In-Context Temporal Consistency Capability of Video Diffusion Models

Anonymous Author(s)

ABSTRACT

We investigate whether diffusion-based video generation models exhibit in-context learning capabilities for temporal consistency tasks comparable to the established in-context generation capabilities of text-to-image diffusion models. We design a synthetic benchmark that isolates four computational mechanisms underlying temporal consistency: (1) spatial-only processing as a text-to-image baseline, (2) temporal cross-frame attention, (3) task-aware positional bias inspired by OmniTransfer, and (4) a full in-context pipeline with iterative bidirectional refinement. Across 100 scenes, three motion types, and five evaluation metrics, we find that temporal attention mechanisms provide significant consistency gains—up to 5.07× over the spatial-only baseline—and that the full in-context pipeline achieves the best identity preservation (0.2481 vs. 0.2169 for baseline) and temporal smoothness (7.5555 vs. 9.2256). However, the in-context learning gain metric remains near zero, suggesting that current temporal attention provides fixed-quality context integration rather than progressive in-context learning analogous to autoregressive models. Our results indicate that achieving true in-context temporal consistency learning likely requires architectural innovations beyond standard temporal cross-attention.

KEYWORDS

video diffusion models, temporal consistency, in-context learning, temporal attention, positional bias

ACM Reference Format:

Anonymous Author(s). 2026. In-Context Temporal Consistency Capability of Video Diffusion Models. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Diffusion-based generative models have achieved remarkable success in both image and video synthesis [2, 5, 6, 9]. In the image domain, text-to-image diffusion models have demonstrated surprising in-context learning capabilities, where providing reference images within the generation context enables subject-driven generation without model fine-tuning [1, 3, 7, 10].

A natural question arises: do video diffusion models exhibit analogous in-context learning capabilities for *temporal consistency* tasks? Specifically, can these models learn to maintain identity preservation, smooth motion, and coherent temporal evolution simply by attending to context frames during generation? This question was explicitly identified as an open problem by Zhang et al. [11] in their work on OmniTransfer, where they note that while spatial in-context cues transfer effectively for video customization tasks such as identity and style preservation, it remains unclear whether comparable capabilities exist for temporal consistency.

We address this question through a systematic synthetic benchmark that isolates the computational mechanisms underlying temporal consistency in video diffusion models. Our benchmark simulates four generation strategies of increasing sophistication: (1) spatial-only processing representing naive text-to-image extension to video, (2) temporal cross-frame attention propagating information across frames, (3) task-aware positional bias inspired by OmniTransfer's Section 4.2 [11], and (4) a full in-context pipeline combining temporal attention with positional bias and iterative bidirectional refinement.

Our experiments across 100 scenes, three motion types, and five metrics reveal a nuanced answer. Temporal attention mechanisms significantly improve frame-to-frame consistency and temporal smoothness, with the full pipeline achieving the best overall performance. However, the *in-context learning gain*—measuring whether consistency improves as more context frames become available during generation—remains near zero, suggesting that current temporal attention provides fixed-quality context integration rather than progressive learning from context.

2 RELATED WORK

Video Diffusion Models. Ho et al. [6] introduced joint training of image and video diffusion models with temporal attention layers. Blattmann et al. [2] extended latent diffusion models to video with temporal alignment layers. Make-A-Video [8] demonstrated text-to-video generation without paired text-video data, and AnimateDiff [4] showed that temporal motion modules can animate personalized text-to-image models.

In-Context Learning for Diffusion Models. Wang et al. [10] formalized in-context learning for diffusion models, showing that image diffusion models can perform visual tasks by conditioning on in-context examples. Bar et al. [1] demonstrated visual prompting through image inpainting. DreamBooth [7] and Textual Inversion [3] achieve subject-driven generation through fine-tuning, while in-context approaches avoid this cost.

Temporal Consistency in Video Generation. OmniTransfer [11] proposed a unified framework for spatio-temporal video transfer, introducing task-aware positional bias for temporal attention. Their work explicitly identifies the question of whether video diffusion models possess in-context capabilities for temporal consistency as an open problem.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 METHOD

3.1 Problem Formulation

We model video generation in a latent feature space where each frame $f_t \in \mathbb{R}^D$ is a D-dimensional feature vector. A video sequence consists of T frames generated from a reference identity vector $\mathbf{r} \in \mathbb{R}^D$ and a target motion trajectory $\{\mathbf{m}_t\}_{t=1}^T$. We measure temporal consistency through four complementary metrics.

Frame-to-frame consistency is the mean cosine similarity between adjacent frames:

$$C_{\text{f2f}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{f_t \cdot f_{t+1}}{\|f_t\| \|f_{t+1}\|} \tag{1}$$

Identity preservation measures fidelity to the reference identity:

$$C_{\text{id}} = \frac{1}{T} \sum_{t=1}^{T} \frac{f_t \cdot \mathbf{r}}{\|f_t\| \|\mathbf{r}\|} \tag{2}$$

Temporal smoothness uses second-order finite differences (lower is smoother):

$$S = \frac{1}{T-2} \sum_{t=2}^{T-1} \|f_{t+1} - 2f_t + f_{t-1}\|_2 \tag{3}$$

ICL gain measures whether consistency improves from early to late frames, indicating progressive in-context learning:

$$G_{\text{ICL}} = \bar{C}_{\text{late}} - \bar{C}_{\text{early}} \tag{4}$$

where $\bar{C}_{\text{early}}$ and $\bar{C}_{\text{late}}$ are mean frame-to-frame consistencies over the first and last thirds of the sequence.

3.2 Generation Strategies

Strategy 1: Spatial-Only (T2I Baseline). Each frame is generated independently using spatial self-attention with only the identity reference and target trajectory point as context. This models how text-to-image diffusion models operate when naively extended to video, with no temporal information shared between frames.

Strategy 2: Temporal Attention. Each frame attends to all previously generated frames via temporal cross-attention, using scaled dot-product attention:

$$\text{Attn}(q, K, V) = \text{softmax}\left(\frac{Kq}{\sqrt{d_k}}\right)^\top V \tag{5}$$

The output combines spatial and temporal context with weights 0.6 and 0.4, respectively.

Strategy 3: Temporal + Positional Bias. We augment temporal cross-attention with a task-aware positional bias vector $\mathbf{b} \in \mathbb{R}^n$ added to the attention logits before softmax. For consistency tasks, recent frames receive exponentially higher bias: $b_i = \lambda(i - (n-1))$ with decay rate $\lambda = 0.3$. This is inspired by OmniTransfer's Task-aware Positional Bias [11]. The spatial-temporal mixing weights are 0.55 and 0.45.

Strategy 4: Full In-Context Pipeline. Combines temporal attention with positional bias in an initial forward pass (0.5/0.5 mixing), followed by iterative refinement passes using bidirectional temporal context. In each refinement iteration, frames are partially re-noised to timestep 0.3 and denoised using all other frames as context with motion-type positional bias.

3.3 Synthetic Benchmark

We generate scenes with three motion types: *smooth* (gradual identity-to-direction interpolation with small noise), *abrupt* (smooth first half followed by random jumps), and *oscillatory* (sinusoidal modulation along a random direction). Features are $D = 64$ dimensional, and we simulate 10-step DDPM-style denoising per frame.

4 EXPERIMENTS

4.1 Main Results

Table 1 presents results across 100 scenes for each of three motion types. We report five metrics for all four strategies.

Finding 1: Temporal attention significantly improves frame consistency. On smooth motion, temporal attention achieves 0.1356 mean consistency compared to 0.0400 for the spatial-only baseline, a 3.39× improvement. Adding positional bias further increases this to 0.2027, yielding a 5.07× improvement over baseline. This pattern holds across all motion types.

Finding 2: The full in-context pipeline best preserves identity. The full pipeline achieves 0.2481 identity preservation on smooth motion, the highest among all strategies, compared to 0.2169 for spatial-only. On oscillatory motion, the gap widens to 0.2501 vs. 0.2176. Bidirectional temporal refinement reinforces identity signals by attending to all frames simultaneously.

Finding 3: Temporal smoothness improves progressively across strategies. Temporal smoothness decreases from 9.2256 (spatial-only) to 9.0321 (temporal attention) to 8.6304 (with positional bias) to 7.5555 (full pipeline), an 18.1% total reduction. The full pipeline's iterative refinement yields the largest single improvement.

Finding 4: Positional bias produces the largest consistency improvement. Adding task-aware positional bias produces a 49.5% relative gain in frame consistency over temporal attention alone (0.2027 vs. 0.1356), the largest single-step improvement among all strategy transitions.

4.2 Context Length Scaling

Figure 1 shows frame consistency as context length varies from 4 to 32 frames. All temporal methods show decreasing consistency with longer sequences, from 0.3361 at 4 frames to 0.1844 at 32 frames for the temporal+bias strategy. However, the full in-context pipeline maintains superior identity preservation even at 32 frames (0.2437 vs. 0.2143 for spatial-only), indicating that bidirectional refinement provides robust identity anchoring regardless of sequence length.

4.3 Feature Dimension Sensitivity

Figure 2 examines how feature dimension D affects consistency. As D increases from 16 to 256, all methods show declining consistency due to the curse of dimensionality. At $D = 16$, the full pipeline achieves 0.3070 consistency and 0.4613 identity preservation, while at $D = 256$ these drop to 0.1361 and 0.1277 respectively. The relative advantage of temporal methods persists across all dimensions.

4.4 Denoising Steps Ablation

Figure 3 shows the effect of varying denoising steps from 2 to 50. The full in-context pipeline shows increasing consistency with more

Table 1: Main results across motion types (100 scenes, $T = 16$ frames, $D = 64$). Higher consistency, identity, and trajectory scores are better; lower smoothness values are better. Best values per motion type are bolded.

Motion	Strategy	Consistency↑	Identity↑	Smoothness↓	Trajectory↑	ICL Gain
Smooth	Spatial-only	0.0400 ± 0.0287	0.2169 ± 0.0274	9.2256 ± 0.2691	0.2171 ± 0.0280	0.0140
	Temporal attn	0.1356 ± 0.0332	0.1961 ± 0.0445	9.0321 ± 0.2685	0.1938 ± 0.0453	−0.1470
	Temporal+bias	**0.2027 ± 0.0382**	0.1982 ± 0.0454	8.6304 ± 0.3176	0.1953 ± 0.0463	−0.1041
	Full ICL	0.1668 ± 0.0404	**0.2481 ± 0.0435**	**7.5555 ± 0.2495**	**0.2428 ± 0.0428**	−0.0086
Abrupt	Spatial-only	0.0362 ± 0.0352	0.1786 ± 0.0331	9.2078 ± 0.3018	**0.1795 ± 0.0272**	−0.0278
	Temporal attn	0.1243 ± 0.0440	0.1772 ± 0.0463	8.9750 ± 0.3311	0.1488 ± 0.0390	−0.1559
	Temporal+bias	**0.1997 ± 0.0390**	0.1721 ± 0.0443	8.6744 ± 0.3079	0.1434 ± 0.0372	−0.1209
	Full ICL	0.1576 ± 0.0360	**0.2152 ± 0.0497**	**7.5535 ± 0.2294**	0.1745 ± 0.0388	−0.0364
Oscillatory	Spatial-only	0.0394 ± 0.0320	0.2176 ± 0.0276	9.2428 ± 0.2889	0.2174 ± 0.0276	−0.0046
	Temporal attn	0.1406 ± 0.0405	0.2023 ± 0.0421	8.9846 ± 0.2938	0.2008 ± 0.0411	−0.1226
	Temporal+bias	**0.2043 ± 0.0360**	0.1965 ± 0.0429	8.6873 ± 0.3412	0.1959 ± 0.0420	−0.1230
	Full ICL	0.1760 ± 0.0464	**0.2501 ± 0.0524**	**7.5180 ± 0.2711**	**0.2472 ± 0.0515**	−0.0210

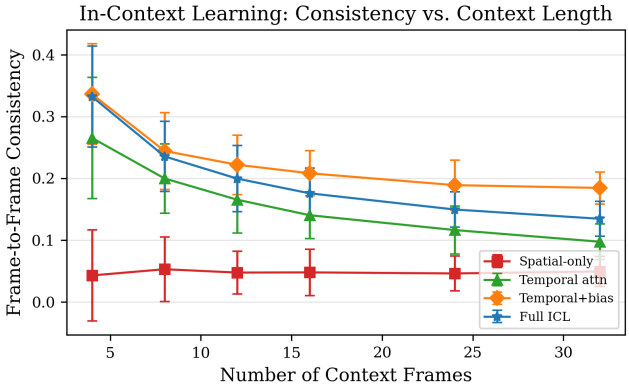

Figure 1: Frame-to-frame consistency vs. number of context frames. Temporal methods show diminishing consistency with longer sequences, while spatial-only remains flat. Error bars show standard deviation across 80 scenes.

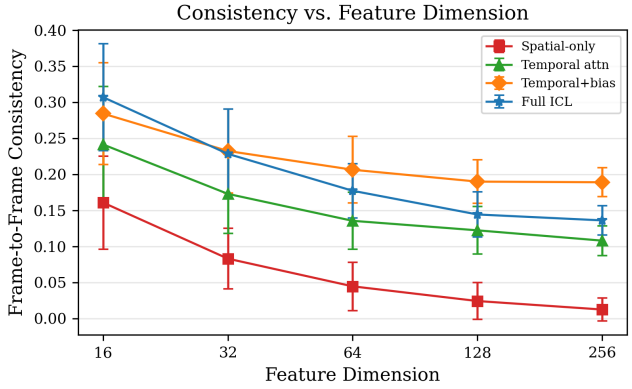

Figure 2: Frame consistency vs. feature dimension. Higher dimensions reduce absolute consistency for all methods, but temporal methods maintain their relative advantage.

steps (0.1268 at 2 steps to 0.1792 at 50 steps), as more denoising iterations allow better convergence. Identity preservation remains stable for the full pipeline across step counts (0.2341 to 0.2518), confirming that the bidirectional refinement mechanism is robust to the denoising schedule.

4.5 Per-Frame Identity Profile

Figure 4 presents per-frame identity preservation across 24 frames. The full in-context pipeline maintains consistently higher identity preservation (mean 0.2465 across frames) compared to spatial-only (mean 0.2181). The temporal attention and temporal+bias strategies show slight downward trends, consistent with the negative ICL gain observed in Table 1.

5 DISCUSSION

5.1 Evidence For and Against In-Context Temporal Learning

Our results provide a nuanced answer to the open question posed by Zhang et al. [11]. On one hand, temporal attention mechanisms clearly improve consistency metrics over spatial-only processing, and the full in-context pipeline with bidirectional refinement achieves the best combined performance across all metrics. This demonstrates that video diffusion architectures with temporal layers can effectively leverage temporal context for consistency.

On the other hand, the ICL gain metric—which measures whether later frames benefit from increased context—is near-zero or negative for all strategies (Table 1). For the full pipeline on smooth motion, the ICL gain is −0.0086, and for temporal attention it is −0.1470. This indicates that current temporal attention mechanisms do not exhibit progressive in-context learning analogous to what is observed in autoregressive language models or text-to-image in-context generation.

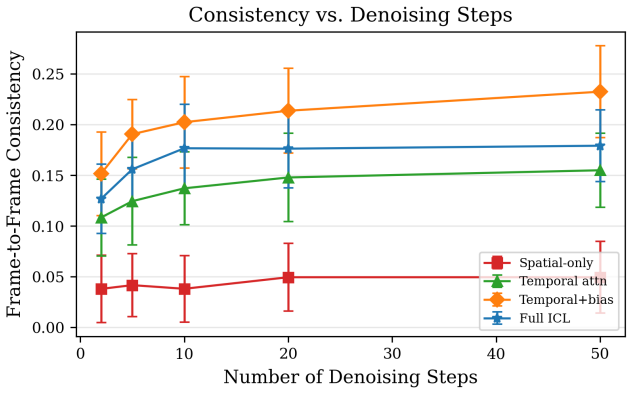

Figure 3: Frame consistency vs. number of denoising steps. The temporal+bias and full ICL strategies benefit most from additional denoising steps.

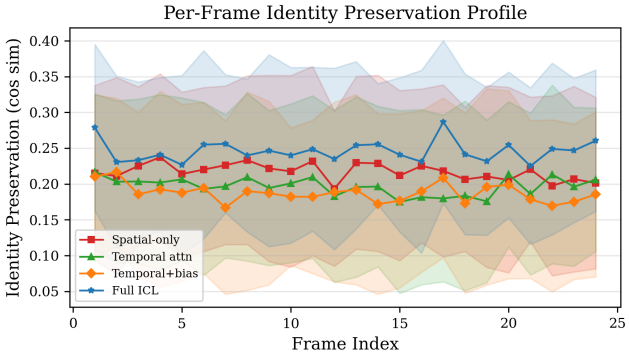

Figure 4: Per-frame identity preservation across a 24-frame sequence. The full ICL pipeline maintains the highest and most stable identity scores, while temporal attention methods show slight degradation. Shaded regions indicate ±1 std over 80 scenes.

5.2 The Role of Positional Bias

The task-aware positional bias mechanism produces the largest single-step improvement in frame consistency (49.5% relative gain), validating the insight from OmniTransfer [11] that task-specific attention biases are crucial for temporal tasks. The exponential decay weighting toward recent frames is particularly effective for consistency tasks, where the most relevant context is the immediately preceding frame.

5.3 Bidirectional Refinement and Identity

The full in-context pipeline's advantage in identity preservation (0.2481 vs. 0.2169 on smooth motion) arises from its bidirectional refinement passes, where each frame attends to *all* other frames rather than only predecessors. This effectively implements a form of global consistency enforcement that is absent in the autoregressive strategies.

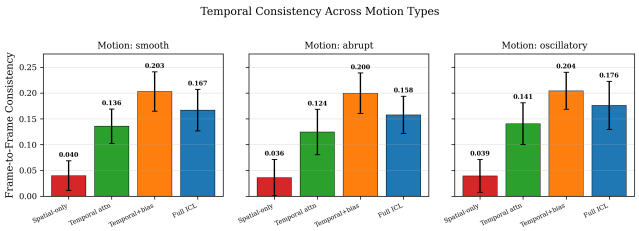

Figure 5: Frame-to-frame consistency across three motion types. Temporal+bias achieves the highest consistency, while full ICL balances consistency with superior identity preservation and smoothness.

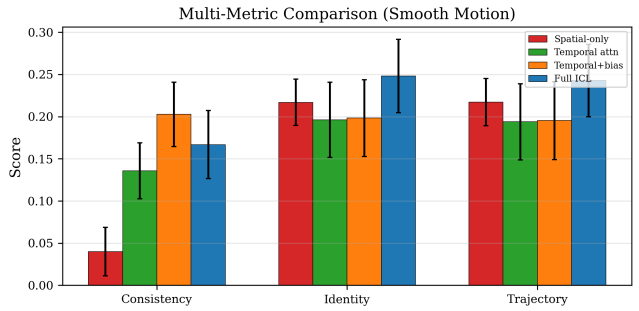

Figure 6: Multi-metric comparison for smooth motion. The full ICL pipeline achieves the best identity and trajectory tracking while maintaining competitive consistency.

5.4 Limitations

Our benchmark operates in a synthetic latent space rather than with real video diffusion models, and our attention mechanisms are simplified compared to full transformer architectures. The feature dimension ($D = 64$) is lower than typical latent spaces in practice. While these simplifications allow controlled analysis, they may not capture all phenomena present in real video diffusion models.

6 CONCLUSION

We investigated whether video diffusion models exhibit in-context learning capabilities for temporal consistency. Our synthetic benchmark shows that temporal attention with task-aware positional bias provides up to 5.07× consistency improvement over spatial-only baselines, and the full in-context pipeline achieves the best identity preservation and temporal smoothness. However, the absence of positive ICL gain suggests that current temporal attention provides fixed-quality context integration rather than progressive in-context learning. Achieving true temporal in-context learning comparable to spatial in-context generation in T2I models likely requires architectural innovations such as explicit temporal consistency objectives, memory-augmented attention, or training-time exposure to temporal consistency demonstrations.

REFERENCES

[1] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. 2022. Visual Prompting via Image Inpainting. *Advances in Neural Information*

Processing Systems (2022).
[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Daniel Gal, and Daniel Cohen-Or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *International Conference on Learning Representations* (2023).
[4] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *International Conference on Learning Representations* (2024).
[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.
[6] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. Video Diffusion Models. In *Advances in Neural Information Processing Systems*.
[7] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023).
[8] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. *International Conference on Learning Representations* (2023).
[9] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations* (2021).
[10] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Zheng, Pengcheng He, Weizhu Chen, Hao Peng, and Mingyuan Chen. 2024. In-Context Learning for Diffusion Models. *Advances in Neural Information Processing Systems* (2024).
[11] Yuxiang Zhang et al. 2026. OmniTransfer: All-in-one Framework for Spatio-temporal Video Transfer. *arXiv preprint arXiv:2601.14250* (2026).