TOKEN FORCING: GEOMETRY-GUIDED TOKENS FOR 3D-CONSISTENT LONG VIDEO GENERATION

Anonymous authors
Paper under double-blind review

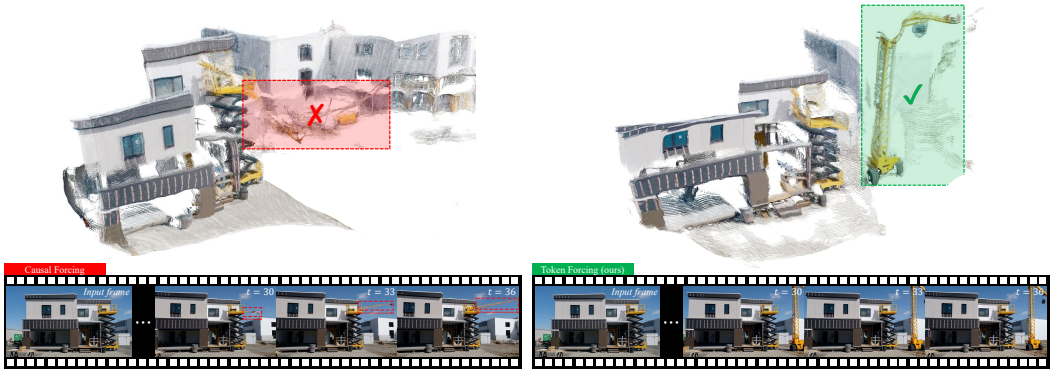

Figure 1: **Long autoregressive rollouts lose 3D structure.** As the rollout proceed, the baseline (Zhu et al., 2026) degrades the crane and its reconstructed geometry collapses (red), while our proposed method Token Forcing keeps the structure intact under an identical camera (green).

ABSTRACT

Autoregressive video diffusion models generate long videos block by block, but over minute-long rollouts their scenes drift as the camera moves. Each block is committed once generated, so geometric errors accumulate, and what remains controllable is which content later blocks rely on. Existing approaches use the generator's own signals, require retraining, or score whole candidates with external rewards. Yet a generated block is rarely uniformly right or wrong, so the question is not which candidate to keep, but which parts of a single generation later blocks should build on. We introduce Token Forcing, a training-free method that answers this question token by token. During the early denoising stages of each block, a frozen recurrent 3D foundation model, whose state summarizes the committed history, keeps the tokens it judges consistent with the scene so far. The generator re-encodes these tokens at the matching noise level and injects their key–value features into the next block at the same denoising stage, where they act as soft geometric anchors. Because the 3D model only decides which tokens to use, neither model is trained. To measure drift over long camera trajectories, we introduce SceneLongBench, 100 static indoor and outdoor scenes for 60-second text-to-video and text-and-image-to-video generation. Relative to Causal Forcing, Token Forcing preserves video quality on motion-rich prompts while reducing reprojection error in all four SceneLongBench settings and epipolar error by up to 21.6%.

1 INTRODUCTION

Autoregressive video diffusion models generate long videos block by block, each conditioned on the blocks before it (Yin et al., 2025; Huang et al., 2025c; Liu et al., 2025b; Zhu et al., 2026). Over minute-long rollouts, the scene drifts as the camera moves: structures deform, spatial relationships change, and newly revealed regions contradict earlier views (Figure 1), even while the video stays temporally smooth. This is partly structural. Unlike a bidirectional model, which denoises all frames of a clip jointly, an autoregressive model fixes each block once it leaves a short rolling

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.

window (Liu et al., 2025b), and all later blocks must build on it. Since these models are also distilled on clips of only a few seconds (Huang et al., 2025c; Liu et al., 2025b), geometric errors that enter the history are carried forward and accumulate (Figure 2).

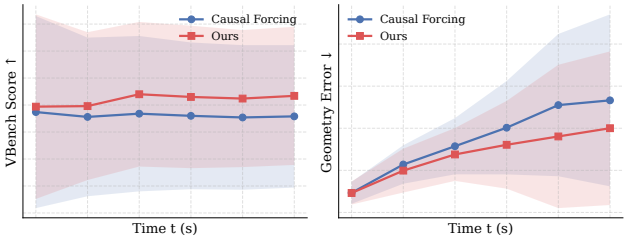

Figure 2: **Geometric consistency and video quality over generation length.** As generation length increases, Causal Forcing exhibits growing geometric error. Token Forcing slows this degradation while achieving higher VBench scores.

Because committed blocks cannot be revised, what remains controllable is which content later blocks rely on. Existing efforts fall into three lines. (i) *Training-free context management* selects cached content using the generator's own signals, such as temporal position, attention scores, or key similarity (Yi et al., 2025; Zhao et al., 2026; Cai et al., 2026), which do not measure geometric consistency. (ii) *Training-based geometric methods* bring in external 3D information through conditions, representation alignment, geometric rewards, or geometry-indexed memory (Huang et al., 2025b; Yu et al., 2025b; Wu et al., 2026; An et al., 2026; Yu et al., 2025a; Li et al., 2025), but require retraining the generator. (iii) *Inference-time reward-guided selection* scores candidates or chunks with external rewards and keeps the best, sometimes also to decide which chunks stay in memory (Liu et al., 2025a; Yin et al., 2026; Tu et al., 2026); it needs no training, but judges each candidate or chunk as a whole.

This whole-unit judgment is the key limitation. A generated block is rarely uniformly right or wrong: in Figure 1, the building stays consistent with earlier views while the crane collapses, and a chunk-level judgment must keep or discard both. The question is therefore not which candidate to keep, but which parts of a single generation later blocks should build on, a judgment that requires neither retraining nor multiple samples.

We introduce **Token Forcing**, a training-free method that answers this question at the token level. It pairs the video generator with a frozen recurrent 3D geometry foundation model (Wang et al., 2025b; Chen et al., 2026b), whose state is updated with every committed block and thus summarizes the scene generated so far. During the early denoising stages of each block, we decode the generator's intermediate clean prediction, evaluate it with the geometry foundation model conditioned on this state, and keep the tokens with high confidence and strong state attention (*history-conditioned token selection*). The generator re-encodes the prediction at the matching noise level, and the key–value (KV) features at the selected positions are injected into the next block at the same denoising stage (*stage-aligned KV injection*). Because the geometry foundation model only decides which tokens to use and the generator encodes them itself, both models remain frozen, and no adapter is needed. The selected tokens thus serve as soft geometric anchors, steering the next block toward the existing scene while its structure is still forming.

Existing benchmarks mostly evaluate clips of a few seconds and cannot reveal geometric drift over long camera trajectories. We therefore introduce a new benchmark, *SceneLongBench*, comprising 100 indoor and outdoor scenes for 60-second text-to-video (T2V) and text-and-image-to-video (TI2V) generation.

Our contributions are summarized as follows:

- We show that improving geometric consistency in autoregressive video generation calls for token-level judgment within a single generation, and propose Token Forcing, a training-free method that couples a frozen video generator with a frozen recurrent 3D foundation model.
- We develop history-conditioned token selection, which evaluates intermediate clean predictions against the accumulated scene state, and stage-aligned KV injection, which transfers the selected content to the next block through noise-matched, generator-native KV features.
- We introduce SceneLongBench, comprising 100 indoor and outdoor scenes for 60-second T2V and TI2V evaluation. Token Forcing leads on all six outdoor T2V geometric metrics, reducing epipolar error by 21.6% and TI2V pointwise reprojection error by up to 13.1%

relative to Causal Forcing (Zhu et al., 2026). It also achieves the highest VBench metric average on motion-rich prompts.

2 Related Work

2.1 Autoregressive Long-Video Generation

Autoregressive video diffusion enables causal generation through per-token noise modeling and diffusion distillation (Chen et al., 2024; Yin et al., 2025; Zhu et al., 2026). Long-horizon consistency is further improved through autoregressive rollout training and extended-context supervision (Huang et al., 2025c; Cui et al., 2025; Yang et al., 2026; Liu et al., 2025b; Chen et al., 2026a). Complementary approaches expand usable history through learned memory and context querying (Henschel et al., 2025; Yu et al., 2026a), or compress historical observations into bounded contexts (Zhang et al., 2025; Oshima et al., 2025). Training-free methods instead manage cached context through attention-guided compression, structured memory, and selective KV retention (Yi et al., 2025; Zhao et al., 2026; Luo et al., 2026; Cai et al., 2026). Token Forcing complements these directions by using external reconstruction signals to select conditioning content during inference.

2.2 Geometry-Aware Video Generation

Geometry-aware methods condition generation on explicit scene representations (Ren et al., 2025; Yu et al., 2025b; Huang et al., 2025b), align generator features with geometric representations (Wu et al., 2026), or optimize geometric objectives through post-training (Kupyn et al., 2025; Yin et al., 2026; An et al., 2026; Ackermann et al., 2026; Li et al., 2026). VIGOR (Yin et al., 2026) also uses geometric rewards for inference-time generation-path verification. Another line organizes scene memory through camera overlap, spatial indexing, and geometry-aware retrieval (Yu et al., 2025a; Li et al., 2025; Huang et al., 2025a; Yu et al., 2026b). These approaches establish geometry as both a generation prior and a cue for context retrieval. Token Forcing uses history-conditioned reconstruction signals to select tokens from intermediate clean predictions and injects their stage-aligned, generator-native KV features as additional context, with both models frozen.

3 Method

We first review the two frozen components (Sec. 3.1), then give an overview of Token Forcing (Sec. 3.2) and detail its token selection (Sec. 3.3) and KV injection (Sec. 3.4).

3.1 Preliminaries

Autoregressive video generation. We represent a long video as a sequence of blocks $\{B_t\}_{t=1}^{T}$ and adopt the rolling diffusion framework (Liu et al., 2025b; **?**). Each block consists of N latent tokens over several latent frames, decodes to F video frames, and passes through K denoising stages. Multiple blocks are denoised simultaneously at different noise levels, with block $t+1$ always one stage behind block t. At stage k, the generator G_θ produces an intermediate clean prediction $\hat{x}_{0,t}^{(k)}$, conditioned on the rolling window and the KV cache. A block is committed after its final denoising stage and re-encoded into the cache as clean context.

Recurrent 3D geometry foundation models. A recurrent geometry foundation model Φ (Wang et al., 2025b; Chen et al., 2026b) processes frames sequentially while maintaining a scene state S. For each frame I_j, Φ predicts camera-frame and world-frame pointmaps with per-pixel confidences c^{cam} and c^{world}, conditioned on the preceding state. Cross-attention between state and image tokens then incorporates the current observation into the state. We write this update function as $S_j = f_{\mathrm{upd}}(S_{j-1}, I_j)$.

3.2 Overview

As motivated in Section 1, a generated block is often only partly consistent with the scene generated so far: some regions agree with earlier views, while others contradict them. Token Forcing lets the

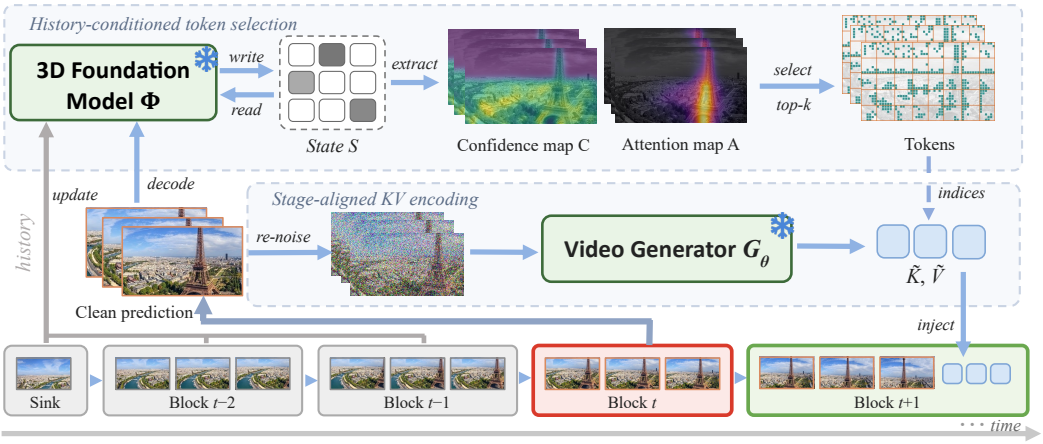

Figure 3: **Overview of Token Forcing.** During autoregressive video generation, Token Forcing identifies regions of the current block that are consistent with the previously generated scene and uses them to condition the next block. *History-conditioned token selection:* the frozen 3D geometry model Φ evaluates the intermediate clean prediction of the current block against the committed history and selects high-confidence tokens. *Stage-aligned KV injection:* the frozen generator G_θ encodes the selected tokens at the matching noise level and injects their KV features into the next block at the same denoising stage.

next block build on the regions that agree. The two frozen models play distinct roles. The geometry model Φ judges which regions of the current block agree with the committed history, and the video generator G_θ encodes those regions into its own key–value (KV) features and uses them as additional context. Because Φ passes only token positions to G_θ, neither model is trained and no mapping between their feature spaces is learned. Figure 3 illustrates the pipeline.

Consider a *source* block t at an early denoising stage $k \in \mathcal{K}$, and the *receiving* block $t+1$, which reaches stage k one step later under the rolling schedule. Let S_{hist} denote the state of Φ after processing all committed blocks, $\mathcal{D}$ the VAE decoder, $x_{t+1}^{(k)}$ the noisy latent of block $t+1$ at stage k, and $\mathcal{C}_t^{(k)}$ the context available to block t at stage k under the base generator, namely the rolling window, the KV cache, and the text condition. Token Forcing performs three steps:

$$\mathcal{A}_t^{(k)} = \text{Select}\big(\Phi(\mathcal{D}(\hat{x}_{0,t}^{(k)}) \mid S_{\text{hist}})\big), \tag{1}$$

$$(\tilde{K}_t^{(k)}, \tilde{V}_t^{(k)}) = \text{KV}_{G_\theta}\big(\text{Renoise}_k(\hat{x}_{0,t}^{(k)}) \,\big|\, \mathcal{C}_t^{(k)}\big)\big|_{\mathcal{A}_t^{(k)}}, \tag{2}$$

$$\hat{x}_{0,t+1}^{(k)} = G_\theta\big(x_{t+1}^{(k)} \,\big|\, \mathcal{C}_{t+1}^{(k)} \cup \{\tilde{K}_t^{(k)}, \tilde{V}_t^{(k)}\}\big). \tag{3}$$

First, Φ scores the decoded source prediction under the committed history and selects the token positions $\mathcal{A}_t^{(k)}$, as in Equation (1). Next, G_θ encodes the re-noised source prediction under the source context $\mathcal{C}_t^{(k)}$ and keeps the KV features at these positions in every layer, as in Equation (2); we omit the layer index ℓ for brevity. Finally, the receiving block takes its ordinary denoising step with these features added to its context, as in Equation (3). Removing the added features recovers the base generator exactly, since the weights, sampler, and noise schedule are unchanged.

We apply these steps only at the early stages $\mathcal{K}$, while the coarse structure of a block is still being formed, and recompute $\mathcal{A}_t^{(k)}$ for every source block and stage. The injected KV features act as soft geometric anchors: they impose no explicit 3D coordinates or correspondence constraints, and the generator's existing attention weights determine how strongly they influence the receiving block. Two design questions remain: which tokens to select, and how to encode them for the receiving pass. Section 3.3 addresses the first with *history-conditioned token selection*, which specifies Select and how S_{hist} is maintained. Section 3.4 addresses the second with *stage-aligned KV injection*, which specifies Renoise_k and the per-layer form of Equation (3).

3.3 HISTORY-CONDITIONED TOKEN SELECTION

To select content that supports consistency across views, we use the committed history to inform the evaluation of the current prediction. The state of Φ provides this historical context, while its confidence and state attention guide the choice of tokens.

History-conditioned rollout. We maintain a persistent state S_{hist} by processing the decoded frames of committed blocks with Φ in temporal order. This state carries scene information beyond the generator's context window and provides the starting point for evaluating a source prediction. Under the rolling schedule, the blocks between the last committed block and the source block are still being denoised and are not yet included in S_{hist}, so the history may lag behind the source block by a few blocks.

An intermediate prediction may still change during denoising, so evaluating it should not alter the committed history. For $k \in \mathcal{K}$, we therefore decode $\hat{x}_{0,t}^{(k)}$ into frames $\{I_{t,j}^{(k)}\}_{j=1}^{F}$ and process them using a temporary copy of S_{hist}:

$$\tilde{S}_{t,0}^{(k)} = S_{\text{hist}}, \qquad \tilde{S}_{t,j}^{(k)} = f_{\text{upd}}\Big(\tilde{S}_{t,j-1}^{(k)}, I_{t,j}^{(k)}\Big). \tag{4}$$

Each frame is thus evaluated conditioned on both the committed history and the preceding frames of the source prediction. We collect the resulting confidences and state attention for scoring, then discard the temporary state. Only committed blocks update S_{hist}.

Reliability score. The confidence of Φ measures its certainty about the predicted geometry. We combine confidence in the camera-frame and world-frame predictions so that selection accounts for both local geometry and its placement in the shared coordinate frame. At image location p, we define

$$C(p) = \sqrt{c^{\text{cam}}(p)\, c^{\text{world}}(p)}. \tag{5}$$

The geometric mean limits the extent to which high confidence in one prediction compensates for low confidence in the other.

Confidence alone does not describe how a region relates to the committed history. We therefore also use *state attention* $A(p)$, obtained by aggregating the pre-softmax cross-attention logits from state queries to image keys. State attention serves as an auxiliary cue, with confidence remaining the primary reliability signal.

We normalize the two signals to $[0, 1]$, denoted $\bar{C}(p)$ and $\bar{A}(p)$, and combine them as

$$R(p) = V(p)\, \bar{C}(p)\, \big[(1-\lambda) + \lambda \bar{A}(p)\big], \qquad 0 \le \lambda \le 1, \tag{6}$$

where the binary mask $V(p)$ removes invalid predictions of Φ. Confidence bounds the score, while state attention modulates selection priority within that bound. Aggregating $R(p)$ over the decoded frames and pixels corresponding to each latent token yields token scores R_i.

Token selection. A token is eligible for selection only if a sufficient fraction of its corresponding pixels is valid under V. Selecting the highest-scoring tokens globally could concentrate the selected tokens in a small part of the scene. To preserve spatial coverage, we partition each latent frame's token grid into spatial cells and rank eligible tokens independently within each cell. For a cell containing n_c tokens, we retain up to $\lceil \rho_k n_c \rceil$ tokens with the highest scores, where ρ_k is the retention ratio at stage k. The union over cells and latent frames forms $\mathcal{A}_t^{(k)}$.

This procedure combines a validity requirement with relative ranking and spatial coverage. Because ranking is performed within each prediction, it does not rely on confidence calibration across blocks and denoising stages.

3.4 STAGE-ALIGNED KV INJECTION

Selection identifies which tokens to reuse; injection must present their content in a form suitable for the receiving block. We obtain this representation through the generator itself, matching the noise level at which the features will be consumed.

Table 1: **Text-to-video (T2V) generation on *SceneLongBench*** (60 seconds). Token Forcing achieves the best results on all geometric metrics for outdoor scenes, reducing EPI by 21.6% relative to Causal Forcing, and the best SSIM, RPX, and RPT for indoor scenes.

Scene	Method	Geometric Consistency						VBench (%)		
		PSNR↑	SSIM↑	LPIPS↓	EPI↓	RPX↓	RPT↓	TQ↑	FQ↑	Avg↑
Indoor	Deep Forcing	17.12	0.53	0.45	**9.93**	0.47	9.16	83.83	**64.19**	77.28
	Causal Forcing	17.27	0.57	**0.39**	11.40	0.41	7.23	96.82	62.06	85.23
	Relax Forcing	**17.59**	0.53	0.56	19.07	0.58	17.09	94.86	60.18	83.30
	Ours	17.50	**0.59**	**0.39**	10.02	**0.39**	**6.62**	**97.01**	62.59	**85.54**
Outdoor	Deep Forcing	16.54	0.51	0.40	10.39	0.42	6.87	85.78	**67.01**	79.52
	Causal Forcing	17.36	0.56	0.36	6.11	0.38	6.15	97.42	64.07	86.30
	Relax Forcing	15.90	0.47	0.53	17.57	0.56	12.95	94.62	65.58	84.94
	Ours	**17.47**	**0.58**	**0.34**	**4.79**	**0.36**	**5.74**	**97.65**	64.42	**86.57**

Noise-matched encoding. The generator's intermediate features depend on the denoising stage, so encoding clean content directly may produce features mismatched to an early, high-noise receiving stage. We instead re-noise the source prediction to the noise level of stage k:

$$\tilde{x}_t^{(k)} = \mathrm{Renoise}_k\big(\hat{x}_{0,t}^{(k)}\big) = \alpha_k \hat{x}_{0,t}^{(k)} + \beta_k \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{7}$$

where α_k and β_k follow the base generator's noise schedule.

An additional forward pass of G_θ encodes $\tilde{x}_t^{(k)}$ under the source context $\mathcal{C}_t^{(k)}$, i.e., the same rolling window, KV cache, and text condition used for its ordinary denoising step. At each layer ℓ, we extract the keys and values at $\mathcal{A}_t^{(k)}$, denoted $\tilde{K}_t^{\ell,(k)}$ and $\tilde{V}_t^{\ell,(k)}$. This pass produces conditioning features without replacing the source block's denoising update.

Injection. The rolling schedule makes these features available before block $t+1$ reaches stage k. At that stage, we concatenate them to the receiving block's attention context:

$$\mathrm{Attn}\Big(Q_{t+1}^{\ell,(k)}, \Big[K_{t+1}^{\ell,(k)}; \tilde{K}_t^{\ell,(k)}\Big], \Big[V_{t+1}^{\ell,(k)}; \tilde{V}_t^{\ell,(k)}\Big]\Big), \tag{8}$$

where $K_{t+1}^{\ell,(k)}$ and $V_{t+1}^{\ell,(k)}$ are the keys and values from $\mathcal{C}_{t+1}^{(k)}$ and the receiving block itself. This context already includes the source block's ordinary rolling-window features, which at this point belong to a later denoising stage. The injected features differ in two ways: they are restricted to the selected positions, and they are encoded at the receiving block's own noise level.

The selected keys are captured after RoPE at the source block's absolute positions and keep those positions during injection. Each set of injected features is used only in the matching forward pass of the receiving block.

Short-range content and long-range history. The injected features supply content from the immediately preceding block, while the committed history determines which parts of that content are selected. Although S_{hist} may lag behind the source block (Sec. 3.3), it still provides the long-range history during selection. Repeating this process block by block lets accumulated scene information guide subsequent generation without requiring the generator to retain the full video history in its attention context.

4 Experiments

4.1 Experimental Settings

Implementation. We build Token Forcing on the released minute-level checkpoint of Causal Forcing (Zhu et al., 2026), based on Wan2.1-T2V-1.3B (Wan et al., 2025), and keep its default sampling configuration, generating 480×832 videos at 16 FPS. We use a frozen TTT3R (Chen et al., 2026b) as the 3D foundation model and apply selection and injection at the first two denoising stages, $\mathcal{K} = \{0, 1\}$, denoted D_0 and D_1, with $\lambda = 0.5$ and retention ratios $\rho_0 = 0.5$ and $\rho_1 = 0.25$. Further details are given in Appendix B.

Table 2: **Text-and-image-to-video (TI2V) generation on *SceneLongBench*** (60 seconds). Compared with Causal Forcing, Token Forcing improves all geometric metrics for indoor scenes and five of six for outdoor scenes.

Scene	Method	Geometric Consistency						VBench (%)		
		PSNR↑	SSIM↑	LPIPS↓	EPI↓	RPX↓	RPT↓	TQ↑	FQ↑	Avg↑
Indoor	Causal Forcing	18.12	0.59	0.35	7.88	0.37	5.96	96.68	59.91	84.42
	Ours	**18.78**	**0.61**	**0.33**	**7.09**	**0.35**	**5.53**	**96.73**	**60.62**	**84.69**
Outdoor	Causal Forcing	17.74	0.56	0.32	**4.95**	0.34	6.16	**96.62**	65.40	**86.21**
	Ours	**18.79**	**0.57**	**0.31**	5.03	**0.32**	**5.35**	95.37	**65.49**	85.41

Benchmarks. We introduce *SceneLongBench*, 100 static scenes in which only the camera moves. Each scene provides a reference image and a caption of the same content, so that text-to-video (T2V) and text-and-image-to-video (TI2V) are conditioned identically. The indoor half consists of 50 RealEstate10K (Zhou et al., 2018) room walkthroughs, using the first frame of each clip as the reference image; the outdoor half consists of 50 landmark photographs from Google Landmarks v2 (Weyand et al., 2020), spanning 50 countries. Every caption specifies a camera move and describes no moving subjects, keeping reconstruction-based metrics well posed. Construction details are given in Appendix C. We also evaluate T2V at 30 and 60 seconds on the motion-rich prompts of Causal Forcing (Zhu et al., 2026).

Evaluation metrics. Following VIGOR (Yin et al., 2026), we evaluate geometric consistency with complementary reconstruction-based and correspondence-based metrics. The former measure PSNR, SSIM, and LPIPS of cross-view renderings from VGGT (Wang et al., 2025a) reconstructions; the latter comprise epipolar error (EPI), pixelwise reprojection error (RPX), and pointwise reprojection error (RPT). We evaluate video quality and motion with six VBench (Huang et al., 2024) metrics: subject consistency (SC), background consistency (BC), motion smoothness (MS), dynamic degree (DD), aesthetic quality (AQ), and imaging quality (IQ). For compact presentation, we report temporal quality (TQ) as the average of SC, BC, MS, and DD; frame quality (FQ) as the average of AQ and IQ; and Avg as the mean of all six metrics.

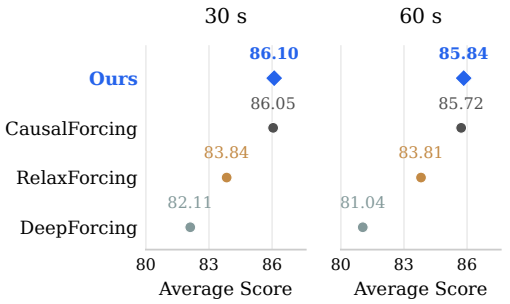

Figure 4: **Evaluation results on motion-rich prompts.** 30- and 60-second T2V generation on the 100 prompts released by Causal Forcing (Zhu et al., 2026). Token Forcing achieves the highest Avg at both durations.

Baselines. For T2V, we compare with Causal Forcing (Zhu et al., 2026), Deep Forcing (Yi et al., 2025), and Relax Forcing (Zhao et al., 2026). Causal Forcing is the unmodified base generator, while Deep Forcing and Relax Forcing represent training-free approaches to long-video context management. For TI2V, we compare only with Causal Forcing, as the other methods do not support image conditioning.

4.2 Main Results

Geometric consistency on SceneLongBench. Tables 1 and 2 report results for 60-second generation. Under outdoor T2V, Token Forcing achieves the best results on all six geometric metrics, including a 21.6% reduction in EPI relative to Causal Forcing. Under indoor T2V, Token Forcing ranks first or second on all geometric metrics: it achieves the best SSIM, RPX, and RPT, ties for the best LPIPS, and is second only to Relax Forcing on PSNR (17.50 vs. 17.59) and to Deep Forcing on EPI (10.02 vs. 9.93).

The improvements extend to image-conditioned generation. Token Forcing outperforms Causal Forcing on all six geometric metrics for indoor TI2V and on five of six for outdoor TI2V, increasing

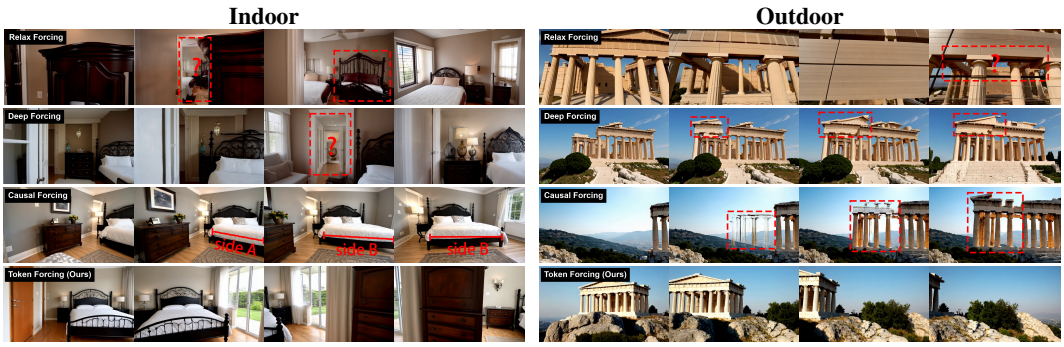

Figure 5: **Qualitative comparison over a 60-second rollout.** Left: indoor. Right: outdoor. Rows top to bottom: Relax Forcing, Deep Forcing, Causal Forcing, Token Forcing. The baselines swap the bedroom for a different one, collapse onto flat masonry, or grow a duplicate colonnade beside the real facade (red dashed boxes); Token Forcing holds one coherent room and one temple stably throughout.

PSNR by 0.66 and 1.05 dB and reducing RPT by 7.2% and 13.1% on indoor and outdoor scenes, respectively. The exception is outdoor TI2V EPI, which increases slightly from 4.95 to 5.03. Overall, Token Forcing improves multiple measures of cross-view consistency, with some variation across scene types and conditioning settings.

Video quality on SceneLongBench. Token Forcing achieves a higher Avg than Causal Forcing on both T2V subsets and on indoor TI2V. Outdoor TI2V exhibits a trade-off: Avg decreases from 86.21 to 85.41, primarily due to lower dynamic degree. Geometric improvements therefore do not uniformly translate into improvements in every aspect of video generation.

For qualitative results, Figure 5 shows frames sampled uniformly from 60-second rollouts on one indoor and one outdoor prompt. The baselines degrade in distinct ways—regenerating the same furniture at inconsistent scales, collapsing onto a degenerate texture, or replicating structures absent from the conditioning frame—all symptoms of context selected by recency rather than geometry. Token Forcing preserves scene layout and object scale in both settings, consistent with its generally lower reprojection errors in Tables 1 and 2.

Motion-rich prompts. To examine whether geometry-guided selection compromises video quality on dynamic, less scene-centric content, we evaluate 30- and 60-second T2V generation on the motion-rich prompts released by Causal Forcing. As shown in Figure 4, Token Forcing achieves the highest Avg among all compared methods at both durations, with 86.10 and 85.84 compared with 86.05 and 85.72 for Causal Forcing. At 60 seconds, it improves SC, BC, and IQ, while DD decreases slightly from 96.83 to 96.13. Given these small margins, we interpret the results as preserving rather than substantially improving video quality beyond scene-centric prompts.

4.3 Ablation Studies

All ablations use 60-second T2V generation on 10 indoor and 10 outdoor prompts randomly sampled from SceneLongBench, with a fixed seed.

Table 3: Effect of anchor selection. Random patches injected with the same side-KV mechanism bring no improvement.

Method	PSNR↑	SSIM↑	LPIPS↓	EPI↓	RPX↓	RPT↓
No injection	**18.00**	0.57	0.36	8.84	0.38	5.66
Random patch	17.92	0.57	0.37	8.84	0.39	5.67
Ours	17.94	**0.58**	**0.35**	**7.52**	**0.37**	**5.22**

Geometry-guided vs. random selection. To test whether the gains come merely from additional KV context, we replace geometry-guided selection with random sampling under the same validity filter, per-region budgets, and injection procedure. Random selection yields no geometric gain over the no-injection baseline (EPI unchanged at 8.84; PSNR 18.00 to 17.92), whereas geometry-guided selection reduces EPI by 15.0% and RPT by 7.6% with PSNR nearly unchanged (Table 3). The benefit thus depends on which content is selected, not on the extra context.

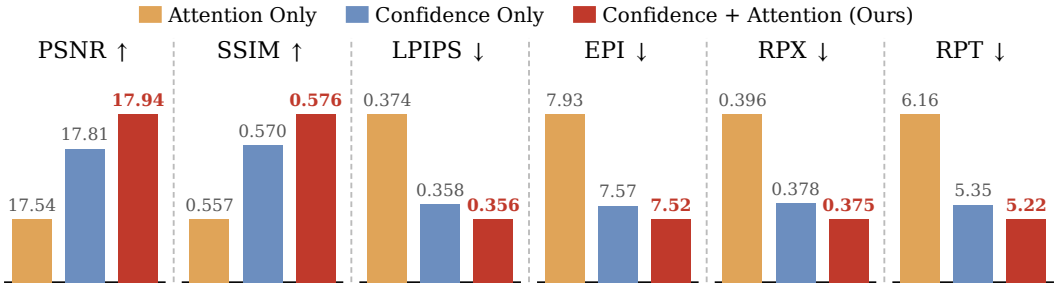

Figure 6: **Ablation of selection signals.** Confidence is the stronger individual signal, and combining it with reconstruction-state attention achieves the best result on all geometric metrics.

Selection signals. Confidence is the stronger individual signal: relative to attention alone, it reduces RPT from 6.16 to 5.35 and EPI from 7.93 to 7.57, and improves all reconstruction metrics (Fig. 6). Adding attention further improves all geometric metrics, reducing RPT to 5.22, which makes it a useful auxiliary signal even though it is insufficient on its own.

Injection stages. Injecting at both D_0 and D_1 outperforms injecting at either stage alone on SSIM and all three correspondence metrics, reducing RPT by 12.1% and 4.2% relative to D_0-only and D_1-only injection, while D_1-only injection attains the highest PSNR (Appendix D). Because joint injection also uses a larger total budget, this comparison does not separate stage coverage from additional context.

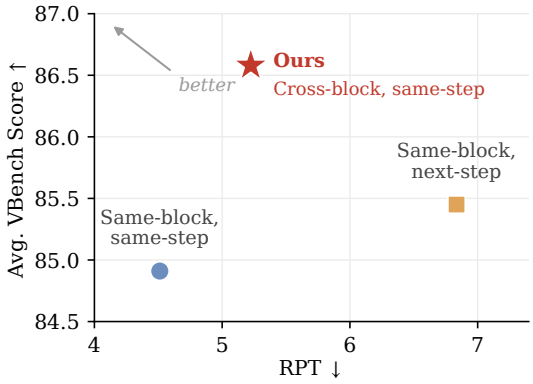

Figure 7: **Ablation of injection configurations.** Cross-block, same-stage injection gives the best balance between geometric consistency and video quality.

Injection location. With selection and encoding fixed at D_0 and D_1, we compare injecting the selected KV features into (t, D_k) by rerunning the current pass (same-block, same-stage), into (t, D_{k+1}) (same-block, next-stage), and into $(t+1, D_k)$ (cross-block, same-stage; our default). Same-stage rerunning attains the lowest RPT but also the lowest Avg (Fig. 7). Our default keeps RPT close to it while raising Avg to 86.58, which is 1.67 and 1.13 points above same-stage rerunning and next-stage injection, respectively, and reduces RPT from 6.84 to 5.22 relative to next-stage injection. Since next-stage injection also changes the receiving stage, this comparison evaluates whole configurations rather than isolating noise-level alignment.

5 Conclusion

We presented Token Forcing, a training-free method that improves the geometric consistency of long-video generation by pairing a frozen recurrent 3D reconstruction model with the generator's own attention. History-conditioned token selection identifies geometrically reliable content, and stage-aligned KV injection passes it to subsequent blocks as soft geometric anchors. On SceneLongBench, Token Forcing improves cross-view consistency in both T2V and TI2V settings, reducing epipolar error by 21.6% on outdoor T2V, while largely preserving video quality on motion-rich prompts. Ablations confirm that these gains come from which content is selected rather than from additional context alone. More broadly, our results suggest that 3D foundation models can guide generation, not only evaluate it.

REFERENCES

Jan Ackermann, Shengqu Cai, Boyang Deng, Zhengfei Kuang, Songyou Peng, and Gordon Wetzstein. GeoFlow: Enforcing implicit geometric consistency in video generation. *arXiv preprint arXiv:2605.18365*, 2026.

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. VGGRPO: Towards world-consistent video generation with 4D latent reward. In *European Conference on Computer Vision (ECCV)*, pp. 305–322, 2026. doi: 10.1007/978-3-032-37447-9_17.

Anthropic. System card: Claude Opus 5. `https://www.anthropic.com/claude-opus-5-system-card`, July 2026. Accessed 2026-09-26.

Peiliang Cai, Evelyn Zhang, Jiacheng Liu, Hao Lin, Ruiqi Zhang, Weile Mo, Yue Ma, Shikang Zheng, Jiehang Huang, et al. Focused forcing: Content-aware per-frame KV selection for efficient autoregressive video diffusion. *arXiv preprint arXiv:2605.18346*, 2026.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 50694–50718, 2026b.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2568–2577. IEEE, 2025.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on Minecraft. *arXiv preprint arXiv:2510.03198*, 2025a.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44(6): 1–15, 2025b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Orest Kupyn, Théo Uscidda, Marta Tintore Gazulla, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699. IEEE, 2025.

Zizun Li, Haoyu Guo, Runzhe Teng, Chunhua Shen, and Tong He. Geo-Align: Video generation alignment via metric geometry reward. *arXiv preprint arXiv:2605.23903*, 2026.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-t1: Test-time scaling for video generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18671–18681. IEEE, 2025a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025b.

Jiayi Luo, Qiyan Liu, Tengyang Wang, JunHao Liu, Jiayu Chen, Cong Wang, Hanxin Zhu, Chen Gao, Xiaobin Hu, et al. Future forcing: Future-aware training-free KV cache policy for autoregressive video generation. *arXiv preprint arXiv:2605.30083*, 2026.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. WorldPack: Dynamic frame compression for long-context video world modeling. *arXiv preprint arXiv:2512.02473*, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yijing Tu, Shaojin Wu, Mengqi Huang, Wenchuan Wang, Yuxin Wang, Chunxiao Liu, and Zhendong Mao. Stream-t1: Test-time scaling for streaming video generation. *arXiv preprint arXiv:2605.04461*, 2026.

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

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2–a large-scale benchmark for instance-level recognition and retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2572–2581. IEEE, 2020.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3D representation for consistent world modeling. In *International Conference on Learning Representations (ICLR)*, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, et al. LongLive: Real-time interactive long video generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tengjiao Yin, Jinglei Shi, Heng Guo, and Xi Wang. VIGOR: VIdeo Geometry-Oriented Reward for Temporal Generative Alignment. In *European Conference on Computer Vision (ECCV)*, 2026. doi: 10.1007/978-3-032-37189-8_14.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *SIGGRAPH Asia 2025 Conference Papers*, 2025a.

Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. MemLearner: Learning to query context memory for video world models. In *European Conference on Computer Vision (ECCV)*, pp. 96–114, 2026a. doi: 10.1007/978-3-032-37595-7_6.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–18, 2025b. doi: 10.1109/TPAMI.2025.3613256.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Sri Siddarth Chakaravarthy P, Dennis Anthony, Yang Ye, Yidi Li, et al. Mosaicmem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 2026b.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025.

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed KV-memory for consistent long video generation. In *British Machine Vision Conference (BMVC)*, 2026.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *International Conference on Machine Learning (ICML)*, 2026.

APPENDIX

A USE OF LARGE LANGUAGE MODELS

In this work, we used generative AI tools only to aid and polish the writing of the manuscript, i.e., to improve the grammar, wording, and readability of text drafted by the authors. We have not used generative AI tools to develop theoretical models, formulate or prove mathematical claims, propose hypotheses, design experiments, implement methods, generate or process data, or interpret results; translation and qualitative data analysis are not applicable to this work. We have reviewed all AI-assisted text: every suggested edit was checked by the authors to ensure it is accurate and faithful to the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

B IMPLEMENTATION DETAILS

The Causal Forcing checkpoint we build on is trained with Rolling Forcing (Liu et al., 2025b). Following its default configuration, we use three latent frames per block and a four-step denoising sampler. Confidence and attention scores are normalized using their 10th and 90th percentiles. Within each latent frame, the 30×52 token grid is partitioned into 5×8 spatial regions, and the retention ratios ρ_0 and ρ_1 are applied within each region; because of rounding at the region level, the realized ratios can be slightly higher than the nominal ones. All methods run on a single NVIDIA RTX PRO 6000 GPU.

C BENCHMARK CONSTRUCTION

For the outdoor half, we select the most widely recognized landmark in each of 50 countries from Google Landmarks v2 and use one photograph of each. All 100 captions were generated with Claude Opus 5 (Anthropic, 2026), from the source clip for indoor scenes and from the photograph for outdoor scenes. Each caption specifies one camera move, and no two outdoor captions share the same move.

D ABLATION OF INJECTION STAGES

Table 4: **Ablation of injection stages.** Injecting at both D_0 and D_1 achieves the best SSIM, LPIPS, and all correspondence metrics.

Method	PSNR↑	SSIM↑	LPIPS↓	EPI↓	RPX↓	RPT↓
D_0 only (50%)	17.70	0.56	0.36	7.88	0.39	5.94
D_1 only (25%)	**18.00**	0.57	0.36	8.98	0.38	5.45
D_0+D_1 (50%/25%)	17.94	**0.58**	**0.36**	**7.52**	**0.37**	**5.22**

Table 4 compares injection at D_0 only, D_1 only, and both stages, following the protocol of Sec. 4.3 with retention ratios $\rho_0 = 0.5$ and $\rho_1 = 0.25$. Joint injection achieves the best SSIM and all three correspondence metrics, while D_1-only injection achieves the highest PSNR. Although D_1-only injection improves the reconstruction metrics over D_0-only injection, its EPI is higher (8.98 vs. 7.88), underscoring the value of complementary geometric metrics.

E LIMITATIONS AND FUTURE WORK.

Token Forcing requires no training, but it is not free at inference time: under our evaluation settings, it roughly doubles the time needed to generate a video. The overhead has three sources: decoding candidate frames, running recurrent 3D reconstruction on them, and re-encoding the selected tokens into the KV cache with additional generator passes. A natural way to remove most of this cost is to train the generator jointly with a CUT3R-style recurrent reconstruction model (Wang et al., 2025b; Chen et al., 2026b).