954	0.1082	**0.2492**	0.1854	0.4010	94.52	99.00	83.13
Commercial Models (15s)									
MiniMax H3 (MiniMax, 2026)	0.1906	4.707	0.1069	0.3343	0.2328	0.4875	92.73	92.00	**83.74**
Seedance 2.0 (Seedance et al., 2026)	**0.1634**	4.257	0.1066	0.3308	0.2296	0.4754	92.07	97.00	83.70
Kling 3.0 (Kling AI, 2026)	0.1694	**3.677**	**0.1011**	**0.2997**	**0.2242**	**0.4662**	**93.56**	93.00	83.00
Streaming Models (60s)									
LongLive (Yang et al., 2025c)	0.4647	**8.356**	**0.1236**	0.4497	0.3092	0.5988	**97.98**	84.00	**82.73**
Helios (Yuan et al., 2026)	**0.2870**	9.066	0.1414	**0.3704**	**0.2748**	**0.5084**	94.33	96.00	77.62
LingBot-World 2.0 (Gao et al., 2026)	0.3350	10.97	0.1424	0.4761	0.3172	0.6403	94.16	**99.00**	81.36

embeddings of its first q valid observations. An endpoint embedding is similarly obtained. We next describe the three global consistency metrics, with formal definitions provided in the Appendix B.3.2

Endpoint Drift (ED): The appearance difference between the start and end of a trajectory. It computes the cosine distance between the reference and endpoint embeddings, averaged across globally eligible trajectories.

Mean Reference Drift (MRD): To capture identity variation throughout a trajectory, Mean Reference Drift compares each valid observation against the initial reference. These distances are averaged within each trajectory and then across globally eligible trajectories.

Worst-frame Drift (WFD): The most severe identity deviation along each trajectory. It takes the maximum distance from the initial reference over valid observations, then averages these maxima across globally eligible trajectories.

The three global metrics respectively characterize final-state identity preservation, average long-term drift, and worst-case deviation. Lower values indicate stronger temporal consistency.

3.2 Controlled synthetic experiments

To assess the discriminative ability of consistency metrics, we construct a controlled synthetic video experiment with explicitly injected local and global inconsistencies. Each video lasts 30s at a resolution of 832×480 and a frame rate of 16 FPS, containing 20 geometric objects that sequentially enter the scene. Their initial shapes are sampled from circles, triangles, squares, and pentagons. Each object moves smoothly along one of two predefined disjoint paths and remains in the scene for 10s.

Fig. 3(left) illustrates our inconsistency injection scheme. Local inconsistency simulates sudden appearance or disappearance and a brief appearance mutation: the object's contour becomes a five-pointed star for 8 frames at midlife before returning to its original shape. Global inconsistency simulates persistent shape drift: each object smoothly transitions to a target shape and retains it until leaving the scene. Each object undergoes one transition according to the fixed mapping circle $\rightarrow$ triangle, triangle $\rightarrow$ square, square $\rightarrow$ pentagon, and pentagon $\rightarrow$ circle. We interpolate radially between the source and target contours. The implementation details are provided in Appendix B.4. Under this setting, restricting evaluation to subjects visible in the initial frames can miss many injected

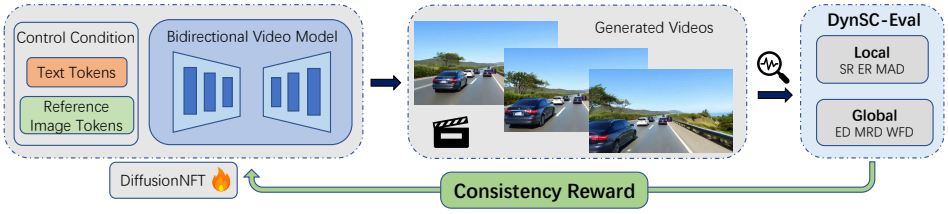

Figure 4: **Overview of consistency reward optimization framework.** Given text or reference-image conditioning, the video model generates videos that DynSC-Eval assesses through object-level local and global consistency metrics. These metrics are combined into a consistency reward for DiffusionNFT optimization, forming a feedback loop to improve temporal consistency. The framework applies to both T2V and I2V generation across short and long durations.

inconsistencies. Initial frames may contain no eligible subjects, or the initially visible subjects may remain consistent while inconsistencies occur among later arrivals, demonstrating its limitation. We next assess the sensitivity of DynSC-Eval and compare it with holistic frame-level assessment.

Let ρ denote the proportion of objects with injected inconsistencies. Starting from clean videos ($\rho = 0$), we progressively increase ρ and examine the resulting metric responses across three random seeds. For evaluation, we use Grounding DINO (Liu et al., 2023a) and SAM2 for automatic object detection and tracking, and DINOv3 features for consistency measurement. Fig. 3(right) shows the representative results when local and global inconsistencies are jointly injected. SR closely follows the ideal relation $\mathrm{SR} = \rho$, with $R^2 = 0.995$, while ED increases clearly with ρ. In contrast, VBench Subject Consistency (SC) remains nearly unchanged. These results demonstrate the sensitivity of DynSC-Eval to inconsistent subjects. We further isolate the two perturbation types and observe that injecting only local inconsistencies produces little change in ED and MRD, while injecting only global inconsistencies has limited effects on local metrics, demonstrating the complementary roles of our metrics. Results for all six metrics and the isolated ablations are provided in Appendix B.5.

3.3 Evaluation

We construct 100 prompts covering animals, pedestrians, football matches, and autonomous driving, with 25 prompts per scenario. Prompt construction details are provided in Appendix B.6. We set the severe appearance mutation detection threshold to $\tau = 0.35$ for general evaluation. We evaluate five open-source models at 5s, three commercial models at both 5s and 15s, and three streaming models at 60s. Table 2 shows the detailed evaluation results.

Longer videos pose greater consistency challenges. For the same commercial models, increasing duration from 5s to 15s raises the mean ED, MRD, and WFD by 25.7%, 22.1%, and 18.5%, respectively, compared with only 3.0% for MAD. SR also increases for all three models, indicating that more trajectories experience severe inconsistencies over longer horizons. The 60s streaming models exhibit substantial inconsistency, with SR ranging from 0.2870 to 0.4647, highlighting the remaining challenges in maintaining subject consistency in long videos.

DynSC-Eval reveals differences obscured by conventional metrics. At 5s, SANA-2B achieves a VBench SC of 94.10, comparable to commercial models (93.92–94.65), despite exhibiting noticeably poorer subject consistency under visual inspection. In contrast, DynSC-Eval clearly separates SANA-2B from the commercial models, with SANA-2B performing worse on all six metrics. A similar discrepancy becomes even more evident for long-video generation. The 60s streaming models achieve VBench SC scores comparable to commercial models at 15s, which is clearly inconsistent with visual perception, while DynSC-Eval reveals the differences. These results demonstrate that DynSC-Eval better reflects the inconsistency patterns and provides stronger discrimination across models.

DynSC-Eval distinguishes complementary consistency capabilities. Seedance 2.0 achieves the lowest SR among commercial models at both durations, whereas Kling 3.0 leads on the remaining five DynSC-Eval metrics at 15s. Among streaming models, LongLive achieves the lowest ER (8.356) and MAD (0.1236), while Helios performs best on SR (0.2870) and all three global metrics. These

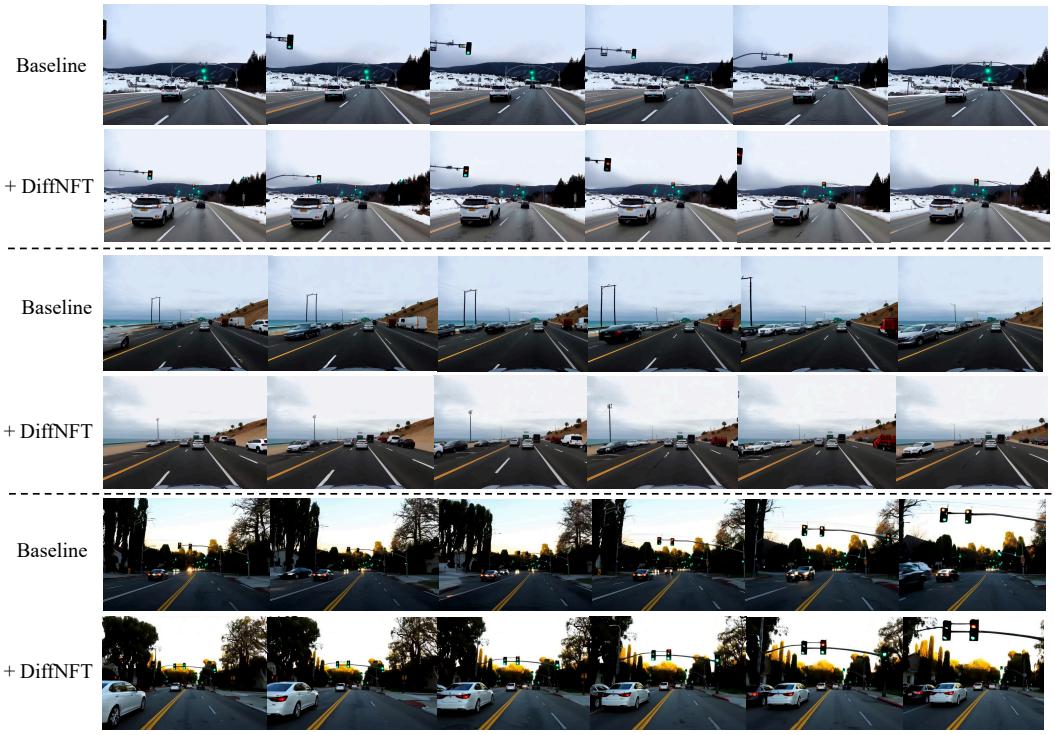

Figure 5: **Qualitative comparison on Wan-2.1-1.3B.** We compare the baseline with our method after DiffusionNFT. Top: our method eliminates the drift of a distant car in later frames. Middle: the parked car on the left maintains a more stable shape. Bottom: the interpenetration and deformation artifact in the fifth snapshot is removed. Overall, our method produces visually more stable and temporally consistent videos. Additional qualitative results are provided in the supporting material.

contrasting profiles demonstrate the value of separately measuring the prevalence of affected subjects, event frequency, short-term variation, and long-term drift.

4 From Consistency Evaluation to Model Optimization

4.1 Testbed

We next design consistency rewards that directly guide model optimization. We adopt autonomous driving as a representative testbed because driving videos contain multiple persistent vehicles undergoing continuous motion and frequent interactions, which provides abundant object trajectories for exposing both abrupt local failures and accumulated global drift. We establish two baselines based on Wan-2.1-1.3B (Wang et al., 2025) and SANA-2B (Chen et al., 2025b). We sample real-world videos from OpenDV and divide them into 5s clips, obtaining approximately 100 hours of data. Each clip is automatically captioned using Qwen3-VL-8B (Bai et al., 2025) to construct text–video pairs. Both models are fine-tuned on the same dataset to obtain the baselines. Refer to Appendix C.1 for details.

4.2 Method

After establishing the baseline model, we design online rewards from DynSC-Eval metrics and optimize consistency with DiffusionNFT (Zheng et al., 2025) to further improve subject consistency of these models. Fig. 4 summarizes our consistency optimization framework.

For online optimization, we select four metrics: beta-smoothed SR, MAD, WFD, and MRD, covering both local and global consistency. The consistency reward is defined as

$$R_{\text{cons}} = -\left(0.30C_{\text{spike}} + 0.30C_{\text{adj}} + 0.30C_{\text{worst}} + 0.10C_{\text{ref}}\right), \tag{3}$$

Table 3: **Results of T2V and I2V models.** +DiffNFT denotes the results after 75 optimization steps. VBench scores use a 0–100 scale. Better DynSC-Eval results are highlighted in bold.

		DynSC-Eval Metrics						*VBench Metrics*			
		Local Consistency			*Global Consistency*						
Model	**Setting**	**SR↓**	**ER↓**	**MAD↓**	**ED↓**	**MRD↓**	**WFD↓**	**Mot↑**	**Dyn↑**	**Aes↑**	**Img↑**
Text-to-Video (T2V)											
Wan-2.1-1.3B (Wang et al., 2025)	Baseline	0.2560	9.177	0.1059	0.3837	0.2393	0.4428	98.35	92.00	50.47	71.40
	+DiffNFT	**0.2220**	**7.139**	**0.0883**	**0.3446**	**0.2129**	**0.4002**	98.49	92.00	50.81	71.82
SANA-2B (Chen et al., 2025b)	Baseline	0.2835	10.50	0.1313	0.3979	0.2733	0.4734	98.84	91.00	53.83	71.92
	+DiffNFT	**0.2691**	**9.868**	**0.1225**	**0.3737**	**0.2570**	**0.4536**	98.94	89.00	53.97	72.19
Image-to-Video (I2V)											
ReSim (Yang et al., 2025b)	Baseline	0.2902	12.87	0.1170	0.3170	0.2075	0.3981	98.33	91.00	50.45	71.70
	+DiffNFT	**0.2758**	**11.55**	**0.1116**	**0.3068**	**0.2013**	**0.3853**	98.44	91.00	50.44	71.26

where each C denotes the normalized inconsistency cost of the corresponding DynSC-Eval metric. A trivial solution would be to generate videos without any target objects. To discourage this behavior, rollouts containing no qualified object trajectories receive a reward of -10.

We freeze the base diffusion transformer and optimize LoRA adapters using DiffusionNFT. An EMA-updated old policy generates grouped rollouts, whose rewards are standardized within each prompt group. A frozen reference policy provides regularization to improve consistency while preserving generative capabilities. Training details are provided in Appendix D.2.

4.3 Experiments

We use Qwen3-8B (Yang et al., 2025a) to construct 300 multi-vehicle prompts and independently train the 5s Wan-2.1-1.3B and SANA-2B baselines with DiffusionNFT for 75 steps. Given the stricter requirements for subject consistency in autonomous driving, we use a lower threshold of $\tau = 0.16$ to increase sensitivity to severe appearance mutation detection. As shown in Table 3, both models improve across all six DynSC-Eval metrics. For Wan-2.1-1.3B, post-training reduces SR, ER and MAD by **13.3%, 22.2%, 16.6%**. Global consistency also improves, with ED, MRD, and WFD decreasing by **10.2%, 11.0%** and **9.6%**. SANA-2B exhibits consistent improvements: its three local metrics decrease by **5.11%, 5.99%, 6.65%** while its three global metrics decrease by **6.08%, 5.96%** and **4.19%**. Meanwhile, VBench metrics indicate our method improves temporal consistency while preserving other capabilities. Qualitative results in Fig. 5 with additional examples in the supporting materials, corroborate the quantitative findings, demonstrating the effectiveness of our method.

4.4 Generalization to Image-to-Video Generation

We next investigate whether our optimization framework can extend to I2V generation. We adopt ReSim (Yang et al., 2025b), an open-source autonomous-driving model at 4.9s, and apply DiffusionNFT to train it with the same 300 prompts. As shown in Table 3, our method improves all six DynSC-Eval metrics, with average reductions of **6.61%** and **3.14%** for local and global inconsistency metrics, respectively. These results extend our findings beyond T2V generation and support the applicability of the proposed feedback rewards across different generation paradigms.

4.5 Scaling Consistency Optimization to Long-video Generation

Long-video generation presents a greater challenge to temporal consistency, as object states must remain coherent over longer durations and errors can accumulate over time. Building on the consistency improvements achieved for 5s videos, we next investigate how to improve temporal consistency as generation scales to longer horizons. We adopt a curriculum learning strategy to progressively extend the driving model from 5s to 10s and then 30s. Specifically, we fine-tune the 5s Wan-2.1-1.3B on an additional 100 hours of 10s OpenDV clips, followed by 200 hours of 30s clips, with no overlap among the training subsets. Details are provided in Appendix C.2.

We then apply the optimization framework introduced in Sec. 4.2 to the full-attention long-video models. Tab. 4 reports the results for 10s and 30s models after 100 training steps, evaluated on fixed

Table 4: **Results of the 10s and 30s models.** +DiffNFT denotes the results after 100 optimization steps. VBench scores use a 0–100 scale. Better DynSC-Eval results are highlighted in bold.

	DynSC-Eval Metrics						*VBench Metrics*			
	Local Consistency			*Global Consistency*						
Setting	**SR↓**	**ER↓**	**MAD↓**	**ED↓**	**MRD↓**	**WFD↓**	**Mot↑**	**Dyn↑**	**Aes↑**	**Img↑**
10s baseline	0.3843	11.77	0.1328	0.4927	0.3291	0.5759	98.03	98.00	50.49	72.12
+DiffNFT	**0.3350**	**9.172**	**0.1119**	**0.4754**	**0.3087**	**0.5547**	97.96	98.00	50.04	72.86
30s baseline	0.4684	13.27	0.1675	0.5588	0.3712	0.6435	96.48	98.00	48.45	67.92
+DiffNFT	**0.4473**	**12.00**	**0.1633**	**0.5088**	**0.3373**	**0.5973**	96.89	98.00	48.85	67.45

sets of 100 and 50 videos, respectively. Our optimization framework consistently improves both local and global consistency. Specifically, for the 10s model, ER, MAD, and SR decrease by **22.09%**, **15.72%**, and **12.81%**, while global MRD and WFD decrease by **6.20%** and **3.69%**. For the 30s model, the same procedure improves all six metrics: SR, ER, and MAD decrease by **4.50%**, **9.54%**, and **2.47%** while global ED, MRD and WFD decrease by **8.94%**, **9.12%** and **7.19%**. VBench scores remain broadly comparable to the baseline, indicating that visual quality and motion dynamics are largely preserved. Additional qualitative comparisons are provided in the supporting materials, further demonstrating the effectiveness of our optimization framework at longer generation horizons.

5 Ablation Study

We examine the contributions of local and global consistency rewards on Wan-2.1-1.3B for 5s generation. The Local-only reward uses the SR and MAD components, while the Global-only reward uses MRD and WFD. As shown in Tab. 5, Local-only optimization achieves the lowest MAD and all three global drift metrics, but reduces Dyn from 92.00 to 85.00, alongside decreases in Aes and Img. This indicates that Local-only optimization tends to converge to a trivial solution of slowing down motion. In contrast, Global-only optimization preserves model capabilities but its local metric SR even worsens. The combined reward achieves the lowest SR and ER, at 0.2220 and 7.139, respectively, while maintaining other capabilities comparable to or above the baseline. These results support the necessity and effectiveness of combining local and global rewards.

Table 5: **Ablation of consistency reward components.** All optimized models are evaluated at Step 75 using the same configuration. VBench scores use a 0–100 scale.

	DynSC-Eval Metrics						*VBench Metrics*			
	Local Consistency			*Global Consistency*						
Reward	SR↓	ER↓	MAD↓	ED↓	MRD↓	WFD↓	Mot↑	Dyn↑	Aes↑	Img↑
5s baseline	0.2560	9.177	0.1059	0.3837	0.2393	0.4428	98.35	92.00	50.47	71.40
Local-only	0.2443	7.549	0.0865	0.3434	0.2110	0.3929	98.61	85.00	49.81	70.84
Global-only	0.2691	9.155	0.0970	0.3583	0.2262	0.4205	98.23	91.00	50.59	73.12
Ours (local+global)	0.2220	7.139	0.0883	0.3446	0.2129	0.4002	98.49	92.00	50.81	71.82

6 Conclusion and Future works

We highlight the importance of evaluating subject consistency over dynamic subject sets in long video generation. Accordingly, we introduce **DynSC-Eval**, which tracks eligible subjects and measures local and global consistency, including inconsistency events detection. Synthetic experiments and evaluations across diverse models and durations validate its sensitivity and discriminative power, exposing limitations of existing methods. Using autonomous driving as a testbed, we show that rewards derived from DynSC-Eval improve subject consistency across T2V and I2V models, with gains extending to 10s and 30s generation while preserving other capabilities.

Future work will extend this evaluation to multi-view generation, jointly assessing temporal continuity and consistency across viewpoints. We also aim to incorporate rewards for broader physical constraints into our optimization framework to improve the plausibility of generated dynamics and interactions.

AI USE STATEMENT

We used generative AI tools for writing refinement, literature retrieval and discovery, and research implementation. For writing, AI was used to polish author-written text for clarity, grammar, and readability, rather than to draft sections of the paper. For literature discovery, we first conducted a manual review, then used AI to identify additional relevant work, and manually verified the references and associated descriptions against the original sources. For research implementation, AI assisted with evaluation scripts and post-training code based on evaluation rules and optimization methods designed by the authors. We reviewed the implementations and checked their behavior against the intended designs.

We did not use generative AI to draft sections of the paper, prove mathematical claims, or generate the controlled synthetic dataset. The synthetic videos were generated programmatically using human-designed rules for object motion and inconsistency injection.

We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, code, and experimental artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main text describes the DynSC-Eval framework, consistency metrics, and reward design. The appendix provides additional implementation details, training configurations, hyperparameter comparisons, and prompts used in the experiments. Upon acceptance, we will release our evaluation and training code, model checkpoints, and evaluation data to facilitate reproduction of our results. We will not redistribute the autonomous driving training data due to copyright restrictions.

REFERENCES

Takehiro Aoshima, Yusuke Shinohara, and Byeongseon Park. Video consistency distance: Enhancing temporal consistency for image-to-video generation via reward-based fine-tuning. *ArXiv*, abs/2510.19193, 2025. URL https://api.semanticscholar.org/CorpusID:282272651.

Shuai Bai, Yuxuan Cai, Rui-Zhe Chen, Ke qin Chen, Xiong-Hui Chen, Zesen Cheng, Liang-Hao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Qidong Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Ming-Sheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xue-Jing Liu, Jiawei Liu, Cheng-Long Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chen-Xu Lv, Rui Men, Lingchen Meng, Lingchen Meng, Xin yi Ren, Sibo Song, Yu chen Sun, Jun Tang, Jianhong Tu, Jian-Qiang Wan, Peng Wang, Pengfei Wang, Qiu-Yue Wang, Yuxuan Wang, Tian-Bao Xie, Yihe Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bo-Wen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Botao Zheng, Humen Zhong, Jingren Zhou, Fanxi Zhou, Jingren Zhou, Yuanzhi Zhu, and Keming Zhu. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025. URL https://api.semanticscholar.org/CorpusID:283262018.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zhenghao Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhi-Xin Xu, Yuzhe Jin, Yu Liang, Yu-Bo Song, Peng Zhao, Bo Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model. *ArXiv*, abs/2504.13074, 2025a. URL https://api.semanticscholar.org/CorpusID:277856899.

Jun-Song Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Jun-Yu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, Haozhe Liu, Hongwei Yi, Hao Zhang, Muyang Li, Yukang Chen, Han Cai, Sanja Fidler, Ping Luo, Song Han, and Enze Xie. Sana-video: Efficient video generation with block linear diffusion transformer. *ArXiv*, abs/2509.24695, 2025b. URL https://api.semanticscholar.org/CorpusID:281674126.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and re-

ward feedback learning. 2023. URL https://api.semanticscholar.org/CorpusID:271855272.

Ze-Lin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jianjun Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Ouyang Hao. Infinite worlds with versatile interactions. *ArXiv*, abs/2607.07534, 2026. URL https://api.semanticscholar.org/CorpusID:289978604.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, David-Pur Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *ArXiv*, abs/2501.00103, 2024. URL https://api.semanticscholar.org/CorpusID:275212083.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li jia Li, and Yongxin Ni. Video-bench: Human-aligned video generation benchmark. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18858–18868, 2025. URL https://api.semanticscholar.org/CorpusID:277622044.

Xuan He, Dongfu Jiang, Ge Zhang, Max W.F. Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi "Richard" Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://api.semanticscholar.org/CorpusID:270688037.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qin Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2023. URL https://api.semanticscholar.org/CorpusID:265506207.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *ArXiv*, abs/2411.13503, 2024. URL https://api.semanticscholar.org/CorpusID:285614965.

Karthik Inbasekar, Guy Rom, and Omer Shlomovits. Worldjen: An end-to-end multi-dimensional benchmark for generative video models. *ArXiv*, abs/2605.03475, 2026. URL https://api.semanticscholar.org/CorpusID:287963320.

Kling AI. Kling VIDEO 3.0 Model User Guide. https://kling.ai/quickstart/klingai-video-3-model-user-guide, February 2026. Accessed: 2026-09-10.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jia-Liang Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fan Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyuan Li, Shuai Li, Wei-Yan Wang, Wenqing Yu, Xi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhen Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yang-Dan Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *ArXiv*, abs/2412.03603, 2024. URL https://api.semanticscholar.org/CorpusID:274514554.

Da-Cheng Li, Yun-Hao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. *ArXiv*, abs/2502.20694, 2025. URL https://api.semanticscholar.org/CorpusID:276725482.

Haoran Li, Fredreic Li, Shichen Ma, Jie Huang, Yijun Liu, Jiaqi Shi, and Yanwen Ma. Joyai-echo: Pushing the frontier of long audio-visual generation. 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chun yue Li, Jian-Wei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023a. URL https://api.semanticscholar.org/CorpusID:257427307.

Tengfei Liu, Yang Shi, Xuanyu Zhu, Jiafu Tang, Liu Yang, Qixun Wang, Zhuoran Zhang, Yuqi Tang, Fengxiang Wang, Yuhao Dong, Xinlong Chen, Bozhou Li, Bohan Zeng, Yuetong Ding, Xiaohan Zhang, Jialu Chen, Haotian Wang, Yuanxing Zhang, Pengfei Wan, and Leye Wang. Longav-compass: Towards unified evaluation of minute-scale audio-visual generation across t2av, i2av, and v2av. *ArXiv*, abs/2605.26244, 2026. URL https://api.semanticscholar.org/CorpusID:288672545.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond H. Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, 2023b. URL https://api.semanticscholar.org/CorpusID:264172222.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. MiniMax Technical Blog, July 2026. URL https://www.minimax.io/blog/minimax-h3. Accessed: 2026-09-02.

OpenAI. Sora 2 System Card. OpenAI Technical Report, September 2025. URL https://openai.com/index/sora-2-system-card/. Accessed: 2026-09-02.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya K. Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao Wu, Ross B. Girshick, Piotr Doll'ar, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *ArXiv*, abs/2408.00714, 2024. URL https://api.semanticscholar.org/CorpusID:271601113.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

Team Seedance, Deying Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xu Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujia Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Ze Fang, Cheng Feng, Yu Gao, Diandian Gu, Dong Guo, Han Guo, Qiushan Guo, Bo Hao, Hongxiang Hao, Hao He, Jiaao He, Qian He, Tu-Tai Hoang, Hengtong Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhao Huang, Zhongyi Huang, Jishuo Jin, Mingbo Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bing Li, Geng Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, Jerry Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Ji-Hong Liu, Jingtuo Liu, Shu Liu, Si-Yu Liu, Wei Liu, Xue Liu, Zuxin Liu, Ruijie Lu, Lecheng Lyu, Jin Ma, Tianxiang Ma, Xi Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Rong Peng, Xue-Qian Qu, Yuxin Ren, Yucheng Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zi-Meng Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weicheng Wang, Xuan Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Gu-Yeon Wei, Meng Wei, Di Wu, Guohong Wu, Han-Cheng Wu, Hua Wu, Jian Wu, Jie Wu, Ruolan Wu, Shao-Yi Wu, Xiaohui Wu, Xinglong Wu, Yonghui Wu, Rui Xia, Xin Xia, Xu Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yi-Hsuan Yang, Zhixiang Yang, Ziyan Yang, Fulong Ye, B. Yi, Xingxi Yin, Yongbin You, Lin Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chen-Lin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyu Zhang, Shuo Zhang, Xiao-Hui Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Hu zi Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuanyan Zhu, and Feilong

Zuo. Seedance 2.0: Advancing video generation for world complexity. *ArXiv*, abs/2604.14148, 2026. URL https://api.semanticscholar.org/CorpusID:287480844.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://api.semanticscholar.org/CorpusID:267412607.

Liao Shen, Wentao Jiang, Yiran Zhu, Jia-He Li, Tiezheng Ge, Zhiguo Cao, and Bo Zheng. Identity-preserving image-to-video generation via reward-guided optimization. *ArXiv*, abs/2510.14255, 2025. URL https://api.semanticscholar.org/CorpusID:282139566.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. 2025. URL https://api.semanticscholar.org/CorpusID:280649654.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihang Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416, 2024. URL https://api.semanticscholar.org/CorpusID:271310039.

Kling Team. Kling-omni technical report. *ArXiv*, abs/2512.16776, 2025. URL https://api.semanticscholar.org/CorpusID:283934036.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningying Zhang, Pandeng Li, Ping Wu, Ruihang Chu, Rui Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wen-Chao Zhou, Wente Wang, Wen Shen, Wenyuan Yu, Xianzhong Shi, Xiaomin Huang, Xin Xu, Yan Kou, Yan-Mei Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhigang Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *ArXiv*, abs/2503.20314, 2025. URL https://api.semanticscholar.org/CorpusID:277321639.

Thaddaus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *ArXiv*, abs/2509.20328, 2025. URL https://api.semanticscholar.org/CorpusID:281505752.

An Yang, An-Feng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chang Gao, Chengen Huang, Chen-Xu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jian-Hong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Ke-Qin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Ming-Ze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tian-Yi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Ying-Er Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhen-Ru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. 2025a. URL https://api.semanticscholar.org/CorpusID:278602855.

Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiang-Yu Yue, and Li Chen. Resim: Reliable world simulation for autonomous driving. *ArXiv*, abs/2506.09981, 2025b. URL https://api.semanticscholar.org/CorpusID:279305514.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *ArXiv*, abs/2509.22622, 2025c. URL https://api.semanticscholar.org/CorpusID:281658815.

Xie-Qin Ye, Wenjia Yang, Yangyang Xu, Xiao-Yan Liu, Duo Su, Mengfei Xia, and Jun Zhu. Shift: Motion alignment in video diffusion models with adversarial hybrid fine-tuning. *ArXiv*, abs/2603.17426, 2026. URL https://api.semanticscholar.org/CorpusID:286643817.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Heng-Hui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *ArXiv*, abs/2605.25874, 2026. URL https://api.semanticscholar.org/CorpusID:288670455.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6463–6474, 2023. URL https://api.semanticscholar.org/CorpusID:266374742.

Shenghai Yuan, Yuan-Yang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *ArXiv*, abs/2603.04379, 2026. URL https://api.semanticscholar.org/CorpusID:286239319.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, Hao Liu, Chen Li, Jing Lyu, and Yueqi Duan. Mbench: A comprehensive benchmark on memory capability for video world models. *ArXiv*, abs/2606.00793, 2026. URL https://api.semanticscholar.org/CorpusID:288860951.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Mingying Liu. Diffusionnft: Online diffusion reinforcement with forward process. *ArXiv*, abs/2509.16117, 2025. URL https://api.semanticscholar.org/CorpusID:281411162.

Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven Lake Waslander. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. *ArXiv*, abs/2601.01528, 2026. URL https://api.semanticscholar.org/CorpusID:284487751.

Haoyi Zhu, Hao-Zhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jin-Cheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *ArXiv*, abs/2605.15178, 2026. URL https://api.semanticscholar.org/CorpusID:288254740.

A Related works

A.1 Temporal Consistency Evaluation

Temporal consistency is an important dimension of video generation evaluation. General benchmarks, including VBench (Huang et al., 2023), VBench++ (Huang et al., 2024), and EvalCrafter (Liu et al., 2023b), primarily assess temporal consistency through holistic frame-level comparisons. A complementary line uses human judgments and learned evaluators (He et al., 2024; Han et al., 2025; Li et al., 2025; Inbasekar et al., 2026). LongAV-Compass (Liu et al., 2026) extends evaluation to minute-level audio-visual generation using MLLM assessment and complementary perceptual metrics. These approaches provide broad perceptual and semantic coverage. However, such holistic evaluations struggle to localize inconsistencies in individual subjects and have limited sensitivity to discrete inconsistency events.

Subject-level evaluations address more specific aspects of consistency. T2V-CompBench (Sun et al., 2024) evaluates compositional generation via MLLM judges, subject detection, and tracking, targeting consistent attribute binding and object interactions. JoyAI-Echo (Li et al., 2026) emphasizes character identity consistency across shots in long-form audio-visual generation. Recent works, including DrivingGen (Zhou et al., 2026), MBench (Zhang et al., 2026), and WBench (Ying et al., 2026), evaluate subject consistency using designated targets or first-frame-based protocols, yet comprehensive coverage of the evolving subject set, including subjects entering later as videos grow longer, remains insufficiently addressed. In contrast, DynSC-Eval explicitly evaluates dynamic subject sets, including all eligible subjects that first appear after the initial frame. It separately measures local disruptions and global drift and detects inconsistency events, providing a more comprehensive and accurate assessment of subject consistency throughout the video.

A.2 Post-Training for Temporal Consistency

Policy optimization methods such as PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) provide foundations for learning from reward feedback. For diffusion models, DiffusionNFT (Zheng et al., 2025) incorporates reward feedback into forward-process flow-matching updates by contrasting positive and negative generations. In video generation, InstructVideo (Yuan et al., 2023) adapts image-based preference rewards through segmental sampling and temporal attenuation while Control-A-Video (Chen et al., 2023) incorporates reward feedback into controllable generation. Some works explicitly design consistency signals to improve temporal consistency in video generation. Video Consistency Distance (Aoshima et al., 2025) uses a frequency-domain discrepancy between generated frame features and a conditioning image for I2V fine-tuning. IPRO (Shen et al., 2025) targets facial identity preservation using identity-based rewards. SHIFT (Ye et al., 2026) introduces pixel-motion rewards for instantaneous and long-term motion consistency within a hybrid supervised and advantage-weighted fine-tuning framework. Our work adopts DiffusionNFT as the optimization framework and designs a dynamic subject consistency reward that combines local and global signals, substantially improving subject consistency in generated videos, with gains further validated on long videos.

B Implementation Details of DynSC

B.1 Local and Global Temporal Consistency

For a continuous single-shot video, temporal consistency requires more than visually smooth frame transitions. A generated subject should remain locally coherent from one moment to the next, while also preserving its identity and appearance over a much longer temporal horizon. Accordingly, we decompose temporal consistency into two dimensions: *local consistency* and *global consistency*.

Local temporal consistency. Local consistency measures whether an subject's state evolves continuously over short temporal intervals. For nearby observations along the same trajectory, changes caused by motion should occur smoothly rather than abruptly. Local inconsistency manifests as transient events, such as sudden appearance or disappearance and severe deformation, which should be explicitly detected and penalized. This dimension directly reflects the model's ability to maintain short-term temporal continuity.

Global temporal consistency. Global consistency measures whether the generative model preserves an subject's established state and identity throughout its complete trajectory. Despite substantial motion and viewpoint changes, later observations should remain compatible with the same underlying subject. This dimension reflects the model's capacity for long-term memory.

Conventional approaches often collapse these behaviors into a single aggregate score, obscuring their distinct failure mechanisms and providing ambiguous supervision for model optimization. In contrast, a more effective approach is to evaluate local discontinuities and global trajectory drift separately and use the resulting subject-trajectory signals as complementary feedback for optimization.

B.2 Robustness Techniques

Temporal smoothing. Recall that $\boldsymbol{f}_{i,k}$ denotes the ℓ_2-normalized DINOv3 representation of subject i at frame k. Frame-level visual representations may contain small variations. To obtain a more stable estimate of short-term appearance, we temporally average every two consecutive embeddings with a stride of two as $\bar{\boldsymbol{f}}_{i,k} = \mathrm{Norm}(\frac{\boldsymbol{f}_{i,2k-1}+\boldsymbol{f}_{i,2k}}{2})$, where $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization. Adjacent appearance variation is then measured between consecutive pooled representations:

$$d_{i,k}^{\mathrm{adj}} = 1 - \bar{\boldsymbol{f}}_{i,k}^{\top} \bar{\boldsymbol{f}}_{i,k+1}. \tag{4}$$

Trajectory Boundary Truncation. Track initialization and termination are intrinsically ambiguous because an subject may naturally enter or leave the camera view. Directly including these observations in feature-based computation may introduce noisy distances or false positive signals. Let s_i and e_i denote the first and last frame indices of trajectory $\mathcal{T}_i$. We use a temporal margin of $m = 4$ frames. If $s_i > m$, we exclude the first m observations of $\mathcal{T}_i$ from feature-based consistency computation. Similarly, if $e_i \leq T - m$, we exclude its last m observations. Although these boundary observations are omitted when computing appearance distances, their neighboring frames are retained for determining whether the trajectory constitutes an abnormal appearance or disappearance event.

B.3 Metric Computation

B.3.1 Local Consistency

Local consistency measures both **discrete inconsistency events** and **continuous short-term appearance variation**. We consider three types of inconsistency events: (i) *Sudden appearance*: a trajectory begins away from the temporal and spatial boundaries of the video, while the corresponding region in the preceding frame contains no compatible subject; (ii) *Sudden disappearance*: a trajectory terminates within the valid region, and no compatible subject remains in the corresponding region of the subsequent frame; (iii) *Severe appearance mutation*: an existing subject undergoes an abrupt deformation or identity change.

The first two event types are identified from detected trajectory onsets and terminations, subject to the above spatial and temporal boundary and compatibility constraints. For the third type, a transition

is classified as a severe appearance mutation if

$$d_{i,k}^{\mathrm{adj}} > \tau,$$

where $\tau \in (0, 1)$ is a conservative threshold for capturing large and visually salient changes.

The complete event set for trajectory i is

$$\mathcal{E}_i = \mathcal{E}_i^{\mathrm{appear}} \cup \mathcal{E}_i^{\mathrm{disappear}} \cup \mathcal{E}_i^{\mathrm{mutation}}.$$

Let N_{local} denote the number of trajectories eligible for local evaluation.

Spike Rate (SR). Spike Rate measures the fraction of trajectories containing at least one inconsistency event:

$$\mathrm{SR} = \frac{1}{N_{\mathrm{local}}} \sum_{i=1}^{N_{\mathrm{local}}} \mathbb{I}\left[|\mathcal{E}_i| > 0\right]. \tag{5}$$

Events per 100 Track-seconds (ER). Spike Rate does not distinguish between trajectories containing one failure and those containing repeated failures. We additionally report the number of inconsistency events per 100 seconds of valid subject trajectories computed as $100 \cdot \sum_i |\mathcal{E}_i| / \sum_i L_i$, where L_i denotes the valid duration of trajectory i in seconds.

Mean Adjacent Drift (MAD). Mean Adjacent Drift measures continuous short-term appearance variation, including changes below the event threshold:

$$\mathrm{MAD} = \frac{1}{N_{\mathrm{local}}} \sum_{i=1}^{N_{\mathrm{local}}} \left(\frac{1}{K_i - 1} \sum_{k=1}^{K_i - 1} d_{i,k}^{\mathrm{adj}} \right), \tag{6}$$

where K_i is the number of temporally smoothed observations in trajectory i.

B.3.2 Global Consistency

Global consistency evaluates whether an subject preserves its initially established identity throughout its trajectory. Since global consistency requires a sufficiently long temporal span to establish a reliable reference appearance, we only compute global consistency metrics for trajectories containing at least 16 frames. Trajectories shorter than 16 frames are evaluated only using the local consistency metrics.

Let $\mathcal{R}_i$ denote the first $q = 4$ valid observations of trajectory i, where we use a short multi-frame window rather than a single frame to obtain a robust identity reference. The reference embedding is

$$\boldsymbol{f}_i^{\mathrm{ref}} = \mathrm{Norm}(\frac{1}{q} \sum_{t \in \mathcal{R}_i} \boldsymbol{f}_{i,t}). \tag{7}$$

Similarly, let $\mathcal{Q}_i$ denote the final q valid observations, from which we construct

$$\boldsymbol{f}_i^{\mathrm{end}} = \mathrm{Norm}(\frac{1}{q} \sum_{t \in \mathcal{Q}_i} \boldsymbol{f}_{i,t}). \tag{8}$$

Endpoint Drift (ED). Endpoint Drift measures the appearance difference between the start and end of a trajectory. We macro-average this distance over all globally eligible trajectories and retain:

$$\mathrm{ED} = \frac{1}{N_{\mathrm{global}}} \sum_i (1 - (\boldsymbol{f}_i^{\mathrm{ref}})^\top \boldsymbol{f}_i^{\mathrm{end}}). \tag{9}$$

Mean Reference Drift (MRD). To capture the full evolution of subject identity rather than only its endpoint, each valid observation is compared against the initial reference

$$d_{i,t}^{\mathrm{ref}} = 1 - (\boldsymbol{f}_i^{\mathrm{ref}})^\top \boldsymbol{f}_{i,t}.$$

The average long-term deviation of trajectory i is

$$D_{i,\mathrm{mean}}^{\mathrm{ref}} = \frac{1}{|\mathcal{V}_i|} \sum_{t \in \mathcal{V}_i} d_{i,t}^{\mathrm{ref}},$$

where $\mathcal{V}_i$ denotes the valid evaluation frames. We define

$$\mathrm{MRD} = \frac{1}{N_{\mathrm{global}}} \sum_i D_{i,\mathrm{mean}}^{\mathrm{ref}}. \tag{10}$$

Worst-frame Drift (WFD). Worst-frame Drift captures the most severe identity deviation along each trajectory. We average the worst-frame deviations across trajectories and define:

$$\mathrm{WFD} = \frac{1}{N_{\mathrm{global}}} \sum_{i} \max_{t \in \mathcal{V}_i} d_{i,t}^{\mathrm{ref}}. \tag{11}$$

B.4 Details of the Synthetic Experiment

Scene construction. We construct paired synthetic videos to examine how consistency metrics respond to controlled shape perturbations at different temporal scales. Each video contains 480 frames at 16 FPS, with a resolution of 832×480. A total of 20 geometric objects are introduced sequentially along two predefined motion paths. Their initial shapes are sampled from circles, triangles, squares, and pentagons, with radius parameters ranging from 72 to 84 pixels. Each object has a scheduled lifetime of 10 seconds. This lifetime includes entering and leaving the image boundary; the interval during which its contour is visible can therefore be slightly shorter. In the following, object age t denotes elapsed time since its scheduled birth, rather than since its first visible pixel. Object colors, motion trajectories, birth times, and radius parameters are shared across paired conditions.

Smooth global perturbations. A selected object undergoes one source-to-target shape transition according to the mapping $\mathrm{circle} \to \mathrm{triangle}$, $\mathrm{triangle} \to \mathrm{square}$, $\mathrm{square} \to \mathrm{pentagon}$, or $\mathrm{pentagon} \to \mathrm{circle}$. It retains its source shape for $t \in [0, 2]$, transitions during $t \in (2, 8)$, and retains its target shape for the remainder of its lifetime. Let $r_{\mathrm{src}}(\theta)$ and $r_{\mathrm{tgt}}(\theta)$ denote normalized radial profiles of the source and target shapes. We generate the intermediate contour using

$$\begin{aligned} r(\theta, t) &= (1 - \alpha(t)) r_{\mathrm{src}}(\theta) + \alpha(t) r_{\mathrm{tgt}}(\theta), \\ \alpha(t) &= 3u^2 - 2u^3, \\ u &= \mathrm{clip}\left(\frac{t-2}{6}, 0, 1 \right). \end{aligned} \tag{12}$$

For an object with center $\mathbf{c}(t)$ and radius parameter R, the corresponding contour points are

$$\mathbf{p}(\theta, t) = \mathbf{c}(t) + R\, r(\theta, t) \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}. \tag{13}$$

We sample 360 uniformly spaced angular directions and rasterize the resulting contour. The smoothstep interpolation gives zero temporal derivative at the beginning and end of the transition. The center trajectory, radius parameter, and color remain fixed across conditions, while the contour area may change naturally. All conditions use the same contour renderer, including Clean.

Local inconsistency injection. We assign selected objects as evenly as possible to three types of local inconsistency: severe appearance mutations, sudden appearances, and sudden disappearances. Each selected object receives one perturbation. For appearance mutations, we replace the object's current contour with a deeply concave five-pointed star at an age of approximately 5 seconds. The onset is rounded to the nearest native video frame, and the replacement lasts exactly 8 frames (0.5 seconds). The star has an inner-to-outer radius ratio of 0.1 and preserves the object's center, outer radius parameter, and color. Afterward, the object resumes its unperturbed shape evolution. For abrupt appearances, we hide the object during its natural entry and reveal it at an age of 3 seconds. For abrupt disappearances, we remove the object at an age of 6.5 seconds for the remainder of its scheduled lifetime. Both events occur inside the image, with onsets mapped to the first native frame at or after the prescribed time.

Corruption rates and paired controls. We use three random seeds, 101, 202, and 303, and corruption rates $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The denominator of ρ is always the full set of 20 objects, so the five nonzero rates correspond to 2, 4, 6, 8, and 10 perturbed objects. For an object with scheduled birth time b, eligibility for perturbation requires $b \geq 0.5$ and $b + 10 < 28$. These constraints yield 15 eligible objects per seed and ensure that the complete global transition and subsequent departure occur within the protected temporal boundaries. We fix an ordering of eligible objects before evaluation and select nested prefixes as ρ increases. The selected set is identical across Local-only, Global-only, and Local+Global for each seed and rate.

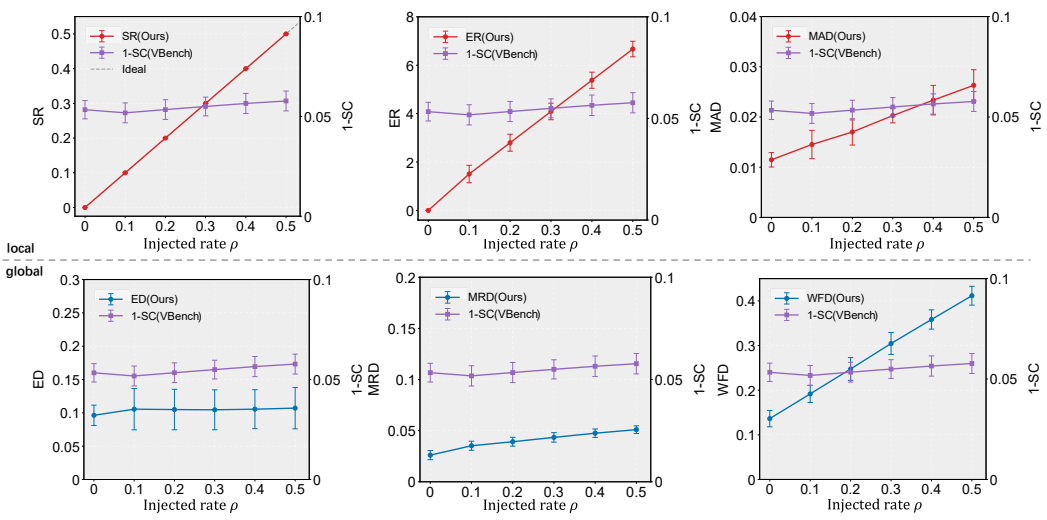

Figure 6: **Metric responses to Local-only inconsistency injection.** As the fraction of objects with local inconsistencies increases, SR and ER rise, while ED remains nearly unchanged. MRD also responds to these temporary deviations from the reference appearance. Each point shows the mean over three random seeds, with error bars indicating one sample standard deviation. VBench is shown as $1 - \mathrm{SC}$ on the right axis.

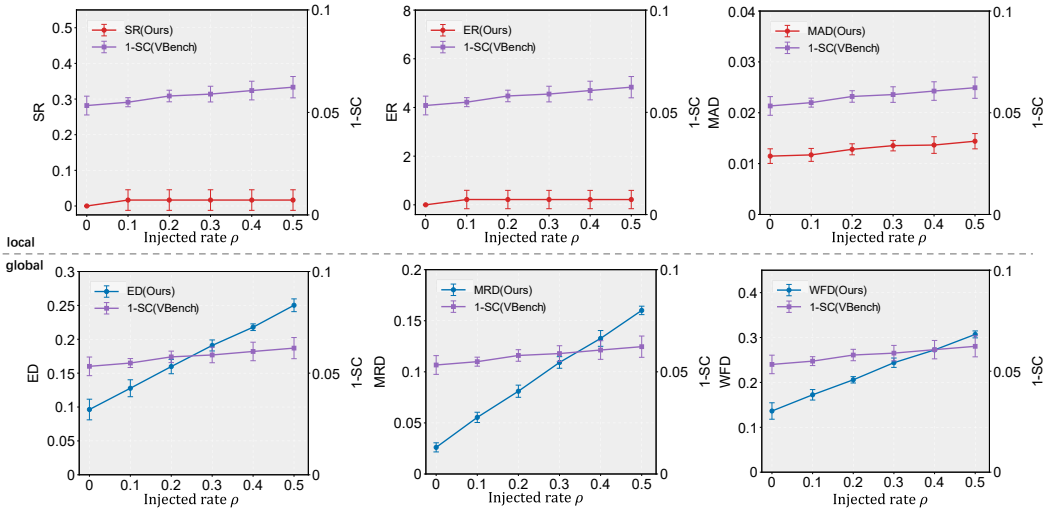

Figure 7: **Metric responses to Global-only inconsistency injection.** As the fraction of objects undergoing smooth shape transitions increases, ED and MRD rise, while SR and ER remain so close to clean baseline and change minimally. Each point shows the mean over three random seeds, with error bars indicating one sample standard deviation. VBench is shown as $1 - \mathrm{SC}$ on the right axis.

B.5 Ablation of Local-only and Global-only Injections

The isolated conditions clarify the different roles of the measurements. Fig. 6 shows the Local-only results. At $\rho = 0.5$, averaged over three seeds, Local-only increases SR from 0 to 0.5000 and ER from 0 to 6.6687 events per 100 track-seconds. ED remains relatively stable, changing from 0.0963 to 0.1071. Conversely, As shown in Fig 7, Global-only increases ED from 0.0963 to 0.2502 and MRD from 0.0261 to 0.1601. In contrast, SR remains very low, with a mean of 0.0167 at every nonzero corruption level and ER only reaches 0.2173 events per 100 track-seconds at $\rho = 0.5$, substantially below its response to Local-only injections. Fig 8 demonstrates the results of local and global inconsistency injection. Under joint injection, SR and ER reach 0.4833 and 6.4790, respectively,

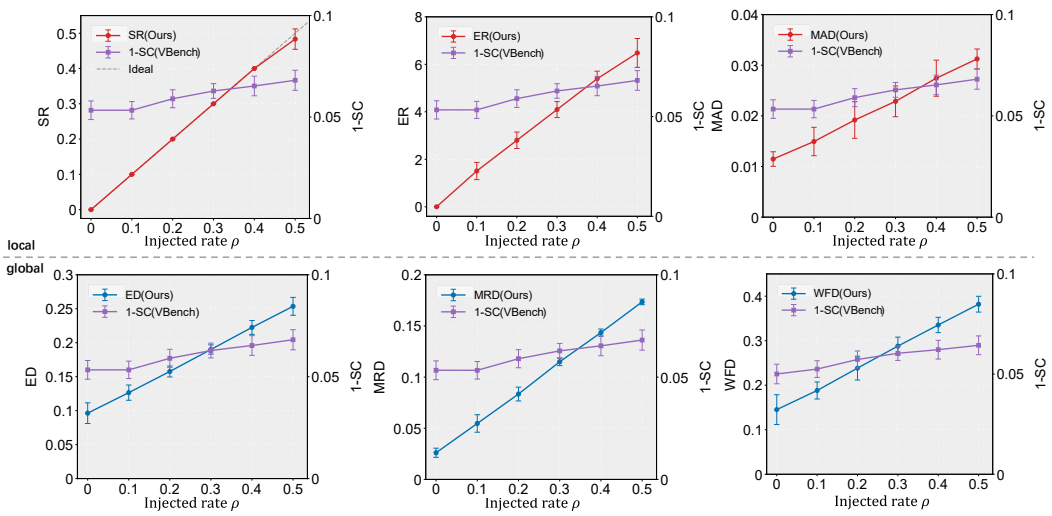

Figure 8: **Metric responses to joint local and global inconsistency injection.** As the fraction of perturbed objects increases, SR closely follows the ideal relation $\mathrm{SR} = \rho$, while ER, ED, and MRD increase. Each point shows the mean over three random seeds, with error bars indicating one sample standard deviation. VBench is shown as $1 - \mathrm{SC}$ on the right axis.

while ED and MRD reach 0.2535 and 0.1735. These observations support the complementary responses of local metrics and global metrics to transient disruptions.

To quantify agreement between measured SR and the injected spike rate, we compare the measurements directly with the identity relation, without fitting a slope or intercept. Let $y_{s,k}$ be the measured SR for seed s at corruption rate ρ_k. For Local-only and Local+Global, we compute

$$R^2_{\text{identity}} = 1 - \frac{\sum_{s,k}(y_{s,k} - \rho_k)^2}{\sum_{s,k}(\rho_k - \bar{\rho})^2}, \qquad \bar{\rho} = 0.25. \tag{14}$$

The calculation includes all 18 seed–rate measurements per condition, including the clean baseline, rather than only the six seed-averaged curve points. For Local+Global, this gives $R^2_{\text{identity}} = 0.9952$, with an MAE of 0.0028 and an RMSE of 0.0118. Local-only yields $R^2_{\text{identity}} = 1.0000$, with both MAE and RMSE equal to zero. All injected local inconsistency episodes were detected, consistent with the intended sensitivity of SR to the injected local inconsistency event. The identity relation is not an appropriate target for Global-only, whose injected spike rate remains zero.

B.6 Prompt Construction

Our evaluation covers four representative scenarios: animals, pedestrians, football matches, and autonomous driving, with 25 prompts per scenario. For reproducibility, the complete set of evaluation prompts used in all experiments is provided in `supporting_materials/prompts/evaluation`.

B.7 Video Generation Prompts

The following boxes demonstrates the text prompts associated with the accompanying video examples in Fig. 1. Paragraphs and headings are formatted for readability.

Prompt: Seedance 2.0 | 15 s

Camera and video style requirements

Prompt: Seedance 2.0 | 15 s (continued)

Every generated video must be a realistic single continuous shot. Required: - The entire video must be captured from one continuous camera trajectory. - The camera should maintain temporal continuity throughout the 15-second duration. - The scene should feel like a real-world observation video. Allowed: - smooth camera tracking; - handheld camera movement; - dashcam view; - fixed surveillance camera; - gradual panning; - natural viewpoint changes. Strictly avoid: - cuts; - scene transitions; - camera switching; - replay; - montage; - time jumps; - sudden viewpoint changes.

Motion requirements

The video should contain normal-speed realistic motion. Required: - Subjects should move at natural real-world speed. - Motion should reflect realistic physics and interactions. Strictly avoid: - slow motion; - slow-motion effects; - cinematic slow-motion shots; - time-lapse; - fast-forward effects; - exaggerated motion; - unrealistic acceleration. The generated video should resemble a real continuous recording rather than a cinematic highlight or edited movie sequence.

Scene specification

An elevated stadium camera records a youth football tournament match between a green team and a white team. Twelve young players are visible, including anchor players: a green captain wearing number 10, a green midfielder wearing number 8, a green defender wearing number 2, a white forward wearing number 9, and a white midfielder wearing number 6. Around 3 seconds, players establish their positions and begin moving around the field. Around 6 seconds, the green team performs several short passes while players rotate positions and opponents follow closely. Around 9 seconds, multiple players converge near the ball, briefly overlapping while maintaining individual appearances. Around 12 seconds, players separate into attacking and defensive formations. The final frames preserve jersey numbers, hairstyles, and player identities. Single continuous shot without slow motion.

Prompt: MiniMax H3 | 15 s

Camera and video style requirements

Every generated video must be a realistic single continuous shot. Required: - The entire video must be captured from one continuous camera trajectory. - The camera should maintain temporal continuity throughout the 15-second duration. - The scene should feel like a real-world observation video. Allowed: - smooth camera tracking; - handheld camera movement; - dashcam view; - fixed surveillance camera; - gradual panning; - natural viewpoint changes. Strictly avoid: - cuts; - scene transitions; - camera switching; - replay; - montage; - time jumps; - sudden viewpoint changes.

Motion requirements

The video should contain normal-speed realistic motion. Required: - Subjects should move at natural real-world speed. - Motion should reflect realistic physics and interactions. Strictly avoid: - slow motion; - slow-motion effects; - cinematic slow-motion shots; - time-lapse; - fast-forward effects; - exaggerated motion; - unrealistic acceleration. The generated video should resemble a real continuous recording rather than a cinematic highlight or edited movie sequence.

Scene specification

An elevated stadium camera records a professional football match between a yellow team and a dark blue team. The scene contains 14 visible players. Five anchor players are a yellow goalkeeper wearing number 1, yellow defender wearing number 4, yellow midfielder wearing number 8, dark blue striker wearing number 10, and dark blue defender wearing number 6. Around 3 seconds, players gather near the penalty area preparing for a corner kick. Around 6 seconds, players begin running and marking opponents as the ball enters play. Around 8 seconds, attackers and defenders move tightly together, causing temporary overlaps and occlusions. Around 11 seconds, players separate and continue tracking the ball. Around 13 seconds, the camera follows the movement while preserving the identity of all anchor players. No replay, no camera cuts.

Prompt: Kling 3.0 | 15 s

Camera and video style requirements

Every generated video must be a realistic single continuous shot. Required: - The entire video must be captured from one continuous camera trajectory. - The camera should maintain temporal continuity throughout the 15-second duration. - The scene should feel like a real-world observation video. Allowed: - smooth camera tracking; - handheld camera movement; - dashcam view; - fixed surveillance camera; - gradual panning; - natural viewpoint changes. Strictly avoid: - cuts; - scene transitions; - camera switching; - replay; - montage; - time jumps; - sudden viewpoint changes.

Motion requirements

The video should contain normal-speed realistic motion. Required: - Subjects should move at natural real-world speed. - Motion should reflect realistic physics and interactions. Strictly avoid: - slow motion; - slow-motion effects; - cinematic slow-motion shots; - time-lapse; - fast-forward effects; - exaggerated motion; - unrealistic acceleration. The generated video should resemble a real continuous recording rather than a cinematic highlight or edited movie sequence.

Scene specification

A realistic dashcam video shows a busy downtown avenue during daytime with a continuous single camera shot from a moving vehicle. The scene contains 12 visible road vehicles, including a red compact sedan, a white SUV, a blue delivery van, a silver city bus, a black pickup truck, a yellow taxi, a gray hatchback, a green motorcycle, a white cargo truck, a dark blue sedan, a silver minivan, and a brown crossover SUV. The red compact sedan, white SUV, blue delivery van, silver city bus, and black pickup truck are anchor vehicles that remain visible throughout the entire 15-second video. Around 3–5 seconds, the yellow taxi accelerates from the right lane and moves ahead of the gray hatchback, while the white SUV gradually slows down due to a traffic light. Around 6–10 seconds, the blue delivery van changes lanes and briefly passes close to the red sedan, causing temporary partial occlusion between the two vehicles while both vehicles remain recognizable. Around 10–13 seconds, the black pickup truck overtakes the silver minivan while the city bus continues straight and the relative positions of the anchor vehicles keep changing naturally. Near the end of the video, all anchor vehicles remain visible with consistent appearance, colors, and identities.

Prompt: LongLive | 60 s

Generate one realistic 60-second video in a single continuous, uninterrupted shot at normal speed. Keep 10-15 subjects visible at detector-friendly scale in a wide or medium-wide view. At least five distinctive anchor subjects must remain continuously visible and preserve identity, color, shape, scale, and markings for the full minute. Use one smooth camera trajectory and physically plausible motion. No cuts, transitions, camera switches, replay, montage, time jumps, slow motion, abrupt viewpoint changes, severe occlusion, tiny distant subjects, teleportation, duplication, disappearance, identity swaps, transformations, text overlays, or watermarks.

A 60-second handheld walking camera video inside a large outdoor shopping street with stores on both sides. There are 13 visible pedestrians. Five anchor pedestrians include a woman with short blonde hair wearing a blue jacket, a man with a red backpack wearing a black shirt, a woman wearing a pink hat and beige coat, a teenager wearing a green hoodie, and an elderly person wearing a navy sweater with glasses. During 0-10 seconds, pedestrians browse shops and walk along the street. During 10-25 seconds, several pedestrians change directions, one person stops briefly near a store entrance, and some pedestrians pass behind others causing temporary partial occlusion. The camera continues moving forward. Single shot, realistic shopping street activity. During 25-40 seconds, the pedestrians vary walking speed and relative position while the camera changes viewpoint gradually and keeps the five anchors in view. During 40-52 seconds, the five anchor pedestrians take part in a second natural crossing and grouping interaction, briefly overlap, then separate with the same clothing, accessories, faces, and body proportions. During 52-60 seconds, normal walking continues and all five anchor pedestrians remain visible, recognizable, and unchanged through the final frame.

Prompt: LingBot-World 2.0 | 60 s

Generate one realistic 60-second video in a single continuous, uninterrupted shot at normal speed. Keep 10-15 subjects visible at detector-friendly scale in a wide or medium-wide view. At least five distinctive anchor subjects must remain continuously visible and preserve identity, color, shape, scale, and markings for the full minute. Use a locked-off static camera with zero translation, zero rotation, and no zoom; only the subjects move, with physically plausible motion. No cuts, transitions, camera switches, replay, montage, time jumps, slow motion, abrupt viewpoint changes, severe occlusion, tiny distant subjects, teleportation, duplication, disappearance, identity swaps, transformations, text overlays, or watermarks.

A 60-second fixed wide-angle camera video showing pedestrians through a university campus pathway during afternoon. The scene contains 11 visible pedestrians. Five anchor pedestrians remain throughout: a student wearing a red jacket with a black backpack, a student wearing a white sweater and glasses, a tall person wearing a blue cap and denim jacket, a woman with braided hair carrying a laptop bag, and an older professor wearing a gray coat. During 0-10 seconds, students walk between buildings, some talking in pairs and others walking alone. During 10-25 seconds, two groups pass each other, several pedestrians briefly overlap, and one student walks behind another before becoming visible again. The locked-off camera maintains a stable wide view. Natural campus environment, single continuous shot. During 25-40 seconds, the pedestrians vary walking speed and relative position. During 40-52 seconds, the five anchor pedestrians take part in a second natural crossing and grouping interaction, briefly overlap, then separate with the same clothing, accessories, faces, and body proportions. During 52-60 seconds, normal walking continues and all five anchor pedestrians remain visible, recognizable, and unchanged through the final frame.

Prompt: Helios | 60 s

Generate one realistic 60-second video in a single continuous, uninterrupted shot at normal speed. Keep 10-15 subjects visible at detector-friendly scale in a wide or medium-wide view. At least five distinctive anchor subjects must remain continuously visible and preserve identity, color, shape, scale, and markings for the full minute. Use one smooth camera trajectory and physically plausible motion. No cuts, transitions, camera switches, replay, montage, time jumps, slow motion, abrupt viewpoint changes, severe occlusion, tiny distant subjects, teleportation, duplication, disappearance, identity swaps, transformations, text overlays, or watermarks.

A realistic 60-second single continuous shot from a moving dashcam records a large urban intersection in daylight. Keep 13 road vehicles visible at detector-friendly scale, including five anchors: a red bus, a white SUV, a blue taxi, a silver cargo truck, and a black sedan. The five anchors remain continuously visible and retain the same colors, body types, proportions, and details. During 0-10 seconds, establish every anchor in normal traffic. During 10-25 seconds, the signal changes and traffic begins crossing; vehicles cross in depth with only brief partial overlap. During 25-40 seconds, turning cars cross the camera view while a bus continues straight, while the camera viewpoint changes smoothly and realistic spacing is maintained. During 40-52 seconds, the anchors complete another gradual merge or ordering change, briefly overlap, and reappear as the same vehicles. During 52-60 seconds, traffic continues naturally and all five anchors remain recognizable through the final frame. No collisions, cuts, teleportation, duplication, disappearance, vehicle transformation, or abrupt camera motion.

C CURRICULUM TRAINING

C.1 SUPERVISED FINE-TUNING OF 5S MODELS

We fine-tune the pretrained Wan-2.1-1.3B and SANA-2B using the same dataset and training protocol. The dataset comprises 72,000 nominal five-second OpenDV clips, totaling approximately 100 hours. A fixed split with seed 314159 partitions the data at the parquet-file level into 68,400 training clips and 3,600 validation clips (95%/5%). Each input contains 77 frames at a resolution of 480×832 (height $\times$ width), corresponding to 20 latent temporal positions. We use precomputed video latents and update all transformer parameters with full bidirectional attention.

Given a clean latent $\boldsymbol{z}$, Gaussian noise $\boldsymbol{\epsilon}$, noise level σ, and text condition $\boldsymbol{c}$, we optimize the conditional flow-matching objective:

$$\boldsymbol{z}_\sigma = (1-\sigma)\boldsymbol{z} + \sigma\boldsymbol{\epsilon}, \qquad \mathcal{L} = \mathbb{E}\left[\frac{1}{D}\,|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_\sigma, \sigma, \boldsymbol{c}) - (\boldsymbol{\epsilon} - \boldsymbol{z})|_2^2\right], \tag{15}$$

where D denotes the number of latent elements per sample. We sample scheduler indices uniformly, use a flow shift of 3, and disable conditioning dropout.

Table 6: Settings for 5s, 10s, and 30s fine-tuning.

Setting	**5s**	**10s**	**30s**
Training clips	68,400	34,208	24,016
Validation clips	3,600	1,792	48
Optimizer	AdamW		
Adam betas	$(0.9, 0.999)$		
Adam epsilon	10^{-8}		
Learning-rate schedule	Cosine decay with linear warm-up		
Weight decay	0.01		
Gradient clipping	1.0		
Training epochs	1		
Peak learning rate	10^{-5}	3×10^{-6}	3×10^{-6}
Warm-up steps	428 (5%)	107 (5%)	150 (10%)
Training updates	8,550	2,138	1,501
Clips per validation	128	128	32

Both models are trained for one epoch (8,550 updates) on eight A100-GPUs, with one sample per GPU and no gradient accumulation, yielding an effective batch size of eight. The learning rate increases linearly over the first 428 updates and subsequently follows a cosine decay. To monitor training stability, we evaluate 128 fixed held-out clips at initialization and every 100 updates, using fixed noise samples and timesteps with seed 314159. We record the overall velocity MSE and its low-, mid-, and high-noise components. Fig. 9 and Fig. 10 show the training curves for Wan-2.1-1.3B and SANA-2B, respectively. Both models exhibit stable training and reduced validation loss. However, validation velocity MSE alone does not reliably indicate generated video quality: despite these improvements in both models, our final consistency evaluation and visual inspection favor the fine-tuned Wan-2.1-1.3B model. Given its stronger empirical performance and our limited computational budget, we use Wan-2.1-1.3B exclusively for subsequent fine-tuning experiments at 10s and 30s.

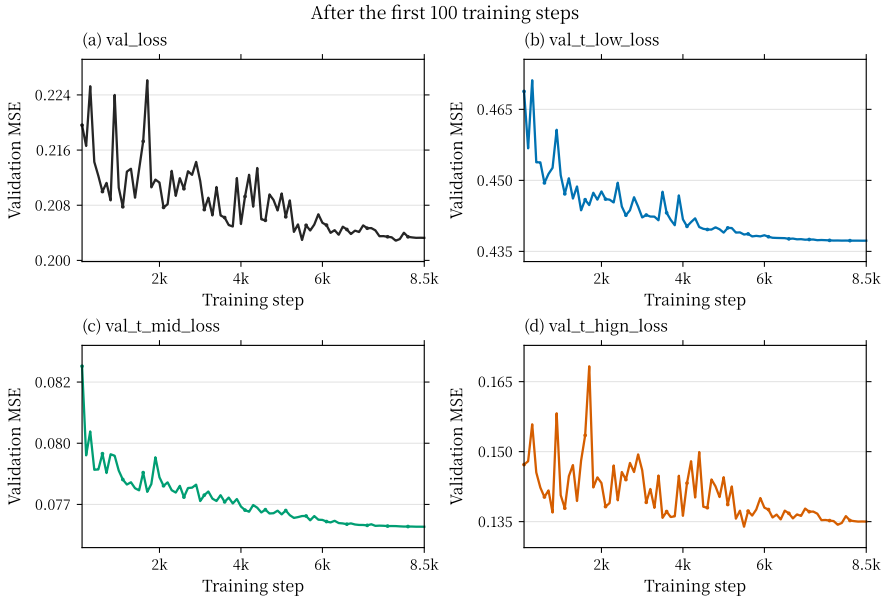

Figure 9: **Validation loss during 5s Wan-2.1-1.3B fine-tuning.**

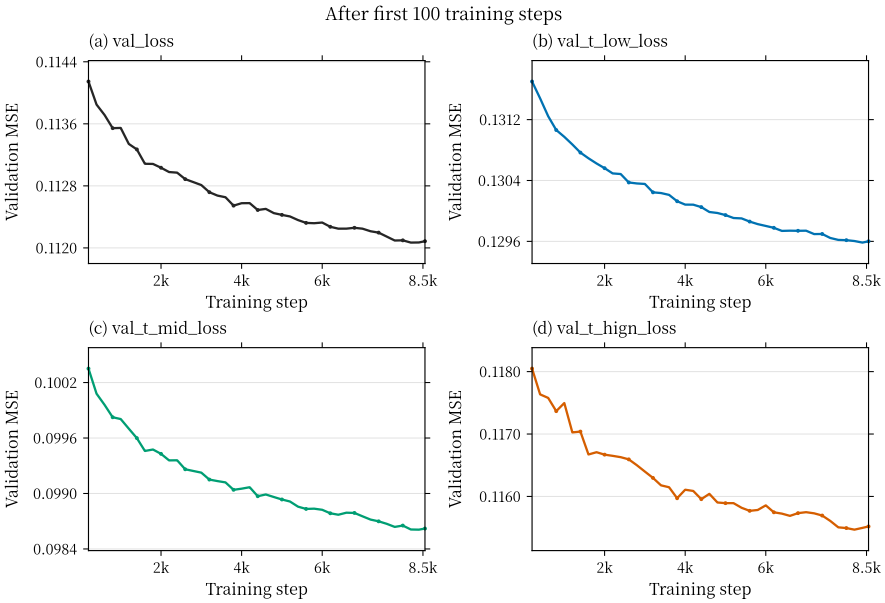

Figure 10: **Validation loss during 5s SANA-2B fine-tuning.**

C.2 Extending to Longer Videos

Fine-tuning on 10s and 30s clips follows the same optimization setup as the 5s stage, with the corresponding hyperparameters summarized in Tab. 6. We use dense spatiotemporal attention for both durations to enable information exchange across the entire video sequence. The 10s experiments are conducted on two nodes with eight A100 GPUs each, while the 30s experiments use four such nodes. As shown in Fig. 11, the validation loss decreases steadily for both durations and gradually stabilizes toward the end of training, indicating stable optimization when extending the model from 5s to longer video generation.

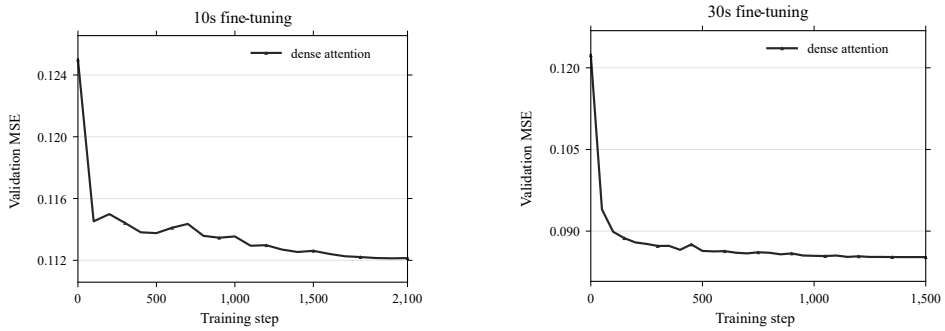

Figure 11: **(a). Validation loss during 10s fine-tuning. (b). Validation loss during 30s fine-tuning.**

D Details of Consistency Post-Training

D.1 Prompt Generation for DiffusionNFT Post-Training

All prompts used for DiffusionNFT post-training at 5s, 10s, and 30s are provided in `supporting_materials/prompts/diffusionnft/train`. The separate prompt sets used exclusively for evaluation are provided in `supporting_materials/prompts/diffusionnft/validation`. These files contain the complete duration-specific training and evaluation prompts used in our experiments.

D.2 Implementation Details of Consistency Post-Training

Policy parameterization. We freeze the base diffusion transformer and optimize only LoRA adapters inserted into the `to_q`, `to_k`, `to_v`, and `to_out` attention projections. The LoRA rank is 32, with a scaling factor of $\alpha = 64$. This lightweight setup aims to improve temporal consistency while preserving the model's overall generative capabilities and maintaining efficient training.

Three-policy optimization. We maintain three policies during training. The *current policy* contains the trainable LoRA parameters and is optimized using AdamW. The *old policy* generates training rollouts and is updated from the current policy using an exponential moving average after each optimizer step. The *reference policy* is the frozen domain-adapted model obtained by disabling the LoRA adapters. The old policy stabilizes rollout generation, while the reference policy constrains the optimized model from deviating excessively from its original generation behavior. For each prompt, the old policy generates a group of $G = 8$ videos. Their rewards are standardized within the group, enabling the current policy to learn from relative consistency differences under the same conditioning prompt. In the 5s setting, we accumulate 4 prompt groups before each optimizer update, resulting in 32 rollouts per update. Each rollout uses 10 diffusion sampling steps. For efficient optimization, we stratify the rollout trajectory and sample 4 diffusion timesteps for computing the training objective.

Table 7: **Ablation on regularization for Wan-2.1-1.3B on 5s generation.** Both optimized variants use the same combined consistency reward and are evaluated at Step 75. Regularization consistently improves DynSC-Eval metrics while preserving motion dynamics. DynSC-Eval metrics use four significant figures, and VBench scores use a 0–100 scale.

	DynSC-Eval Metrics						*VBench Metrics*			
	Local Consistency			*Global Consistency*						
Setting	SR↓	ER↓	MAD↓	ED↓	MRD↓	WFD↓	Mot↑	Dyn↑	Aes↑	Img↑
5s baseline	0.2560	9.177	0.1059	0.3837	0.2393	0.4428	98.35	92.00	50.47	71.40
w/o Reg.	0.2397	7.853	0.09514	0.3514	0.2223	0.4096	98.36	91.00	51.10	73.13
w/ Reg.	**0.2220**	**7.139**	**0.08829**	**0.3446**	**0.2129**	**0.4002**	98.49	92.00	50.81	71.82

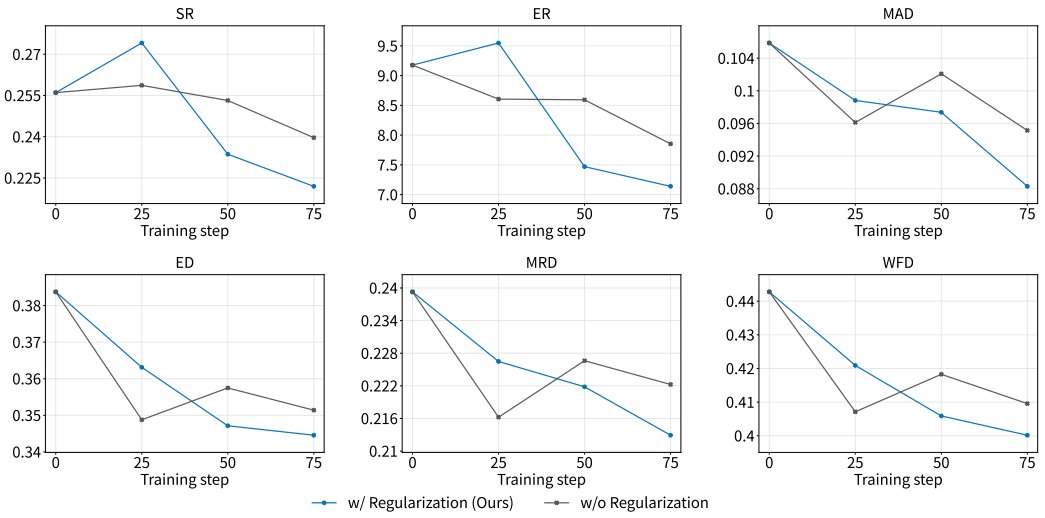

Figure 12: **Effect of reference regularization on video consistency during fine-tuning**

Reference regularization. Optimizing consistency alone may encourage degenerate solutions, such as motion suppression or excessive deviation from the pretrained model. We therefore regularize the current policy against a frozen reference policy using spatial and temporal constraints. Table 7 compares consistency optimization with and without regularization and Fig. 12 shows the validation curves of all six metrics. Even without regularization, optimizing the DynSC-Eval reward improves all six consistency metrics over the 5s baseline. Adding regularization further improve both local and global metrics, decreasing SR from 0.2397 to 0.2220 and ER from 7.853 to 7.139, while also improving MAD, ED, MRD, and WFD. Importantly, these gains do not come from suppressing motion: the VBench motion score remains nearly unchanged (98.36 vs. 98.49), while dynamic degree increases from 91.00 to 92.00. This indicates that regularization stabilizes reward optimization and yields additional consistency improvements without sacrificing video dynamics.