LaVieID: Local Autoregressive Diffusion Transformers for Identity-Preserving Video Creation

Wenhui Song*
Shenzhen Campus of Sun Yat-sen University
Shenzhen, China
wenhuisong16@gmail.com

Hanhui Li*
Shenzhen Campus of Sun Yat-sen University
Shenzhen, China
lihh77@mail.sysu.edu.cn

Jiehui Huang
Hong Kong University of Science and Technology (HKUST)
Hong Kong, China
jhhuang117@gmail.com

Panwen Hu
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
Panwen.Hu@mbzuai.ac.ae

Yuhao Cheng
Lenovo Research
Beijing, China
chengyh5@lenovo.com

Long Chen
Lenovo Research
Beijing, China
chenlong12@lenovo.com

Yiqiang Yan
Lenovo Research
Beijing, China
yanyq@lenovo.com

Xiaodan Liang†
Shenzhen Campus of Sun Yat-sen University
Shenzhen, China
liangxd9@mail.sysu.edu.cn

Abstract

In this paper, we present LaVieID, a novel local autoregressive video diffusion framework designed to tackle the challenging identity-preserving text-to-video task. The key idea of LaVieID is to mitigate the loss of identity information inherent in the stochastic global generation process of diffusion transformers (DiTs) from both spatial and temporal perspectives. Specifically, unlike the global and unstructured modeling of facial latent states in existing DiTs, LaVieID introduces a local router to explicitly represent latent states by weighted combinations of fine-grained local facial structures. This alleviates undesirable feature interference and encourages DiTs to capture distinctive facial characteristics. Furthermore, a temporal autoregressive module is integrated into LaVieID to refine denoised latent tokens before video decoding. This module divides latent tokens temporally into chunks, exploiting their long-range temporal dependencies to predict biases for rectifying tokens, thereby significantly enhancing inter-frame identity consistency. Consequently, LaVieID can generate high-fidelity personalized videos and achieve state-of-the-art performance. Our code and models are available at https://github.com/ssugarwh/LaVieID.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Video Synthesis, Diffusion Model, Spatio-temporal Consistency

ACM Reference Format:
Wenhui Song, Hanhui Li, Jiehui Huang, Panwen Hu, Yuhao Cheng, Long Chen, Yiqiang Yan, and Xiaodan Liang. 2025. LaVieID: Local Autoregressive Diffusion Transformers for Identity-Preserving Video Creation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754943

1 Introduction

In recent years, diffusion models [5] have demonstrated remarkable capabilities in modeling data of various modalities, such as texts [12], time series [51], images [46], 3D models [58] and videos [23]. These advances have significantly contributed to the development and application of content synthesis, especially for tasks such as text-to-image (T2I) [42] and text-to-video (T2V) [47] generation. However, the emerging task of identity-preserving T2V (IPT2V) generation conditioned on a single subject image [2, 17, 66], remains challenging for existing generative models. This is because it needs to explore large latent spaces and requires higher spatio-temporal consistency, compared with personalized T2I generation [15] or subject animation with reference motions [27, 56].

Several pioneering studies seek to address IPT2V generation by leveraging cutting-edge diffusion transformers (DiTs) [41] trained on large-scale data, and enhancing them with carefully designed modules and strategies, such as textual inversion [39], frequency decomposition [66], cross-video pairing [70], and reinforcement learning [34]. Although these methods improve spatio-temporal

*Both authors contributed equally to this research.
†Corresponding author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754943

Figure 1: Given a single reference image of a target person, the proposed LaVieID framework can follow diverse instructions to generate vivid videos with identity fidelity and visual quality superior to that of state-of-the-art methods [66].

consistency to some extent, they still suffer from the loss of identity information due to the intrinsic architecture designs of DiTs. Particularly, current DiTs [60] primarily rely on global attention modules to capture spatio-temporal feature correlations, which are relatively insensitive to fine-grained facial structures and temporal dependencies. Hence, there is a substantial risk that identity-related features may be disrupted by irrelevant information, yielding issues like identity incoherence, abrupt motions, and flickering frames.

Therefore, in order to address the aforementioned challenges, this paper proposes LaVieID, a local autoregressive video diffusion framework to tackle the above issues. LaVieID improves the identity preservation capabilities of DiTs from both spatial and temporal perspectives. First, LaVieID introduces a local router that leverages fine-grained facial structures, such as eyebrows, eyes, and mouth, to provide spatial guidance on modeling feature correlations. Specifically, the proposed local router learns to reconstruct and refine the latent tokens of subjects based on local facial structures, with adaptive weights that reflect the relative importance of these structures. In this way, latent states related to facial structures and subject identity are emphasized, while those that are irrelevant are suppressed, and consequently mitigate feature interference. Furthermore, a temporal autoregressive module is introduced to enhance temporal identity consistency. This module splits the denoised latent tokens of videos into multiple sequential chunks and explicitly establishes correlations between adjacent chunks, thus improving long-range temporal dependencies across frames. With the local router and the temporal autoregressive module, our LaVieID framework effectively overcomes the limitations of existing diffusion models in identity preserving, and achieves state-of-the-art performance in IPT2V generation, as shown in Figure 1.

In summary, LaVieID provides conventional DiTs with global and unstructured attention with the ability to exploit local facial structures and model long-range temporal dependencies, facilitating video customization with enhanced spatio-temporal identity consistency and visual quality. The main contributions of this paper are summarized as follows:

- We propose LaVieID, a novel framework that effectively addresses the identity-preserving text-to-video generation task.
- We introduce a local router to provide spatial structural guidance by leveraging fine-grained facial cues.
- We devise a temporal autoregressive module to model long-range temporal dependencies among frames.
- Comprehensive quantitative and qualitative analyses and a user study validate that LaVieID outperforms state-of-the-art methods.

2 Related Work

2.1 Generative Models

Diffusion models. Diffusion-based video generation models can be broadly categorized into UNet-based and transformer-based approaches. Among *UNet-based methods*, an early approach [23] extends the UNet architecture along the temporal dimension (using a dilated 3D UNet) to adapt it for video generation. Building on this, the latent video diffusion model [19] shifts from pixel space to latent space to represent video frames by a lower dimension. This effectively reduces computational and memory requirements, albeit at the cost of certain fine details. Recently, *transformer-based video diffusion models* have attracted considerable attention due to their scalability. Notably, Latte [37] introduces one of the first T2V diffusion models using a DiT architecture, which paves the way for a series of subsequent models, such as Sora [40], Open-Sora [32], and CogVideoX [60]. These models have achieved impressive progress in simulating the physical world. However, they still face severe challenges in maintaining identity consistency across frames.

Autoregressive models. The dominant paradigm of autoregression models is *Next-token prediction* that sequentially generates discrete tokens obtained via vector quantization methods [54]. Representative autoregression-based models include CogVideo [24] and

VideoPoet [30], and both of them employ transformer-based architectures to generate video tokens autoregressively conditioned on previous context tokens. Specifically, CogVideo integrates hierarchical and multi-frame-rate training to enhance temporal coherence aligned with textual prompts, while VideoPoet leverages multimodal inputs for flexible video synthesis. MAGVIT [64] employs masked token prediction techniques to accelerate inference speeds, and NOVA [7] uses continuous-valued tokens and intra-frame bidirectional modeling for efficient generation. Teller [69] introduces an autoregressive motion generation framework for real-time streaming portrait animation, driven by audio inputs. It leverages causal temporal attention and motion prediction priors to ensure responsiveness and consistency under streaming conditions. Meanwhile, Neighboring Autoregressive Modeling[18] proposes a local temporal conditioning scheme that only attends to nearby frame tokens, significantly improving short-term motion realism and reducing autoregressive redundancy in long sequences.However, the discreteness of tokens inherently causes spatio-temporal incoherence. More recently, a new *next-scale prediction* paradigm [36, 52] tries to handle this issue via predicting multi-scale latent representations of images, yet its employment to video generation remains unexplored. Several studies [11, 13, 16, 43] design strategies for improving temporal content consistency in the paradigm of frame interpolation or extrapolation. Nevertheless, the general performance of autoregression-based methods still requires further improvement to rival that of diffusion models.

Hybrid models. Hybrid models integrate autoregression methods with diffusion models to exploit the temporal consistency of autoregression approaches alongside the high-quality generation capabilities of diffusion models. For example, ACDiT [25] employs a block-wise conditional diffusion mechanism within an autoregression framework, preserving temporal coherence across video segments. MaskFlow [9] introduces discrete token-based flow matching within an autoregression setting, significantly accelerating video generation and ensuring frame continuity. CausVid [62] applies autoregression diffusion distillation, considerably reducing diffusion inference steps and enabling near real-time generation without compromising visual quality or temporal coherence. VideoWorld [44] integrates interactive prompting into an autoregression-diffusion hybrid framework, facilitating user-guided dynamic video generation. AR-Diffusion [50] introduces an asynchronous latent diffusion framework conditioned on past latent features in a chunk-wise autoregressive manner. By decoupling motion prediction from high-fidelity synthesis, it enables temporally coherent and efficient video generation with fewer sampling steps.Yet these methods are complex and inevitably modify the original latent space of diffusion models, which requires expensive computational costs and training data. Moreover, they still lack structured modeling strategies of spatio-temporal identity consistency, which is essential for the IPT2V task.

2.2 Personalized Content Synthesis

Personalized image synthesis. Recent advancements in image synthesis have significantly progressed towards producing high-quality, identity-consistent visuals. OneDiffusion [33] formulates personalized synthesis tasks within a unified sequential modeling framework, enabling tasks such as inpainting and novel view synthesis by treating various conditional images as diverse viewpoints of a single target. To further enhance the quality and personalization, InfiniteYou [29] introduces a multi-stage training approach incorporating supervised fine-tuning on synthesized personalized images, thus notably improving visual fidelity and identity consistency. ConsistentID [26] exploits local facial components to preserve identity information by inserting their visual features into text tokens. This strategy is effective in identity preservation, but it also yields text contamination and is insufficient for feature disentanglement. To address contamination of identity features, PuLID [15] leverages contrastive alignment to facilitate clean insertion of identity embeddings, mitigating undesirable feature blending.

Personalized video synthesis. Personalized video synthesis emphasizes both spatial and temporal identity consistency, thus facing additional challenges compared with personalized image synthesis. ConsisID [66] proposes to decompose facial information into low-frequency and high-frequency components, which are modeled by a keypoint-based global extractor and a local extractor supervised by face recognition, respectively. ID-Animator [17] addresses identity preservation in a zero-shot manner by leveraging latent facial queries and identity-oriented datasets without requiring task-specific fine-tuning. PersonVideo [35] uses multi-level personalized tokens to maintain identity across video sequences while allowing precise text following. Concat-ID [70] introduces a scalable framework that integrates variational autoencoders with video latents through 3D self-attention, leveraging cross-video pairing and multi-stage training to tackle diverse and complex video scenarios. Magic Mirror [67] integrates dual-branch facial feature extraction with conditioned adaptive normalization within a DiT, effectively balancing identity consistency and dynamic motion generation. Similarly, FantasyID [68] incorporates explicit 3D facial geometry priors and multi-view augmentation into diffusion transformers, enhancing the structural stability of facial identities. In parallel, approaches oriented toward animation tasks, such as AniPortrait [55], FaceShot [10], Hallo2 [6], and LivePortrait [14], have focused primarily on facial motion synchronization or expression control. Although these aforementioned methods substantially enhance the flexibility and quality of video generation, the global and unstructured feature correlations in their architectures still hinder their abilities in identity preservation.

3 Methodology

We present the details of LaVieID in this section. We begin by formulating the identity-preserving text-to-video (IPT2V) generation task and providing the overall framework of LaVieID in Sec. 3.1. We then introduce the two key components of LaVieID, including a local router (Sec. 3.2) and a temporal autoregressive module (Sec. 3.3) for enhancing spatial and temporal identity consistency, respectively. At last, the learning objectives of LaVieID are provided in Sec. 3.4.

3.1 Overall Framework

Task formulation. Given a reference image of a target person and a text prompt, our goal is to generate a corresponding video of the target person with spatially and temporally coherent identity.

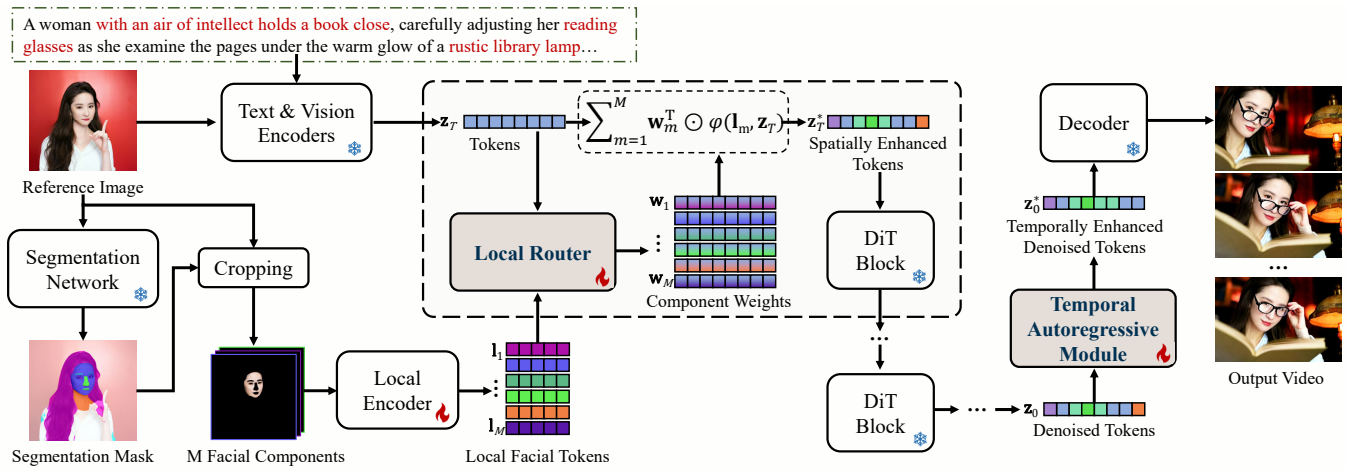

Figure 2: The proposed LaVieID framework. Two key components are introduced to improve the baseline DiT in spatio-temporal identity preservation: (i) A local router exploits local facial tokens to incorporate spatial structures into latent tokens; (ii) A temporal autoregressive module establishes long-range temporal dependencies in denoised tokens.

Formally, we formulate the video generation process by a latent diffusion model (LDM) [21, 60] guided by the following learning objective:

$$\mathcal{L}_{diff} = \mathbb{E}_{t,\epsilon,\mathbf{z}_0}||\epsilon - \epsilon_\theta(\sqrt{\bar{a}_t}\mathbf{z}_0 + \sqrt{1-\bar{a}_t}\epsilon, t)||^2, \tag{1}$$

where ϵ_θ denotes the LDM with trainable parameters θ. ϵ_θ can be interpreted as an iterative process consisting of T denoising steps that gradually restores the joint latent representation of the reference image and the text prompt $\mathbf{z}_0$ from a random noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. $t \in [1, T]$ denotes an arbitrary denoising step and $\bar{a}_t$ is a hyperparameter controlling the noise scale at the t-th step.

In this paper, we implement ϵ_θ based on a cutting-edge text-to-video model [60], in which multiple DiT blocks with 3D hybrid full attention are used to model global spatio-temporal contexts. However, these attention modules disregard facial structures and calculate correlations across all latent tokens, which are prone to feature interference and loss of target identity. Hence, we develop the following LaVieID framework to address these limitations.

LaVieID framework. The framework of LaVieID is shown in Figure 2. In parallel with the text and vision encoders of the baseline model, an off-the-shelf segmentation network [63] is employed to extract M local facial components (such as eyes, nose, and lips) from the target image. Subsequently, a local encoder embeds these facial components into M corresponding token sequences. As the subject structures in synthesized videos are mainly influenced by front DiT blocks [4], we propose to inject the facial token sequences into the first DiT block of the baseline model via a local router. This router is designed to model the correlations between the joint latent tokens and the local facial token sequences, so that the facial structures can be leveraged adaptively to refine the joint latent tokens and better preserve identity information. The latent tokens processed by all DiT blocks after T iterations are noise-free. Before decoding them into the output video, we further improve their temporal consistency via an autoregressive module. This module splits the denoised tokens into multiple chunks temporally (i.e., along the frame axis), and progressively refines the tokens in one chunk by predicting their biases conditioned on those in the preceding chunk. This ensures explicit long-range temporal dependencies among the denoised tokens, thus overcoming the limitations of the temporally independent and unstructured denoising process of the baseline model. Consequently, the proposed LaVieID framework is capable of generating latent representations with high spatio-temporal identity consistency to address the IPT2V task.

3.2 Local Router

The proposed local router is devised to disentangle global facial information into fine-grained local representations, thereby enhancing spatial identity consistency. The pioneering research on personalized image synthesis [26] also proposes to exploit local facial components. Nevertheless, it simply incorporates component-level features into text tokens, which contaminates textual information and hinders the ability of its model in feature disentanglement. In contrast, our local router estimates and modulates the influences of local facial components on the joint latent tokens, thus it can emphasize identity-related tokens while suppressing those that are irrelevant to better preserve identity information.

Specifically, let $\mathbf{l}_m \in \mathbb{R}^{L \times D}$ denote one of the local facial token sequences, where L is the number of tokens, D is the feature dimension, and $m \in [1, M]$. For the conciseness of presentation, we omit the index of denoising time step and denote the joint latent tokens as $\mathbf{z} \in \mathbb{R}^{L' \times D'}$. The proposed local router calculates the token-wise weights for each component $\mathbf{w}_m \in [0, 1]^{1 \times L'}$ as follows:

$$\mathbf{w}_m = \sigma(\mathbf{W}_m \tilde{\mathbf{l}_m} \mathbf{W}_l \mathbf{W}_z^{\mathrm{T}} \tilde{\mathbf{z}}^{\mathrm{T}}), \tag{2}$$

where $\tilde{\mathbf{l}_m}$ and $\tilde{\mathbf{z}}$ are $\mathbf{l}_m$ and $\mathbf{z}$ after layer normalization [1]. $\mathbf{W}_l \in \mathbb{R}^{D \times D''}$ and $\mathbf{W}_z \in \mathbb{R}^{D' \times D''}$ are two linear transformations that project $\mathbf{l}_m$ and $\mathbf{z}$ into the same latent space. T represents the matrix transpose operator. It is easy to see that $\tilde{\mathbf{l}_m} \mathbf{W}_l \mathbf{W}_z^{\mathrm{T}} \tilde{\mathbf{z}}^{\mathrm{T}}$ actually models the correlations among all tokens in $\mathbf{l}_m$ and $\mathbf{z}$, while $\mathbf{W}_m \in \mathbb{R}^{1 \times L}$ is used to aggregate the correlations w.r.t. each token in $\mathbf{z}$. σ denotes

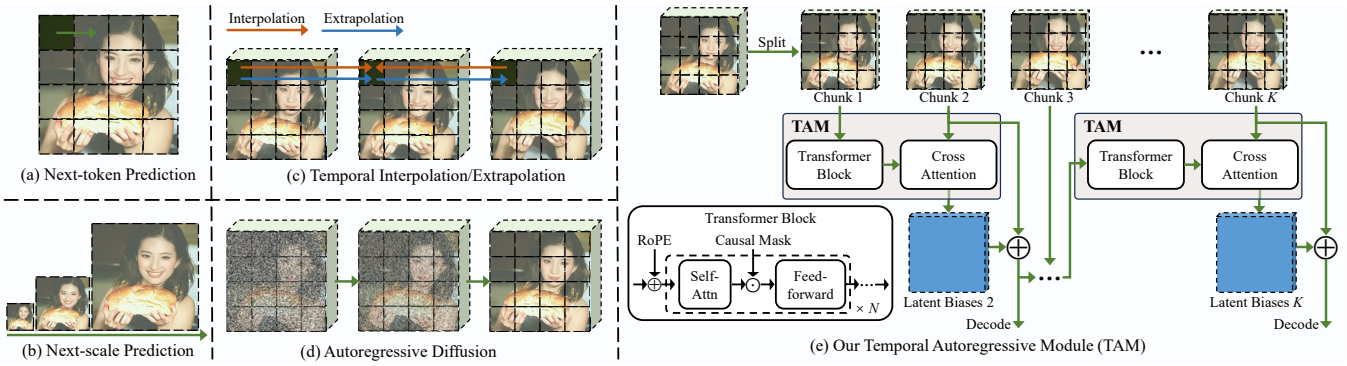

Figure 3: The proposed temporal autoregressive module against other autoregression-based mechanisms. Existing autoregression methods rely on predicting (a) discrete tokens [31, 65], (b) multi-scale tokens [36, 52], or (c) inter-frame tokens [11, 16], of which the visual quality remains inferior to that of diffusion models. (d) Hybrid models [3, 23, 38, 45] leverage the advantages of diffusion models in visual quality and autoregression models in temporal modeling. Yet their schemes are complicated and require intensive resources. (e) Our module models long-range temporal dependencies across frames on the chunk-wise level. Note that it is applied on denoised tokens, and hence it will not alter the latent space of the baseline DiT while enjoying the advantages of both diffusion models and autoregression models efficiently.

the softmax function, which is used to normalize $\mathbf{w}$ across all M facial components.

Subsequently, we leverage the component weights to refine $\mathbf{z}$ and improve spatial consistency as follows:

$$\mathbf{z}^* = \mathbf{z} + \alpha \sum_{m=1}^{M} \mathbf{w}_m^{\mathrm{T}} \odot \varphi(\mathbf{l}_m, \mathbf{z}). \tag{3}$$

Here, $\varphi : \mathbb{R}^{L\times D} \times \mathbb{R}^{L'\times D'} \rightarrow \mathbb{R}^{L'\times D'}$ is a function that reconstructs $\mathbf{z}$ using the tokens in $\mathbf{l}_m$ as the bases, akin to the cross-attention mechanism for feature interactions [15], where $\mathbf{l}_m$ serves as both the keys and the values, and $\mathbf{z}$ are considered as the queries. $\odot$ denotes the (broadcastable) element-wise product. α is a hyperparameter controlling the scale of the spatial enhancement. Based on Eqs. (2) and (3), we can expect that a properly optimized router will assign higher weights to identity-related and discriminative tokens, thereby facilitating effective identity preservation.

As for the local encoder, we adopt the architecture proposed in [66], which was originally designed for extracting high-frequency patterns, such as contours. However, as the goal of our local encoder is to embed local facial structures, we fine-tune it alongside the local router and the temporal autoregressive module during training our model, rather than keeping it fixed.

3.3 Temporal Autoregressive Module

In addition to the local router for improving frame-wise identity consistency, we introduce the temporal autoregressive module, which aims at cross-frame identity consistency to ensure overall video quality. The core idea of the proposed autoregressive module is to incorporate explicit temporal dependencies into latent tokens to alleviate abrupt changes in appearance and motion.

Figure 3 demonstrates the architecture of the proposed autoregressive module. This module takes the denoised latent tokens prior to decoding as input, splits them along the frame axis into multiple chunks, and predicts their corresponding biases sequentially for temporal refinement. Moreover, at the beginning of each chunk, the last enhanced frame from the preceding chunk is inserted, thereby forming a "teacher forcing" paradigm [57] that prevents the enhanced tokens from deviating excessively from the original content. Assume the shape of denoised latent tokens $\mathbf{z}_0$ remains $L' \times D'$ and the number of chunks is K, we denote an arbitrary chunk as $\mathbf{c}_k \in \mathbb{R}^{(1+\frac{L'}{K})\times D'}$, $k \in [1, K]$. The temporally enhanced chunk $\mathbf{c}_k^*$ is obtained as follows:

$$\mathbf{c}_k^* = \mathbf{c}_k + \beta \mathbf{b}_k, \quad \mathbf{b}_k = \varphi(\psi(\mathbf{c}_{k-1}^*), \mathbf{c}_k), \tag{4}$$

where $\mathbf{b}_k \in \mathbb{R}^{(1+\frac{L'}{K})\times D'}$ denotes the predicted latent biases. φ is a cross-attention module with the same architecture used in the local router. ψ is an efficient transformer block that alternately stacks N multi-head self-attention modules and feedforward layers. We employ two techniques within ψ to facilitate temporal context modeling. First, we add the input tokens of ψ with the rotary position embeddings (RoPE) [49] corresponding to their frame indexes to incorporate temporal information. Second, to impose the temporal order constraint on the frames, we employ causal masking [59] on the outputs of the self-attention modules. Consequently, ψ can capture the temporal contexts in $\mathbf{c}_{k-1}^*$ and interact with $\mathbf{c}_k$ via φ, similar to the spatial enhancement paradigm proposed in Eq. (3). β is a hyperparameter for weighting the biases.

Discussion. Figure 3 summarizes the differences between the proposed autoregressive module with other autoregression methods employed in mainstream generative models. Compared with existing approaches, the advantages of our module are threefold: (i) Our module can fully leverage the superior generation abilities of well-developed diffusion models, compared with methods that merely rely on autoregression [31, 36, 52, 65]. (ii) Instead of directly predicting latent tokens, our method predicts biases. This is more stable compared to [11, 16] because the latent tokens initially generated by the baseline LDM serve as a favorable starting point and help to address the accumulation of deviations. (iii) In contrast to

hybrid models [3, 23, 38, 45], our module is employed as a plug-in appended to the end of the denoising process. This allows TAM to maintain the original latent space of the base model and improve the temporal consistency efficiently and allows us to train it efficiently (with only one GPU). Besides, temporal correlations in TAM are calculated on the chunk level instead of frame-by-frame, which alleviates drastic latent changes in noisy frames and improves the temporal stability of our synthesized videos.

3.4 Network Optimization

Besides the standard diffusion loss $\mathcal{L}_{diff}$ defined in Eq. (1), we propose to use the cross-entropy loss between the output logits of the local router and the ground-truth face segmentation masks for training, namely,

$$\mathcal{L}_{route} = -\sum_{m=1}^{M} \mathbf{y}_m \odot \log \mathbf{w}_m. \tag{5}$$

Here $\mathbf{y}_m \in \{0,1\}^{1\times L'}$ is the ground-truth, where an element equals 1 means that the corresponding token belongs to the m-th component, and 0 otherwise. Hence, minimizing $\mathcal{L}_{route}$ encourages the router to recognize local facial structures and facilitate identity preservation.

The total objective function for training our LaVieID framework is given as follows,

$$\mathcal{L}_{\text{total}} = \lambda_{\text{diff}}\,\mathcal{L}_{\text{diff}} \;+\; \lambda_{\text{route}}\,\mathcal{L}_{\text{route}}, \tag{6}$$

where λ_{diff} and λ_{route} are hyperparameters balancing overall video quality and identity preservation.

4 Experiment

In this section, we validate the proposed LaVieID framework via extensive experiments. Interested readers can refer to our project website for more results and details.

4.1 Setup

Implementation details. All our experiments are conducted on a single A100 GPU. We use $M = 6$ facial component classes, including eyebrows, eyes, mouth, nose, facial skin region, and hair. For local facial token sequences, the number of tokens L is set to 32 and the feature dimension D is set to 2048. The length and feature dimension of latent tokens,L' and D', are set to 17750 and 3072. The inner feature dimension in the local router $D'' = 2048$ and the transformer block in the temporal autoregressive module contains $N = 6$ layers. The number of chunks K is set to 4. The hyperparameters controlling spatial and temporal enhancement, α and β, are set to 1 and 0.2, respectively. The trade-off weights between $\mathcal{L}_{diff}$ and $\mathcal{L}_{route}$, are set as $\lambda_{\text{diff}} = \lambda_{\text{route}} = 1$. We adopt AdamW as the optimizer and train our models with a batch size of 1 and a learning rate of 3×10^{-6} for 10K steps, which takes approximately 60 hours. We employ classifier-free guidance (CFG) [22] with random null text prompts at a ratio of 0.1. We use DDIM [48] for inference with $T = 50$ denoising steps and a CFG scale of 6.

Datasets and Metrics. We follow the experiment settings of [66] and use its dataset for training and evaluation, including approximately 30K short video clips for training; 30 different subjects disjoint from the training set, with about 5 reference images per subject, and 90 curated prompts for evaluation. We resize all videos to 720×480 with the frame rate of 8 frames per second and 49 frames per video. To evaluate the proposed method comprehensively, we adopt multiple widely-used metrics, including the subject/background consistency from [28] that measures temporal quality, FID [20] that evaluates the reality of synthesized videos, and FaceSim-Curricular/ArcFace from [8] that calculate the ID similarity between reference images and generated videos.

4.2 Comparison with State-of-the-arts

We compare LaVieID with four open-source state-of-the-art IPT2V methods, including ConsisID [66], ID-animator [17], CogvideoX+IPA (i.e., CogvideoX [60] with IP-Adapter [61]), and Concat-ID [70] using the same evaluation protocol.

Quantitative analysis. Table 1 summarizes the performance of LaVieID and the baselines. LaVieID outperforms all competitors on identity-centered metrics, including FaceSim-Curricular and FaceSim-ArcFace. It also excels at temporal quality and obtains the highest subject and background consistency values. As for the FID metric, LaVieID is the second best but still slightly worse than CogvideX+IPA (174.121 vs. 167.117). This is because the IPT2V task does not have matched training data and hence can be considered as an out-of-distribution generalization task per se. In this context, CogvideX+IPA is less generalizable, and its results are more likely to follow the distributions of its original large-scale training data. Consequently, its synthesized videos seem more realistic. This can be validated by its FaceSim-Curricular and FaceSim-Arcface scores (0.105 and 0.074), which are much worse than those of LaVieID (0.425 and 0.401). These scores indicate that CogVidex+IPA is less affected by target subjects and is confined to in-domain predictions. On the contrary, LaVieID better balances video quality and identity preservation, obtaining competitive performance on all five metrics.

Qualitative analysis. Figure 4 provides the visual comparison of video examples generated by different methods. Compared to the baselines, LaVieID produces identity-faithful content of satisfactory quality, even with complex conditions involving diverse poses, motions, expressions, and scenarios. In contrast, ID-animator fails to maintain coherent face representations, and occasionally cannot focus on the subject (e.g., in the first example). ConsisID, while exhibiting relatively stable identity, suffers from poor text alignment in some cases, such as missing the scene description in the second example (e.g., "golden autumn trees" and "leaves"). Moreover, we observe that it changes some facial attributes accidentally. For example, in Figures 1 (the second example) and 4 (the third example), it mistakenly turns both the rather "young" subjects into elders by adding lots of wrinkles. The qualitative examples of CogVideoX+IPA support our opinion that it is ineffective in out-of-distribution generalization and preserving target identity. In summary, the quantitative results of LaVieID align well with its qualitative results, suggesting it is a considerable solution for IPT2V generation.

4.3 Ablation Studies

To validate the effectiveness and contributions of each proposed component in LaVieID, we conduct a series of ablation studies. Limited by our computational resources, we select a subset of the

Figure 4: Visual comparisons of the proposed LaVieID against state-of-the-art methods.

Table 1: Quantitative comparisons of the proposed LaVieID framework against state-of-the-art methods.

Method	Subject Consistency ↑	Background Consistency ↑	FaceSim Curricular ↑	FaceSim ArcFace ↑	FID ↓	Human Evaluation			
						VQ ↑	TA ↑	DD ↑	IDS ↑
ID-animator [17]	0.691	0.704	0.052	0.054	201.33	0.148	0.154	0.134	0.049
Concat-ID [70]	0.708	0.714	0.291	0.278	204.465	0.160	0.166	0.150	0.186
CogvideoX+IPA [60]	0.743	0.768	0.105	0.074	**167.117**	0.212	0.188	0.227	0.106
ConsisID [66]	0.731	0.765	0.412	0.389	178.447	0.131	0.144	0.163	0.155
LaVieID (Ours)	**0.747**	**0.773**	**0.425**	**0.401**	174.121	**0.350**	**0.348**	**0.326**	**0.504**

Table 2: Ablation study on the proposed components.

Module	Subject Consistency ↑	Background Consistency ↑	FaceSim Curricular ↑	FaceSim ArcFace ↑	FID ↓
Baseline	0.672	0.736	0.312	0.291	203.04
w/ LR	0.740	0.742	0.361	0.334	178.663
w/ TAM	0.725	0.756	0.358	0.327	186.374
LaVieID	**0.755**	**0.775**	**0.428**	**0.398**	**168.714**

test dataset for evaluation in this section. More ablation study results can be found in our project website.

To evaluate the contribution of each key component in LaVieID, we constructed two variants for assessment: one combining the baseline DiT with the local router (denoted as w/ LR) and the other with the temporal autoregressive module (denoted as w/ TAM). The quantitative results of these two variants are reported in Table 2, from which we can observe that both components yield considerable performance gains in all metrics. It should be noted that the objectives of these components differ: the local router aims at spatial identity consistency, whereas the temporal autoregressive module, as its name suggests, is designed for improving temporal identity consistency. Hence, it is reasonable that the baseline w/ LR outperforms its counterpart w/ TAM on identity-centered metrics (FaceSim-Curricular/ArcFace). This can also be supported by a more intuitive comparison based on temporal quality metrics. Specifically, while the baseline w/ LR exhibits higher subject consistency, the baseline w/ TAM achieves superior performance in background consistency. These results suggest that the temporal autoregressive module tends to improve the overall temporal consistency, while the local router focuses more on the subject. Therefore, we can conclude that both components effectively fulfill their intended purposes, and their integration within the LaVieID framework yields the best overall performance.

The visual comparison between the above variants of our method is shown in Figure 5. Compared with the visual examples generated by the baseline w/ LR, those by the baseline w/ TAM exhibit a few identity distortions. For instance, the hairs of both subjects are changed by the latter variant. This is in accordance with our above discussion that the identity consistency benefits more from the local router, while the temporal consistency is improved by the proposed autoregressive module. Another interesting phenomenon is that the baseline w/ TAM tends to generate videos with smaller yet more dynamic subjects, which may also be the reason for its lower performance in subject consistency compared with that of

Figure 5: Visual ablation studies on the proposed components.

the baseline w/ LR. On the other hand, this also means that the temporal autoregressive module facilitates more diverse and smoother motions. For example, in the left column of Figure 5, the results of the baseline and the variant w/ LR do not exhibit the motion of "softly gazes downward", which is achieved by the baseline w/ TAM successfully. This suggests a potential solution to alleviate the limitation of IPT2V models in motion control. That is, the precise generation of subject motions merely by texts is possible, if we can construct the long-range and appropriate temporal correlations between the text descriptions of motions and latent tokens. Finally, LaVieID with the full architecture overcomes the limitations of both the variants and achieves the best visual results.

4.4 User Study

In the field of video synthesis, whether manually crafted metrics can be fully aligned with human preferences is a controversial and unsolved issue [28, 53]. Therefore, we further conduct a user study to provide a subjective evaluation of different methods. Specifically, we assess the quality of generated videos across four dimensions, including visual quality (VQ), text alignment (TA), dynamic degree (DD), and identity similarity (IDS), to provide a comprehensive performance evaluation from the human perspective. We design a questionnaire of 50 questions and invite 30 participants. In each question, we randomly show the synthesized videos of LaVieID and the baselines, and ask the participants to choose the best one based on the above four criteria.

The results of this user study are reported in Table 1, from which we can see that the synthesized videos of LaVieID are favored by most participants across all four criteria. Particularly, more than half of the participants agree that LavieID achieves the best identity-preserving results, which we owe to the proposed local router and the temporal autoregressive module for enhancing spatio-temporal identity consistency.

5 Conclusion

This paper introduces a spatio-temporal identity-enhancing framework named LaVieID to complete personalized text-to-video generation. The proposed approach tackles the intrinsic limitations of existing diffusion transformers with two carefully devised components, including a local router and a temporal autoregressive module. The local router dynamically emphasizes fine-grained facial structures to refine latent tokens. The temporal autoregressive module ensures temporal coherence through explicit modeling of long-range dependencies. Extensive quantitative and qualitative experiments and a user study validate that LaVieID significantly surpasses state-of-the-art methods, particularly in terms of identity consistency and visual quality.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant No. 62372482 and Shenzhen Science and Technology Program under Grant No. GJHZ20220913142600001.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[2] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. 2024. Still-moving: Customized video generation without customized video data. *ACM Transactions on Graphics* 43, 6 (2024), 1–11.

[3] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. 2024. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems* 37 (2024), 24081–24125.

[4] Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. 2024. $Delta$-DiT: A Training-Free Acceleration Method Tailored for Diffusion Transformers. *arXiv preprint arXiv:2406.01125* (2024).

[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10850–10869.

[6] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. 2024. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718* (2024).

[7] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. 2024. Autoregressive Video Generation without Vector Quantization. *arXiv preprint arXiv:2412.14169* (2024).

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[9] Michael Fuest, Vincent Tao Hu, and Björn Ommer. 2025. MaskFlow: Discrete Flows For Flexible and Efficient Long Video Generation. *arXiv preprint arXiv:2502.11234* (2025).

[10] Junyao Gao, SUN Yanan, Fei Shen, Xin Jiang, Zhening Xing, Kai Chen, and Cairong Zhao. [n. d.]. FaceShot: Bring Any Character into Life. In *The Thirteenth International Conference on Learning Representations*.

[11] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. 2022. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *Proceedings of the European Conference on Computer Vision*. 102–118.

[12] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2022. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *International Conference on Learning Representations*.

[13] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. 2025. Long-Context Autoregressive Video Modeling with Next-Frame Prediction. *arXiv preprint arXiv:2503.19325* (2025).

[14] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. 2024. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168* (2024).

[15] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. 2024. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in Neural Information Processing Systems* 37 (2024), 36777–36804.

[16] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. 2022. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3615–3625.

[17] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. 2024. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275* (2024).

[18] Yefei He, Yuanyu He, Shaoxuan He, Feng Chen, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. 2025. Neighboring autoregressive modeling for efficient visual generation. *arXiv preprint arXiv:2503.10696* (2025).

[19] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221* (2022).

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[22] Jonathan Ho and Tim Salimans. [n. d.]. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

[23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.

[24] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022).

[25] Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. 2024. ACDiT: Interpolating Autoregressive Conditional Modeling and Diffusion Transformer. *arXiv preprint arXiv:2412.07720* (2024).

[26] Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, et al. 2024. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771* (2024).

[27] Jiancheng Huang, Mingfu Yan, Songyan Chen, Yi Huang, and Shifeng Chen. 2024. MagicFight: Personalized Martial Arts Combat Video Generation. In *Proceedings of the ACM International Conference on Multimedia*. 10833–10842.

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

[29] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. 2025. InfiniteYou: Flexible Photo Recrafting While Preserving Your Identity. *arXiv preprint* arXiv:2503.16418 (2025).

[30] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125* (2023).

[31] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming Chang Chiu, et al. 2024. VideoPoet: A Large Language Model for Zero-Shot Video Generation. *International Conference on Machine Learning* 235 (2024), 25105–25124.

[32] PKU-Yuan Lab, Tuzhan AI, et al. 2024. Open-sora-plan.

[33] Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. 2025. One Diffusion to Generate Them All. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[34] Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. 2025. MagicID: Hybrid Preference Optimization for ID-Consistent and Dynamic-Preserved Video Customization. *arXiv preprint arXiv:2503.12689* (2025).

[35] Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. 2024. PersonalVideo: High ID-Fidelity Video Customization without Dynamic and Semantic Degradation. *arXiv preprint arXiv:2411.17048* (2024).

[36] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. 2024. ControlVAR: Exploring Controllable Visual Autoregressive Modeling. *arXiv preprint arXiv:2406.09750* (2024).

[37] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2024. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048* (2024).

[38] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. 2025. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2025).

[39] Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. 2024. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368* (2024).

[40] OpenAI. 2023. Video generation models as world simulators.

[41] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE International Conference on Computer Vision*. 4195–4205.

[42] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. 8821–8831.

[43] Shuhuai Ren, Shuming Ma, Xu Sun, and Furu Wei. 2025. Next Block Prediction: Video Generation via Semi-Auto-Regressive Modeling. *arXiv preprint arXiv:2502.07737* (2025).

[44] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. 2025. VideoWorld: Exploring Knowledge Learning from Unlabeled Videos. *arXiv preprint arXiv:2501.09781* (2025).

[45] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. 2024. Rolling diffusion models. In *International Conference on Machine Learning*. 42818–42835.

[46] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *Proceedings of the ACM SIGGRAPH Conference*. 1–10.

[47] 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. In *International Conference on Learning Representations*.

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

[49] J Su, H Zhang, X Li, J Zhang, and Y RoFormer Li. 2021. Enhanced transformer with rotary position embedding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*. 1–6.

[50] Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, SiYu Zhou, Qian He, and Jing Liu. 2025. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7364–7373.

[51] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34 (2021), 24804–24816.

[52] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in Neural Information Processing Systems* 37 (2024), 84839–84865.

[53] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9568–9578.

[54] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in Neural Information Processing Systems* 30 (2017).

[55] Huawei Wei, Zejun Yang, and Zhisheng Wang. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694* (2024).

[56] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. 2024. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6537–6549.

[57] Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1, 2 (1989), 270–280.

[58] Zhenyu Xie, Haoye Dong, Yufei Gao, Zehua Ma, and Xiaodan Liang. 2024. DreamVTON: Customizing 3D Virtual Try-on with Personalized Diffusion Models. In *Proceedings of the ACM International Conference on Multimedia*. 10784–10793.

[59] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157* (2021).

[60] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).

[61] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).

[62] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. 2024. From slow bidirectional to fast autoregressive video diffusion models. *arXiv preprint arXiv:2412.07772* 2 (2024).

[63] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*. 325–341.

[64] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. 2023. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10459–10469.

[65] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. 2024. Language Model Beats Diffusion-Tokenizer is key to visual generation. In *International Conference on Learning Representations*.

[66] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. 2025. Identity-Preserving Text-to-Video Generation by Frequency Decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[67] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. 2025. Magic Mirror: ID-Preserved Video Generation in Video Diffusion Transformers. *arXiv preprint arXiv:2501.03931* (2025).

[68] Yunpeng Zhang, Qiang Wang, Fan Jiang, Yaqi Fan, Mu Xu, and Yonggang Qi. 2025. Fantasyid: Face knowledge enhanced id-preserving video generation. *arXiv preprint arXiv:2502.13995* (2025).

[69] Dingcheng Zhen, Shunshun Yin, Shiyang Qin, Hou Yi, Ziwei Zhang, Siyuan Liu, Gan Qi, and Ming Tao. 2025. Teller: Real-Time Streaming Audio-Driven Portrait Animation with Autoregressive Motion Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 21075–21085.

[70] Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. 2025. Concat-ID: Towards Universal Identity-Preserving Video Synthesis. *arXiv preprint arXiv:2503.14151* (2025).