Identity-Preserving Text-to-Video Generation by Frequency Decomposition

Shenghai Yuan[1,4], Jinfa Huang[3], Xianyi He[1,4], Yunyang Ge[1,4],
Yujun Shi[5], Liuhan Chen[1,4], Jiebo Luo[3], Li Yuan[1,2,†]

[1] Peking University, [2] Peng Cheng Laboratory, [3] University of Rochester,
[4] Rabbitpre Intelligence, [5] National University of Singapore

{yuanshenghai,HeXianyi,yunyang,chenliuhan}@stu.pku.edu.cn, shi.yujun@u.nus.edu,
yuanli-ece@pku.edu.cn, {jhuang90@ur,jluo@cs}.rochester.edu

https://pku-yuangroup.github.io/ConsisID

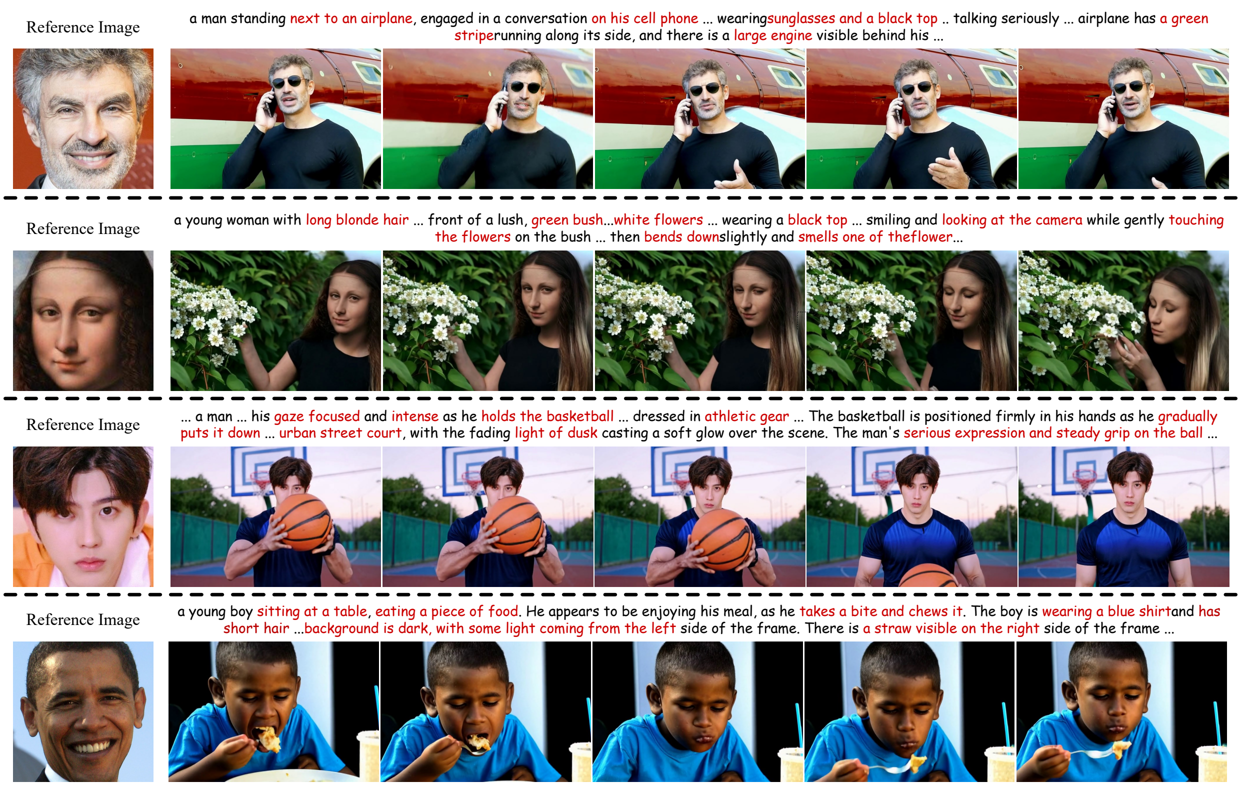

Figure 1. **Examples of identity-preserving video generation (IPT2V) by our ConsisID.** Given a reference image, our method can generate realistic and personalized human-centered videos while preserving identity. **Red** indicates that attributes in long instructions.

Abstract

*Identity-preserving text-to-video (IPT2V) generation aims to create high-fidelity videos with consistent human identity. It is an important task in video generation but remains an open problem for generative models. This paper pushes the technical frontier of IPT2V in two directions that have not been resolved in the literature: (1) A tuning-free pipeline without tedious case-by-case finetuning, and (2) A frequency-aware heuristic identity-preserving Diffusion Transformer (DiT)-based control scheme. To achieve these goals, we propose **ConsisID**, a tuning-free DiT-based controllable IPT2V model to keep human-**id**entity **consis**tent in the generated video. Inspired by prior findings in fre-*

*quency analysis of vision/diffusion transformers, it employs identity-control signals base on frequency domain, since facial features can be decomposed into low-frequency global features (*e.g.*, profile, proportions) and high-frequency intrinsic features (*e.g.*, identity markers that remain unaffected by pose changes). Extensive experiments demonstrate that our frequency-aware heuristic scheme provides an optimal control solution for DiT-based models, making strides towards more effective IPT2V.*

1. Introduction

Large-scale pre-trained video diffusion models [30, 66, 77, 78] have facilitated a variety of downstream applications [49, 54, 69, 70, 72, 74], particularly in identity-preserving text-to-video (IPT2V) [7, 38, 58, 60, 61]. However, existing methods face significant challenges, particularly the high overhead associated with the need for case-by-case finetuning, which diminishes their applicability. Within the open-source community, only the ID-Animator [15] can implement tuning-free IPT2V, but it can only generate videos similar to talking head [57] and has poor id preservation.

Additionally, the above efforts are predominantly based on U-Net and cannot be adapted to the emerging DiT-based video model [30, 63, 66, 77, 78]. This challenge may stem from the inherent limitations of DiT compared to U-Net, including greater difficulty in training convergence and weakness in perceiving facial details. From some prior findings in frequency analysis of vision/diffusion transformers [2–4, 45, 50, 55, 71], we can know that the reason is: **Finding 1:** *Shallow (*e.g.*, low-level, low-frequency) features are essential for pixel-level prediction tasks in diffusion models, as they ease model training.* U-Net facilitates model convergence by aggregating shallow features to the decoder via long skip connections, a mechanism that DiT does not incorporate; **Finding 2:** *Transformers have limited perception of high-frequency information, which is important for preserving facial features.* The encoder-decoder architecture of U-Net naturally possesses multi-scale features (*e.g.*, richness in high-frequency), while DiT lacks a comparable structure. To develop a DiT-based control model, these must be addressed first. Please see Appendix for more details.

For ID-preserving video generation, the challenges stem from the requirement for each frame to incorporate both high-frequency (*e.g.*, age- and make-up-independent identity markers) and low-frequency information (*e.g.*, facial shape) derived from the reference image, which can just be used to make up for the DiT defects mentioned above. Therefore, we propose **ConsisID**, to keep the **identity consis**tency in video generation by frequency decomposition, based on the previously **Findings of DiT** in frequency analysis. Thanks to the large-scale pre-trained DiT, we can use its powerful capabilities to achieve tuning-free effects. ConsisID decouples identity features into high- and low-frequency signals, which are injected into specific locations within the DiT, facilitating efficient IPT2V generation. Specifically, in line with *Finding 1*, we first convert the reference image and the facial key points to the low-frequency signal, then concatenate them with input noise latent to ease the training. Following *Finding 2*, we utilize a dual-tower feature extractor to capture high-frequency facial information, which is integrated with vision tokens within the transformer block, thereby enhancing the DiT's high-frequency perception capabilities. Finally, to transform the pre-trained model into an IPT2V model and improve its generalization, we further introduce a hierarchical training strategy.

Our contributions can be summarized as follows:

- We introduce **ConsisID**, a tuning-free identity-preserving DiT-based IPT2V model, which preserves the identity of the main subject of the video using control signals from frequency decomposition.
- We propose a hierarchical training strategy, including coarse-to-fine training, dynamic mask loss, and dynamic cross-face loss, which work together to facilitate training and enhance generalization effectively.
- Extensive experiments demonstrate that our **ConsisID** can generate high-quality, editable, consistent identity-preserving videos, benefiting from our frequency-aware identity-preserving T2V DiT-based control scheme.

2. Related Work

Tuning-based Identity-preserving T2V Models. Diffusion models are widely recognized for their strong generative capabilities [19, 34–37, 44, 72, 73], significantly advancing the development of identity-preserving generative models [8, 39, 59, 74]. Initially, the researchers used tuning-based methods to generate content that matched the input ID. This process requires finetuning pretrained model for each new person during inference. For example, DreamBooth [47] introduced a novel loss function to fine-tune the entire network, embedding identity information while preserving the original generative capabilities. LoRA [21], similar to DreamBooth [47], requires training only a small subset of network parameters. In contrast, Textual Inversion [11] freezes the pretrained network and embeds identity information into a trainable word embedding. Subsequent tuning-based methods, including both image and video models based on U-Net or DiT architectures [7, 26, 48, 58, 60–62, 65, 76], generally follow three main approaches. While these models demonstrate substantial effectiveness, the requirement to fine-tune for each new identity restricts their practical applicability.

Tuning-free Identity-preserving T2V Models. To address the issue of high resource consumption, several tuning-free diffusion models have recently emerged in the field of image generation [13, 14, 29, 56, 67]. These models

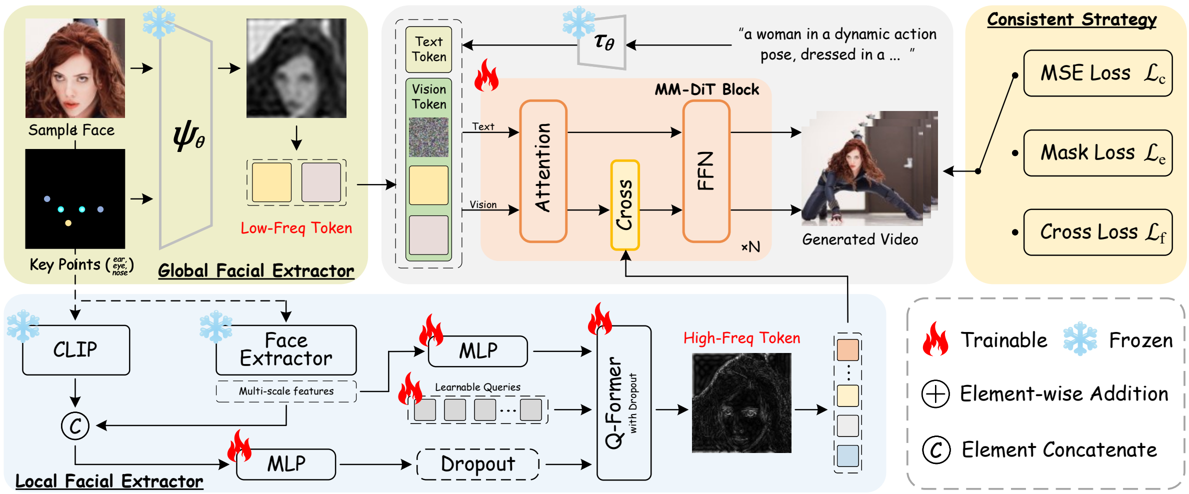

Figure 2. **Overview of the proposed ConsisID.** Based on *Findings of DiT*, low-frequency facial information is embedded into the shallow layers, while high-frequency information is incorporated into the vision tokens within the attention blocks. The ID-preserving Recipe is applied to ease training and improve generalization. The *cross face*, *DropToken* and *Dropout* are executed based on probability.

do not require finetuning parameters for newly introduced IDs during inference. For instance, IP-Adapter [67] utilizes the CLIP [43] features of the identity image through cross-attention to guide the pretrained model in generating identity-preserving images. InstantID [56] extends this approach by replacing CLIP [43] features with Arcface [10] features and integrating a pose network to adjust facial proportions. Unlike these initial methods, which introduce control signals via visual tokens, PhotoMaker [29] and Imagine Yourself [16] leverage text tokens. Specifically, PhotoMaker [29] concatenates identity features obtained from the CLIP encoder [43] to the text embedding, while Imagine Yourself [16] uses element-wise addition for feature fusion. In the domain of video generation, only MovieGen [42] and ID-Animator [15] currently support ID-preserving text-to-video (IPT2V) generation. MovieGen is closed-source, whereas ID-Animator is open-source but uses a methodology similar to image models, leading to lower-quality identity preservation in the generated videos. We select the emerging DiT architecture [30, 33, 66, 77] and optimize it for IPT2V, drawing on conclusions from prior frequency analyses [2–4, 45, 50, 55]. This enables high-quality, editable, and consistent ID-preserving video generation.

3. Methodology

3.1. Preliminaries

Diffusion Model. Text-to-video generation models usually utilize the diffusion paradigm, which gradually transforms noise ϵ into a video x_0. Originally, denoising was conducted directly within the pixel space [20, 51, 52]; however, due to significant computational overheads, recent methods predominantly employ latent space [12, 25, 46, 72]. The optimization process is defined:

$$\mathcal{L}_a = \mathbb{E}_{x_0,t,y,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(x_0, t, \tau_\theta(y)\right)\right\|_2^2\right], \quad (1)$$

where y is text condition, ϵ is sampled from a standard normal distribution (*e.g.*, $\epsilon \sim \mathcal{N}(0,1)$), and $\tau_\theta(\cdot)$ is the text encoder. By replacing x_0 with $\mathcal{E}(x_0)$, the latent diffusion is derived, which is used by ConsisID.

Diffusion Transformer. The DiT-based video generation model shows significant potential in simulating the physical world [6, 66, 78]. Despite being a novel architecture, research on controllable generation has been limited, and current methods [9, 13, 42, 75] largely resemble U-Net based approaches [11, 39, 74]. However, no study has yet examined why this approach works with DiT. Drawing from prior analyses of Diffusion and Transformer from a frequency domain perspective [2–4, 45, 50, 55], we conclude that: **(1)** Low-frequency (*e.g.*, shallow-layer) features are essential for pixel-level prediction tasks in diffusion models, which helps facilitate model training; **(2)** Transformers have limited perception for high-frequency information, which is important for controllable generation. Based on these, we decouple ID features into high- and low-frequency parts and inject them into specific locations, achieving effective identity-preserving text-to-video generation.

3.2. ConsisID: Keep Your Identity Consistent

The overview is illustrated in Figure 2. Given a reference image, the global facial extractor and local facial extractor inject both high- and low-frequency facial information into model, which then generates identity-preserving videos with the assistance of the consistency training strategy.

3.2.1. Low-frequency View: Global Facial Extractor

In light of *Finding 1*, enhancing low-level (*e.g.*, shallow, low-frequency) features accelerates model convergence. To easily adapt a pre-trained model for the IPT2V task, the most direct approach is concatenating the reference face with the noise input latent [5]. However, the reference face contains both high-frequency details (*e.g.*, eye and lip textures) and low-frequency information (*e.g.*, facial proportions and contours). From *Finding 2*, prematurely injecting high-frequency information into the Transformer is inefficient and may hinder the model's processing of low-frequency information, as the Transformer focuses primarily on low-frequency features. In addition, feeding the reference face directly into the model could introduce irrelevant noise such as lighting and shadows. To mitigate this, we extract facial key points, convert them to an RGB image, and then concatenate it with the reference image, as shown in Figure 2. This strategy focuses the model's attention on the low-frequency signals in the face, while minimizing the impact of extraneous features. We found that when this component is discarded, the model is difficult to convergen. The objective function is changed to:

$$\mathcal{L}_b = \mathbb{E}_{x_0,t,y,f,\epsilon}\left[\|\epsilon - \epsilon_\theta\left(x_0, t, \tau_\theta(y), \psi_\theta(f)\right)\|_2^2\right], \quad (2)$$

where $\psi_\theta(\cdot)$ is a variational autoencoder, f represents the reference image, we ignore key points here for simplicity.

3.2.2. High-frequency View: Local Facial Extractor

In light of *Finding 2*, we recognize that Transformers have limited sensitivity to high-frequency information. It can be concluded that relying solely on global facial extractor is insufficient for IPT2V generation, as global facial features lack of high-frequency information. So we use a face recognition backbone [10] to extract high-frequency features, as these are invariant to non-ID attributes (*e.g.*, expression, posture, and shape). We refer to these features as intrinsic identity features (*e.g.*, high-frequency), since age and makeup do not alter an individual's core identity. Following [14], we utilize the penultimate layer of the backbone, rather than its output, as it retains more spatial information pertinent to identity. However, our experiments reveal that while the face recognition backbone improves identity consistency, it lacks the semantic features required for editing. This task demands not only maintaining identity consistency but also incorporating the ability to edit, such as generating videos of faces with the same identity but varying age and makeup. Previous research [15, 16] relies solely on the CLIP encoder [43] to enable editing capabilities. However, since CLIP is not specifically trained on face datasets, the features it extracts include irrelevant non-face information, which can compromise identity fidelity [29, 56, 67].

To address these challenges, we first use a facial recognition backbone to extract features that strongly represent intrinsic identity, and a CLIP image encoder to capture semantically rich features. We then employ the Q-Former [27, 28, 64] to fuse these features, resulting in intrinsic identity representations enriched with high-frequency semantic information. To mitigate the impact of irrelevant features from CLIP, dropout [1, 22] is applied post-processing. Additionally, we follow [14] to concatenate the shallow, multi-scale features from the facial recognition backbone (after interpolation) with the CLIP features. This approach ensures that the model captures essential intrinsic identity features while filtering out extraneous noise unrelated to identity. Finally, we apply cross-attention to facilitate interaction between this feature set and the visual tokens produced by each attention block of the pre-trained model, thereby enhancing the high-frequency information in the DiT:

$$Z_i' = Z_i + \text{Attention}(Q_i^v, K_i^f, V_i^f), \quad (3)$$

where i represents the layer number of the attention block, $Q^v = Z_i W_i^q$, $K^f = F W_i^k$, and $V^f = F W_i^v$, where Z_i is the visual token, F represents the intrinsic identity features, and W_q, W_k, and W_v are trainable parameters. The objective function is changed to:

$$\mathcal{L}_c = \mathbb{E}_{x_0,t,y,f,\epsilon}\left[\|\epsilon - \epsilon_\theta\left(x_0, t, \tau_\theta(y), \psi_\theta(f), \varphi_\theta(f)\right)\|_2^2\right], \quad (4)$$

where $\varphi_\theta(\cdot)$ is the local facial extractor.

3.2.3. Consistency Training Strategy

During training, we randomly select a frame from the training frames and apply the Crop & Align [10] to extract the facial region as reference images, which is subsequently used as an identity-control signal, alongside the text as control.

Coarse-to-Fine Training. Compared to Identity-preserving image generation, video generation requires maintaining consistency in both spatial and temporal dimensions, ensuring that high and low-frequency facial information matches the reference image. To mitigate the complexity of training, we propose a hierarchical strategy where the model learns information globally before refining it locally. In the coarse-grained phase (*e.g.*, corresponding *Finding 1*), we employ the global facial extractor, enabling the model to prioritize low-frequency features, such as facial contours and proportions, thereby ensuring rapid acquisition of identity information from the reference image and consistency across the video sequence. In the fine-grained phase (*e.g.* corresponding to *Finding 2*), the local facial extractor shifts the model's focus to high-frequency details, such as the texture details of eyes and lips (*e.g.*, intrinsic identification), improving the fidelity of facial expressions and the overall similarity of the generated face.

Dynamic Mask Loss. The objective of our task is to ensure that the identity of the person in the generated video remains consistent with the input reference image. However, Equation 1 considers the entire scene, encompassing both high- and low-frequency identity information as well as redundant background content, which introduces noise that

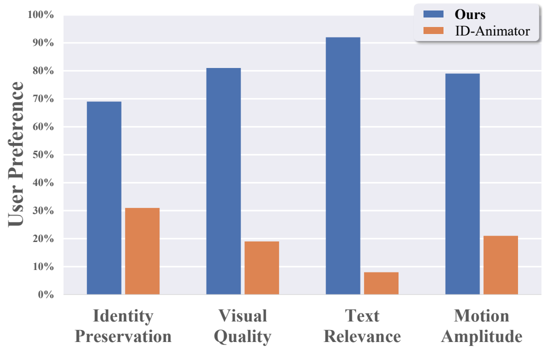

Figure 3. **User Study between ConsisID and state-of-the-art methods.** ConsisID is preferred by voters in all dimensions.

interferes with model training. To address this, we propose to focus the model's attention on face regions. Specifically, we first extract the facial mask from the video, apply trilinear interpolation to map it to the latent space, and finally use this mask to constrain the computation of $\mathcal{L}_c$:

$$\mathcal{L}_d = M \odot \mathcal{L}_c, \tag{5}$$

where M represents a mask with the same shape as ϵ. However, if Equation 5 is used as the supervisory signal for all training data, the model may fail to generate a natural background during inference. To mitigate this issue, we apply Equation 5 with a probability p of α, resulting in:

$$\mathcal{L}_e = \begin{cases} \mathcal{L}_d, & \text{if } p > \alpha \\ \mathcal{L}_c, & \text{if } p \leq \alpha \end{cases} \tag{6}$$

Dynamic Cross-Face Loss. After training with Equation 6, we observed that the model struggled to generate satisfactory results for persons not present in the data domain during inference. This issue arises because the model, trained exclusively on faces from the training frames, tends to overfit by adopting a "copy-paste" shortcut—essentially replicating the reference image without alteration. To improve the model's generalization capability, we introduce slight Gaussian noise ζ to the reference images and use cross-face (*e.g.*, reference images are sourced from video frames outside the training frames) as inputs with probability β:

$$\mathcal{L}_f = \begin{cases} \mathcal{L}_e \quad \text{where} \quad x_0 \cdot \zeta, & \text{if } p > \beta \\ \mathcal{L}_e \quad \text{where} \quad x_c \cdot \zeta, & \text{if } p \leq \beta \end{cases} \tag{7}$$

where x_0 is the reference image extracted from the training frames, and x_c is extracted from outside the training frames.

4. Experiments

4.1. Setup

Implementation details. **ConsisID** selects DiT-based generation architectures CogVideoX-5B [66] as our baseline for validation. We use an in-house human-centric dataset for training, which differs from previous datasets [40, 57, 68] that focus only on the face. In the training phase, we set the resolution to 480×720 and extracted 49 consecutive frames at a stride of 3 from each video as training data. We set the batch size to 80, the learning rate to 3×10^{-6}, and the total number of training steps to 1.8k. The randomly discarded text rate is set to 0.1, with AdamW serving as an optimizer and *cosine_with_restarts* as a learning rate scheduler. The training strategy is the same as Section 3.2.3. We set α and β in the dynamic cross-face loss ($\mathcal{L}_e$) and dynamic mask loss ($\mathcal{L}_f$) to 0.5, respectively. In the inference phase, we employ DPM [51] with a sampling step of 50, and a classifier free guidance ratio of 6.0. For more detials and results, please refer to Appendix.

	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
ID-Animator [15]	0.32	0.33	24.97	**117.46**
ConsisID	**0.58**	**0.60**	**27.93**	151.82

Table 1. **Quantitative comparison with state-of-the-art methods.** ConsisID achieve well-aligned results across most metrics. "↓" denotes lower is better. "↑" denotes higher is better.

Benchmark. Since there is an absence of an evaluation dataset, we select 30 persons who were not included in the training data and sourced five high-quality images for each ID from the internet. We then design 90 distinct prompts, encompassing a variety of expressions, actions, and backgrounds for evaluation. Building on previous works [15, 42], we evaluate four dimensions: **(1)**. Identity Preservation: We use FaceSim-Arc [10] and introduce FaceSim-Cur, which assesses identity preservation by measuring feature differences between face regions in the generated videos and those in real face images within the ArcFace [10] and CurricularFace [23] feature spaces. **(2)** Visual Quality: We utilize FID [18] by calculating feature differences in the face regions between the generated frames and real face images within the InceptionV3 [53] feature space. **(3)** Text Relevance: We utilize CLIPScore [17] to measure the similarity between the generated videos and the input prompts. **(4)**. Motion Amplitude: Due to the lack of reliable metrics [24, 73], we evaluate through the user study.

4.2. Qualitative Analysis

In this section, we compare our method, ConsisID, with ID-Animator [15] (*e.g.*, the only available open-source model) for tuning-free IPT2V tasks. We randomly select images and text prompts of four individuals for qualitative analysis, all of which are absent from the training data. As shown in Figure 4, ID-Animator cannot generate human body parts beyond the face and is unable to generate complex actions or backgrounds in response to text prompts (*e.g.*, action, attribute, background), which significantly limits its practical application. In addition, the preservation of the identity is inadequate; for example, in case 1, the reference image

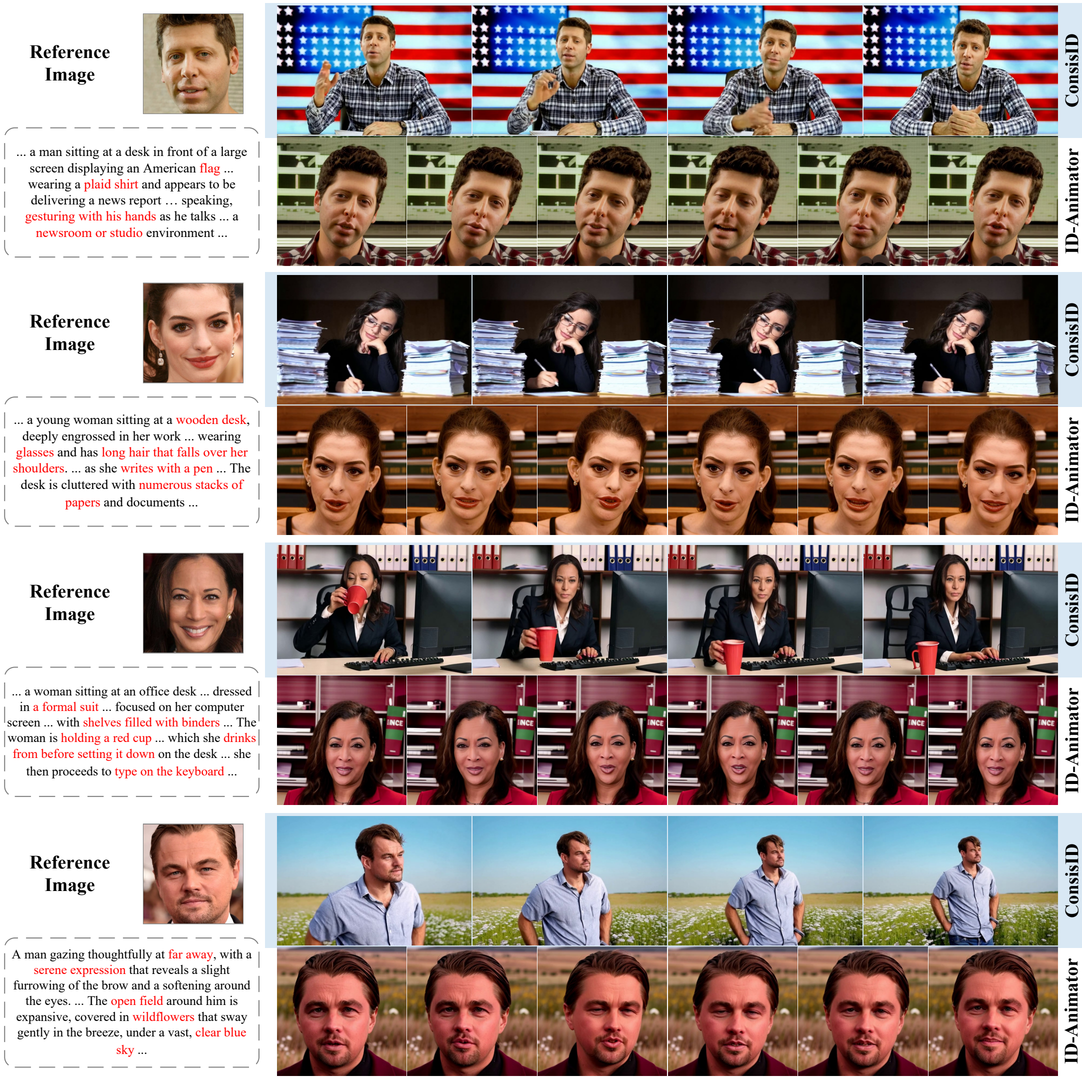

Figure 4. **Qualitative analysis between ConsisID and ID-Animator [15].** ID-Animator can only generate videos of the face region, and the identity Preservation is poor (*e.g.*, shape, texture). Additionally, it cannot generate specified content according to the text prompt (*e.g.*, action, decoration, background). ConsisID achieves advantages in identity preservation, visual quality, motion amplitude, and text relevance. Moreover, our ConsistID can generate more frames rather than ID-Animator (49 480×720p frames v.s. 16 512×512p frames).

appears to be processed with skin smoothing. In case 2, wrinkles have been introduced which detract from the aesthetic quality. In cases 3 and 4, the face is distorted due to the lack of low frequency information, which compromises identity consistency. In contrast, the proposed ConsisID consistently produces high-quality, realistic videos that accurately match the reference identity and adhere to prompt.

4.3. Quantitative Analysis

We present a comprehensive quantitative evaluation of different methods, with results displayed in Table 1. Consistent with Figure 4, our method outperforms state-of-the-art methods across five metrics. For identity preservation, ConsisID achieves a higher score by designing appropriate identity signals for DiT from a frequency perspective. By contrast, ID-Animator [15] is not optimized for IPT2V and only partially retains facial features, resulting in lower FaceSim-Arc [10] and FaceSim-Cur scores. For Text Relevance, ConsisID not only controls expressions via prompts but also adjusts actions and backgrounds, achieving higher CLIPScore [17]. Regarding visual quality, the FID is presented solely as a reference due to its limited alignment [31, 32, 41, 73] with human perception. Please refer to Figure 4 and 3 for qualitative analysis of the visual quality.

4.4. User Study

Building on previous work, we conduct a human evaluation using a binary voting strategy, with each questionnaire con-

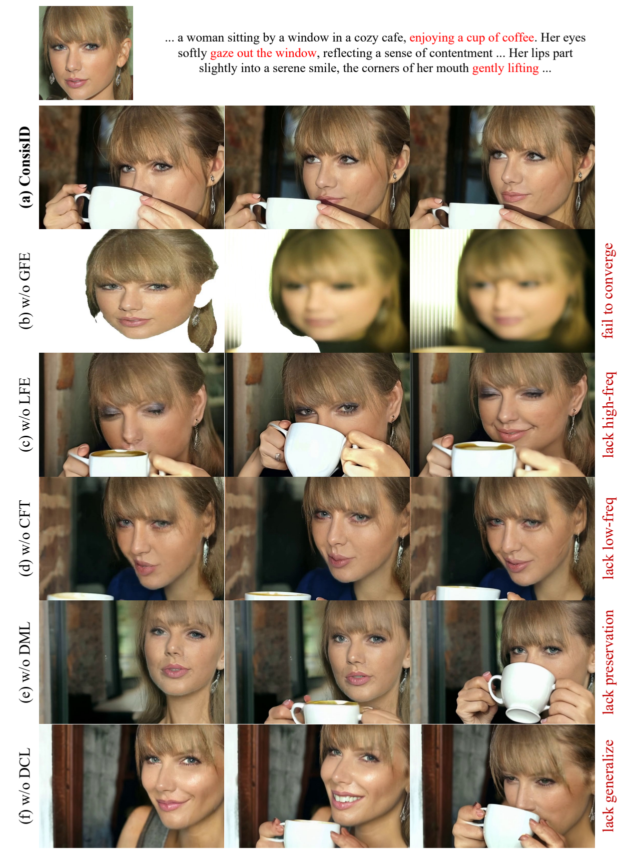

Figure 5. **Effect of Different Components via Qualitative Analysis.** Removing any component may result in the loss of high- or low-frequency facial information, or hinder the ability to modify video content based on the text prompt.

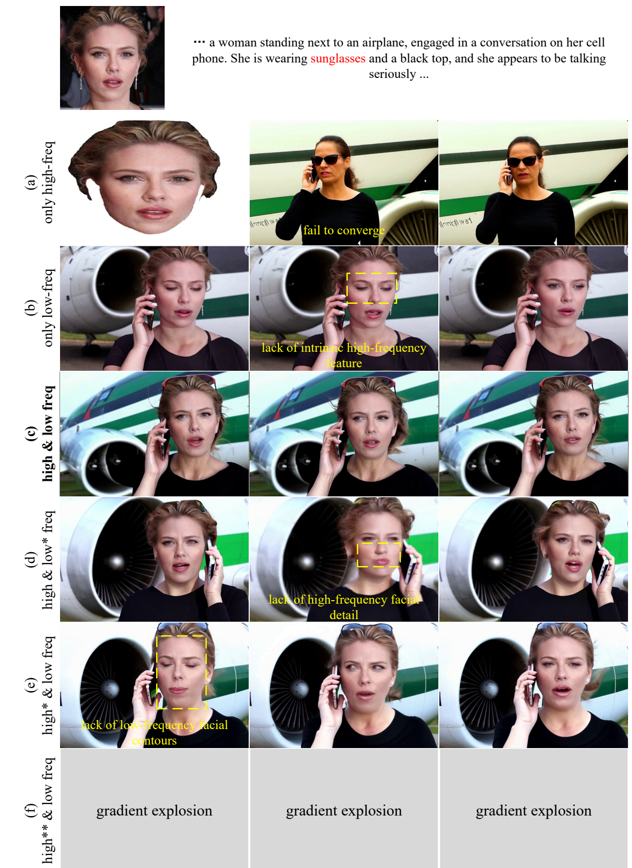

Figure 6. **Effect of Different Control Signal Injection Way via Qualitative Analysis.** Only (c), which injects both high & low-freq face signals into the suitable location, performs best.

taining only 80 questions. Participants are required to view 40 video clips, a setup designed to improve both engagement and questionnaire validity. For the IPT2V task, each question requires participants to separately judge which option performs better in terms of Identity Preservation, Visual Quality, Text Alignment, and Motion Amplitude. This composition ensures the accuracy of the human evaluation. Owing to the extensive participant base required for this evaluation, we successfully gathered 103 valid questionnaires. The results, depicted in Figure 3, demonstrate a significant superiority of our method over ID-Animator [15], verifying the effectiveness of the designed DiT for IPT2V generation.

4.5. Effect of the Identity Signal Injection in DiT

To assess the effectiveness of *Finding 1* and *Finding 2*, we perform ablation studies on different methods of injecting control signals into DiT. Specifically, these experiments involved **(a)** injecting only low-frequency face information with key points into the noise latent, **(b)** injecting only high-frequency face signals within the attention block, **(c)** combining (a) and (b), **(d)** based on (c), but the low-frequency

	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
w/o GFE	0.05	0.05	34.86	269.88
w/o LFE	0.66	0.68	34.48	**104.34**
w/o CFT	0.54	0.58	34.47	144.62
w/o DML	0.62	0.67	34.23	187.78
w/o DCL	0.65	0.69	32.21	117.80
ConsisID	**0.73**	**0.75**	**36.77**	127.42

Table 2. **Effect of *Local Facial Extractor (LFE)*, *Global Facial Extractor (GFE)*, *coarse-to-fine training (CFT)*, *dynamic mask loss (DML)* and *dynamic cross-face loss (DCL)* by Automatic Metrics.** Removing any of the above methods significantly reduces identity preservation, text relevance, and visual quality.

face information does not contain key points, and **(e - f)** based on (c), but the high-frequency signal is injected at the output or input of the attention block. **(g)** injecting only high-frequency face signals before the attention block. To reduce overhead, for each identity, we only select 2 reference images each with 90 text prompts for the evaluation. The results are shown in Figure 6 and Table 3. For *Finding 1*, we observe that only injecting high-frequency signals (a)

Plan	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
a	0.05	0.05	34.86	269.88
b	0.66	0.68	34.48	**104.34**
c	**0.73**	**0.75**	**36.77**	127.42
d	0.64	0.68	30.69	177.65
e	0.62	0.66	33.61	164.15
f	*unstable training process*			
g	*unstable training process*			

Table 3. **Effect of Different Control Signal Injection Way via Quantitative Analysis.** Only **plan c**, which injects both high and low-frequency face information into the model, performs best.

greatly increases the training difficulty, causing the model to fail to converge due to the lack of low-frequency signal injection. In addition, the inclusion of facial key points (d) allows a greater focus on low-frequency information, thereby facilitating training and improving model performance. For *Finding 2*, when only low-frequency signals are injected (b), the model lacks high-frequency information. This reliance on low-frequency signals causes the generated face in the video to copy the reference image, making it difficult to control facial expressions, movements, and other features through prompts. Furthermore, injecting identity signals into the attention block input (f - g) disrupts the intended frequency domain distribution of DiT, resulting in a gradient explosion. Embedding control signals in the attention block (c) is preferable to embedding them in the output (e) because attention block processes predominantly low-frequency information. By embedding high-frequency information internally, the attention block is guided to highlight intrinsic facial features, whereas injecting it into the output merely concatenates features without directing focus, reducing DiT's modeling capacity. Moreover, we apply a Fourier transform to the generated videos (only the face region) to visually compare the influence of different components to extract facial information. As shown in Figure 7, the Fourier spectrum and the log amplitude of the Fourier transform reveal that injecting high or low-frequency signals can indeed enhance the corresponding frequency information of the generated face. Moreover, the low-frequency signal can be further enhanced by matching with the face key points, and injecting the high-frequency signal into the attention block has the highest feature utilization rate. Our method (c) shows strongest high and low frequency, further validating the efficiency benefit from *Findings 1 and 2*.

4.6. Ablation on the Consistency Training Strategy

To reduce overhead, for each identity, we only select 2 reference images for the following experiments. To demonstrate the benefits of the proposed consistency training strategy, we perform ablation experiments on coarse-to-fine training (CFT), dynamic mask loss $\mathcal{L}_e$ (DML), and dynamic cross-face loss $\mathcal{L}_f$ (DCL), with the results presented in Figure

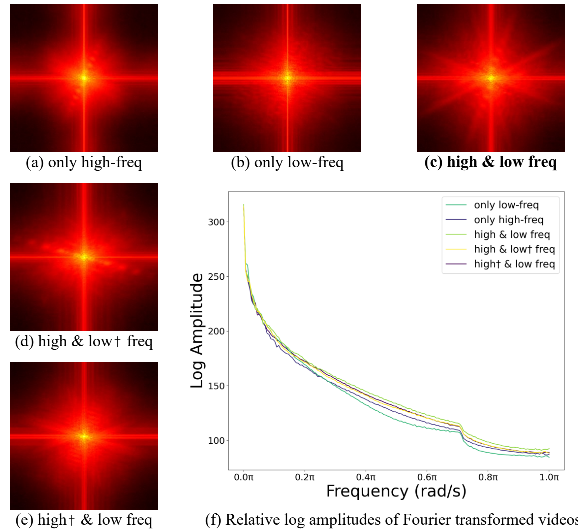

Figure 7. **(a - e) Fourier spectrum of different id signal injection.** The center area represents low frequencies and the surrounding area represents high frequencies. **(f) Relative log amplitudes of Fourier transformed generated videos.** A larger response value indicates a higher inclusion of frequency information. (a - f) verify the effect of our frequency decomposition.

5 and Table 2. When CFT is removed, GFE and LFE exhibit competing behaviors, complicating the model's ability to prioritize high and low-frequency information accurately, leading to convergence at suboptimal points. Removing DML required the model to simultaneously focus on both foreground and background elements, with background noise negatively affecting training and reducing facial consistency. Similarly, the exclusion of DCL impaired the generalization capability, reducing fidelity for faces, not in the training set and reducing its effectiveness in generating identity-preserving videos as intended.

5. Conclusion

In this paper, we present **ConsisID**, a unified framework for keeping faces consistent in video generation by frequency decomposition. It can seamlessly integrate into existing DiT-based text-to-video models, for generating high-quality, editable, consistent identity-preserving videos. Extensive experiments show that **ConsisID** outperforms the current state-of-the-art identity-preserving T2V models. It reveals that our frequency-aware heuristic DiT-based control scheme is an optimal solution for IPT2V generation.

6. Acknowledgments

We thank all the anonymous reviewers for their constructive comments. This work was supported in part by the Natural Science Foundation of China (No. 62202014, 62332002, 62425101, 62088102).

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*, 34:24206–24221, 2021.

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[3] Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. In *ECCV*, pages 1–18. Springer, 2022.

[4] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, pages 22669–22679, 2023.

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. In *openai*, 2024.

[7] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *arXiv preprint arXiv:2407.08674*, 2024.

[8] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023.

[9] Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802*, 2024.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ICLR*, 2024.

[13] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *arXiv preprint arXiv:2404.16022*, 2024.

[14] Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single- and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024.

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

[16] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024.

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[22] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016.

[23] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *CVPR*, pages 5901–5910, 2020.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[25] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[26] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023.

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022.

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023.

[29] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, pages 8640–8650, 2024.

[30] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[31] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[32] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *NeurIPS*, 36, 2024.

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

[34] Yue Ma, Xiaodong Cun, Yingqing He, Chenyang Qi, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*, 2023.

[35] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024.

[36] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-your-click: Open-domain regional image animation via short prompts. *arXiv preprint arXiv:2403.08268*, 2024.

[37] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

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

[39] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[40] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.

[41] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *CVPR*, pages 14277–14286, 2023.

[42] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021.

[45] Chen Rao, Guangyuan Li, Zehua Lan, Jiakai Sun, Junsheng Luan, Wei Xing, Lei Zhao, Huaizhong Lin, Jianfeng Dong, and Dalong Zhang. Rethinking video deblurring with wavelet-aware dynamic transformer and diffusion model. *arXiv preprint arXiv:2408.13459*, 2024.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *CVPR*, pages 6527–6536, 2024.

[49] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent YF Tan, and Jiashi Feng. Instadrag: Lightning fast and accurate drag-based image editing emerging from videos. *arXiv preprint arXiv:2405.13722*, 2024.

[50] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, pages 4733–4743, 2024.

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015.

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv*, 2020.

[53] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[54] Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Wangbo Yu, Chaoran Feng, Yatian Pang, Bin Lin, and Li Yuan. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle. *arXiv preprint arXiv:2407.19548*, 2024.

[55] Yuchuan Tian, Zhijun Tu, Hanting Chen, Jie Hu, Chao Xu, and Yunhe Wang. U-dits: Downsample tokens in u-shaped

diffusion transformers. *arXiv preprint arXiv:2405.02730*, 2024.

[56] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[57] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, pages 10039–10049, 2021.

[58] Zhao Wang, Aoxue Li, Enze Xie, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024.

[59] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, pages 15943–15953, 2023.

[60] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *CVPR*, pages 6537–6549, 2024.

[61] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*, 2024.

[62] Jing Xiong, Gongye Liu, Lun Huang, Chengyue Wu, Taiqiang Wu, Yao Mu, Yuan Yao, Hui Shen, Zhongwei Wan, Jinfa Huang, et al. Autoregressive models in vision: A survey. *arXiv preprint arXiv:2411.05902*, 2024.

[63] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024.

[64] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024.

[65] Shuo Yang, Kun-Peng Ning, Yu-Yang Liu, Jia-Yu Yao, Yong-Hong Tian, Yi-Bing Song, and Li Yuan. Is parameter collision hindering continual learning in llms? *arXiv preprint arXiv:2410.10179*, 2024.

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

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

[68] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *CVPR*, pages 14805–14814, 2023.

[69] Wangbo Yu, Chaoran Feng, Jiye Tang, Xu Jia, Li Yuan, and Yonghong Tian. Evagaussians: Event stream assisted gaussian splatting from blurry images. *arXiv preprint arXiv:2405.20224*, 2024.

[70] 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. *arXiv preprint arXiv:2409.02048*, 2024.

[71] Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785*, 2024.

[72] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. *arXiv preprint arXiv:2404.05014*, 2024.

[73] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Rui-Jie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *NeurIPS*, 37:21236–21270, 2025.

[74] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

[75] Zhenghao Zhang, Junchao Liao, Menghao Li, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv preprint arXiv:2407.21705*, 2024.

[76] Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2412.02259*, 2024.

[77] Zangwei Zheng, Xiangyu Peng, and Yang You. Open-sora: Democratizing efficient video production for all. In *Github*, 2024.

[78] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024.