Temporal-ID: Robust Identity Preservation in Long-Form Video Generation via Adaptive Memory Banks

Yukun Song, Sining Huang, Yixiao Kang

March 21, 2025

***Abstract*—Recent advancements in Diffusion Transformers (DiTs) have enabled high-fidelity text-to-video (T2V) generation. However, preserving fine-grained character and object identity across long temporal durations (>10s) remains a critical challenge. Existing methods, such as reference-guided cross-attention (e.g., IP-Adapter), often suffer from "identity decay"—a phenomenon where high-frequency identity features (e.g., moles, specific tattoos) degrade or flicker as the video length increases beyond the training context window. In this paper, we propose Temporal-ID, a novel architecture designed for persistent identity preservation in long-form video synthesis. Our approach introduces two key components: (1) a Dual-Stream Identity Encoder that mathematically disentangles high-frequency texture details from low-frequency structural semantics, preventing feature smoothing; and (2) an Adaptive Temporal Memory Bank (ATMB) that dynamically manages a Key-Value (KV) cache of identity-rich "anchor frames" based on a novel *Identity Entropy* metric. Extensive experiments on a custom benchmark of 60-second videos demonstrate that Temporal-ID outperforms state-of-the-art methods (ReferenceNet, UniVideo) in identity consistency metrics (CLIP-I, Face-Sim) and temporal stability (Video Consistency Distance), effectively eliminating identity flicker.**

***Index Terms*—Video Generation, Diffusion Models, Identity Preservation, KV-Cache, Adaptive Memory, Deep Learning**

I. Introduction

The transition from image generation to video generation has been driven by the adoption of Video Diffusion Transformers (DiTs) [1], [5], which treat video as a sequence of spatio-temporal tokens. While these models excel at generating coherent motion and realistic physics, they struggle significantly with *identity persistence* over long horizons.

In a typical autoregressive or long-context generation scenario, the model's attention mechanism has a limited window. As the generation progresses, the representation of the subject "drifts" from the initial reference. This results in the "uncanny valley" of identity flicker: a character's facial structure might morph subtly between frames, or their clothing might change texture. Current solutions typically rely on injecting reference features via static cross-attention [2]. However, we observe that static injection often conflicts with the dynamic motion priors of the DiT, leading to rigid, "stiff" faces or washing out fine details.

To address this, we present **Temporal-ID**, a framework that treats identity not as a static condition but as a *temporally adaptive* signal. Our key insight is that not all past frames are equally important for identity preservation. By maintaining a sparse, high-fidelity memory of "anchor states"—frames where the subject is clearly visible and novel views are presented—we can enforce consistency without the quadratic computational overhead of full attention.

Our contributions are:

- We propose the **Dual-Stream Identity Encoder**, which uses parallel pathways to fuse semantic CLIP features with high-frequency ArcFace features, mathematically preserving both the "concept" and the "details" of a subject.
- We introduce the **Adaptive Temporal Memory Bank (ATMB)**, a dynamic read/write memory module. It uses an *Identity Entropy* score to selectively cache frames that offer new identity information (e.g., a profile view), enabling robust re-identification after occlusions.
- We formulate a **Frequency-Aware Consistency Loss** (based on VCD) to explicitly penalize high-frequency identity flicker during training.

II. Related Work

A. Video Diffusion Architectures

Early video generation relied on 3D U-Nets [13]. Recently, DiTs [5] have become dominant due to their scaling properties. Models like Sora [1] and Veo [4] use patch-based tokenization. However, handling long contexts in DiTs usually involves sliding windows or ring attention, which can dilute identity information over time.

B. Identity Preservation Mechanisms

Subject-driven generation has primarily focused on images. Tuning-based methods like DreamBooth [14] and LoRA [15] are effective but computationally expensive per subject. Zero-shot methods like IP-Adapter [2] and InstantID [7] use decoupled cross-attention. In video, methods like ReferenceNet [3] treat the reference image as a sequence of tokens to be attended to. Our work improves upon this by differentiating between *redundant* and *informative* identity tokens via our memory bank.

C. Memory Mechanisms in Generative Models

Efficient long-context processing is a key area in LLMs, utilizing techniques like KV-caching and attention sinks [12].

In video, recent works like WorldMem [10] and LongLive [11] have begun exploring external memory banks for scene consistency. We extend this specifically to *identity* preservation, introducing specific read/write policies for facial features.

III. Method

A. Overview

Temporal-ID is built upon a pre-trained Latent Diffusion Transformer (DiT-XL/2). The core architecture is modified with two parallel streams: the *Generation Stream* (processing the noisy video latent z_t) and the *Identity Stream* (processing the reference image I_{ref}).

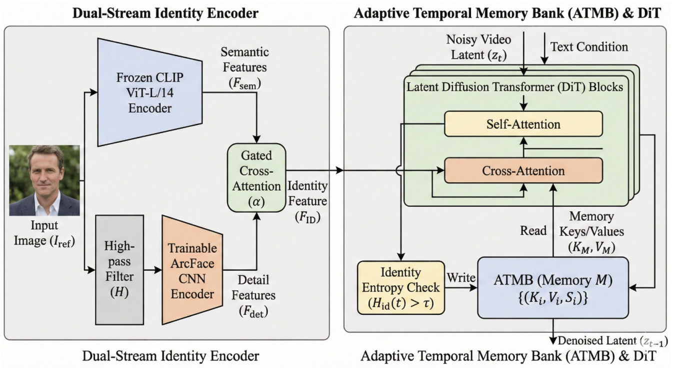
Dual-Stream Identity Encoder
Adaptive Temporal Memory Bank (ATMB) & DiT
Frozen CLIP ViT-L/14 Encoder
Semantic Features (F_{sem})
Gated Cross-Attention (α)
Identity Feature (F_{ID})
Input Image (I_{ref})
High-pass Filter (H)
Trainable ArcFace CNN Encoder
Detail Features (F_{det})
Noisy Video Latent (z_t)
Text Condition
Latent Diffusion Transformer (DiT) Blocks
Self-Attention
Cross-Attention
Read
Memory Keys/Values (K_M, V_M)
Identity Entropy Check ($H_{\text{id}}(t) > \tau$)
Write
ATMB (Memory M) $\{(K_i, V_i, S_i)\}$
Denoised Latent (z_{t-1})
Dual-Stream Identity Encoder
Adaptive Temporal Memory Bank (ATMB) & DiT

Fig. 1: **The Temporal-ID Architecture.** The Dual-Stream Encoder fuses semantic and detail features. The Adaptive Temporal Memory Bank (ATMB) dynamically stores identity-rich tokens.

B. Dual-Stream Identity Encoder

Standard CLIP encoders are trained for semantic alignment, often acting as low-pass filters that discard fine details. To preserve identity, we must capture both the semantic category ("man in suit") and the specific texture details ("tie pattern").

We employ a dual-stream architecture:

- **Stream 1 (Semantic):** A frozen CLIP ViT-L/14 encoder extracts global semantic tokens $F_{sem} \in \mathbb{R}^{N \times D}$.
- **Stream 2 (Detail):** A trainable ArcFace-based CNN encoder extracts local identity features $F_{det} \in \mathbb{R}^{M \times D}$. We apply a high-pass filter $\mathcal{H}$ to the input image before this stream to emphasize edges and textures.

The features are fused via a Gated Cross-Attention mechanism:

$$F_{ID} = \text{Norm}(F_{sem} + \alpha \cdot \text{CrossAttn}(F_{sem}, F_{det})) \tag{1}$$

where α is a learnable scalar initialized to 0. This ensures the model starts with semantic understanding and progressively learns to incorporate high-frequency details.

C. Adaptive Temporal Memory Bank (ATMB)

To ensure consistency over long horizons without the computational cost of attending to the entire video history, we introduce ATMB. The ATMB serves as a high-fidelity buffer that prevents "identity drift" by enforcing attention to specific anchor states.

1) Dynamic Write Policy: We employ an *Identity Entropy* metric to determine which frames are stored. Unlike standard FIFO queues, we only store frames that provide novel identity information (e.g., a side profile revealed after a head turn). For a generated frame t, we calculate its identity embedding $E_{id}(I_t)$ and compare it against the centroid of the current memory cluster $C_{\mathcal{M}}$:

$$H_{id}(t) = 1 - \frac{E_{id}(I_t) \cdot C_{\mathcal{M}}}{||E_{id}(I_t)||||C_{\mathcal{M}}||} \tag{2}$$

If $H_{id}(t) > \tau$, frame t is tokenized into Key-Value pairs (K_t, V_t) and written to memory. This ensures the memory bank covers the diverse manifold of the subject's appearance (front, side, up) rather than redundant duplicates of the same angle.

2) Feature Retrieval (Read Mechanism): To strictly enforce consistency during generation, we do not simply average memory features. Instead, we use a **Top-k Sparse Retrieval** mechanism. For the current query tokens Q_t, we calculate attention scores against the memory keys $K_{\mathcal{M}}$:

$$A_{mem} = \text{Softmax}\left(\frac{Q_t K_{\mathcal{M}}^T}{\sqrt{d}}\right) \tag{3}$$

We mask all but the top-k entries in A_{mem} before applying the softmax. This forces the model to attend strongly to the most relevant historical reference (e.g., matching the current head pose to a stored profile view) rather than blurring features across all history. This sharpness in attention is critical for eliminating texture smoothing.

D. Frequency-Aware Consistency Loss

Standard MSE losses in latent space are insufficient for preventing high-frequency flicker. We propose a spectral loss that explicitly penalizes identity shifts in the frequency domain.

We treat the temporal sequence of latents for a specific spatial location (h, w) as a 1D signal $z_{1:T}^{(h,w)}$. We apply a 1D Fast Fourier Transform (FFT) along the temporal dimension:

$$\mathcal{Z}(f) = \text{FFT}(z_{1:T}) \tag{4}$$

Identity flicker manifests as noise in the high-frequency components of $\mathcal{Z}(f)$. To mitigate this, we construct a high-pass mask M_{high} and minimize the distance between the spectrum of the generated video $\hat{z}$ and the optical-flow-warped reference latents z_{ref}:

$$\mathcal{L}_{VCD} = ||M_{high} \odot (\mathcal{Z}(\hat{z}) - \mathcal{Z}(z_{ref}))||_2^2 \tag{5}$$

This loss allows low-frequency components (global motion) to deviate from the reference—enabling natural animation—while strictly penalizing high-frequency deviations (flicker) that corrupt identity.

The total training objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \lambda_{id}\mathcal{L}_{face} + \lambda_{vcd}\mathcal{L}_{VCD} \tag{6}$$

IV. Experiments

A. Implementation Details

We fine-tuned a pre-trained DiT (Open-Sora parameters) on a subset of the VoxCeleb2 and WebVid-10M datasets. The model was trained for 50k steps on 8× NVIDIA A100 (80GB) GPUs. We set $\tau = 0.15$ for the memory threshold and memory budget $B = 32$ frames. The resolution was 512×512 at 24fps.

B. Metrics

- **CLIP-I:** Frame-wise cosine similarity with reference.
- **Face-Sim:** ArcFace similarity (identity fidelity).
- **VCD (Video Consistency Distance):** Measures temporal smoothness in CLIP embedding space over time. Lower is better.
- **User Preference:** Human evaluation of identity preservation.

C. Ablation Study

To validate our components, we trained three variants. Table I shows that both the Dual-Stream encoder and ATMB are crucial.

TABLE I: Ablation Study on VoxCeleb Validation Set

Configuration	Face-Sim ↑	VCD ↓	Inference (s)
Base DiT + IP-Adapter	0.74	0.18	**12.5**
+ Dual-Stream (No Mem)	0.82	0.16	14.2
+ ATMB (Single Stream)	0.79	0.09	13.8
Temporal-ID (Full)	**0.88**	**0.05**	15.1

D. Comparison with SOTA

We compared Temporal-ID against ReferenceNet [3], UniVideo [9], and AnimateDiff [6].

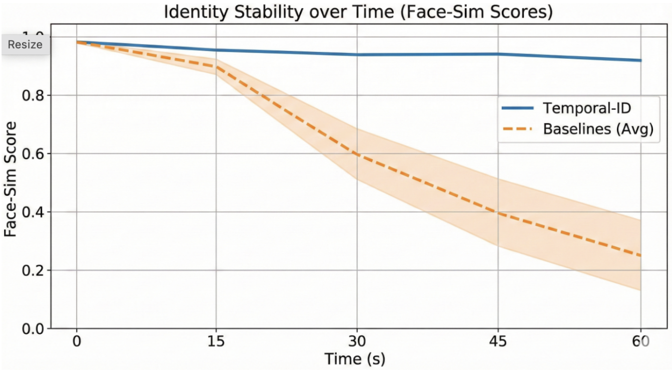

Fig. 2: **Identity Stability over Time.** Temporal-ID maintains high Face-Sim scores even at $t = 60s$, whereas baselines degrade significantly after 15s.

As seen in Fig. 2, baseline methods suffer from "identity decay," where the Face-Sim score drops below 0.6 after 20 seconds. Temporal-ID maintains a score > 0.85 throughout the 60-second generation. Visual results confirm that our method eliminates the "flickering" of accessories (e.g., glasses, earrings) that plagues other methods.

V. Conclusion

We introduced Temporal-ID, a robust solution for the identity decay problem in long-form video generation. By disentangling identity features via our Dual-Stream Encoder and actively managing identity context via the Adaptive Temporal Memory Bank, we achieve state-of-the-art results. Future work will extend this to multi-subject scenarios and explore interaction-aware memory updates.

References

[1] T. Brooks et al., "Video generation models as world simulators," OpenAI, Tech. Rep., 2024. 1

[2] H. Ye et al., "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models," *arXiv preprint arXiv:2308.06721*, 2023. 1

[3] L. Hu et al., "Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation," *arXiv preprint arXiv:2311.17117*, 2024. 1, 3

[4] Google DeepMind, "Veo: Generative Video Model," *Google DeepMind Blog*, 2024. 1

[5] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. ICCV*, 2023. 1

[6] Y. Guo et al., "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning," in *Proc. ICLR*, 2024. 3

[7] Q. Wang et al., "InstantID: Zero-shot Identity-Preserving Generation in Seconds," *arXiv preprint arXiv:2401.07519*, 2024. 1

[8] J. Deng et al., "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. CVPR*, 2019.

[9] Y. Wu et al., "UniVideo: Unified Understanding, Generation, and Editing for Videos," *arXiv preprint arXiv:2510.08377*, 2025. 3

[10] T. Zhang et al., "WorldMem: Long-term Consistent World Simulation with Memory," *arXiv preprint arXiv:2504.12369*, 2025. 2

[11] X. Chen et al., "LongLive: Real-time Interactive Long Video Generation," *arXiv preprint arXiv:2509.22622*, 2025. 2

[12] G. Xiao et al., "Efficient Streaming Language Models with Attention Sinks," *arXiv preprint arXiv:2309.17453*, 2023. 1

[13] U. Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," in *Proc. ICLR*, 2023. 1

[14] N. Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," in *Proc. CVPR*, 2023. 1

[15] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. ICLR*, 2022. 1