FaithfulFaces: Pose-Faithful Facial Identity Preservation for Text-to-Video Generation

Yuanzhi Wang[1,2,*], **Xuhua Ren**[2], **Jiaxiang Cheng**[2], **Bing Ma**[2], **Kai Yu**[2], **Sen Liang**[2,3], **Wenyue Li**[2], **Tianxiang Zheng**[2], **Qinglin Lu**[2,†], **Zhen Cui**[4,†]

1. Nanyang Technological University
2. Tencent Hunyuan
3. University of Science and Technology of China
4. Beijing Normal University

Abstract

Identity-preserving text-to-video generation (IPT2V) empowers users to produce diverse and imaginative videos with consistent human facial identity. Despite recent progress, existing methods often suffer from significant identity distortion under large facial pose variations or facial occlusions. In this paper, we propose *FaithfulFaces*, a pose-faithful facial identity preservation learning framework to improve IPT2V in complex dynamic scenes. The key of FaithfulFaces is a pose-shared identity aligner that refines and aligns facial poses across distinct views via a pose-shared dictionary and a pose variation–identity invariance constraint. By mapping single-view inputs into a global facial pose representation with explicit Euler angle embeddings, FaithfulFaces provides a pose-faithful facial prior that guides generative foundations toward robust identity-preserving generation. In particular, we develop a specialized pipeline to curate a high-quality video dataset featuring substantial facial pose diversity. Extensive experiments demonstrate that FaithfulFaces achieves state-of-the-art performance, maintaining superior identity consistency and structural clarity even as pose changes and occlusions occur. **Project Page:** `https://mdswyz.github.io/FaithfulFaces/`

1 Introduction

Identity-preserving text-to-video generation (IPT2V) is a specialized facet of content creation that aims to generate various videos from the user-provided reference image and text prompts while maintaining consistent human facial identity across consecutive frames [46, 44]. This task showcases the potential to create and author visual content across domains, including but not limited to film and television production, personalized avatars, advertising design, and social multimedia content.

Benefiting from the robust generative capabilities of the large-scale pre-trained video foundational generative models [21, 45, 36], the IPT2V task can seamlessly extend these models to generate videos guided by reference face images. To generate videos with high-fidelity facial identity, researchers have proposed various methods to represent the identity information of the reference image. For example, ID-Animator [8] used a lightweight face adapter to encode the identity-relevant embeddings. ConsisID [46] designed two facial extractors to extract global low-frequency structure and local high-frequency details for IPT2V. At the same time, many commercial tools, such as Vidu [35], Kling [20], have also been adapted to the IPT2V task. Therefore, this task is the focus of the GenAI field and has attracted widespread attention.

[*]Work done during the internship at Tencent Hunyuan.
[†]Corresponding Authors: Qinglin Lu and Zhen Cui

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

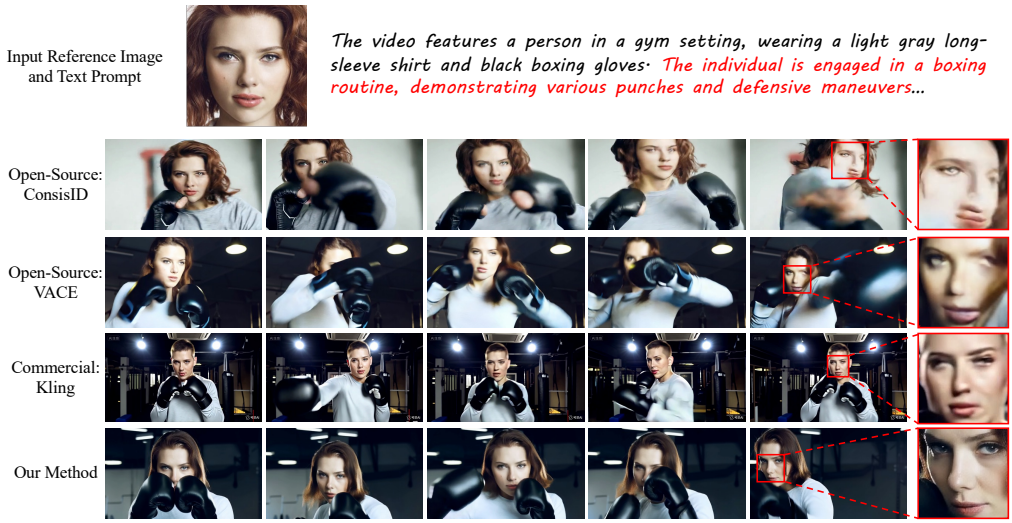


Figure 1: Visualization results from four different IPT2V methods. ConsisID [46] shows severe distortion of the facial structure. VACE [17] and Kling [20] suffer from significant distortion of facial identity details. In contrast to these open-source and commercial methods, our method exhibits clear facial structure and high-fidelity identity details as the facial pose changes and occlusions occur.

Despite their notable success, existing methods still exhibit limitations in effectively handling certain intricate scenarios. As shown in Fig. 1, we visualize the generation results of different methods in a complex dynamic case, where ConsisID and VACE [17] are two representative open-source methods, based on CogVideoX-5B [45] and Wan2.1-14B [36], respectively. Kling is one of the most popular and powerful commercial models. In this case, the goal is to generate a video depicting a subject performing a boxing action, which often involves significant variations in facial pose as well as facial occlusions. We can observe that both open-source and commercial approaches tend to produce noticeable distortion in the facial region as the subject moves and their facial expressions or pose change. This phenomenon may be attributed to the fact that such methods can only capture a single facial pose information from an input reference image, limiting their ability to handle scenarios with significant variations in facial pose. A question arises: *Can we capture global facial pose information from an input single-view image?*

In this paper, we propose a pose-faithful facial identity preservation learning framework, named *FaithfulFaces*, to address the aforementioned problem. We first propose a pose-shared identity aligner to encode global facial pose representation from the input single-view reference image. This aligner establishes a pose-shared dictionary to project diverse facial poses into a refined dictionary space, which is learned by a well-crafted pose variation–identity invariance constraint. In this constraint, face images from the same identity but with different poses are treated as positive pairs, while others serve as negative samples. In particular, we incorporate Euler angle embedding learning into the aligner to provide explicit pose cues during the refinement and alignment processes.

Furthermore, to support our FaithfulFaces learning, we design a new dataset collection and processing pipeline that constructs a high-quality, task-specific video dataset with significant facial pose variations to provide a robust training foundation. Finally, the well-trained framework is capable of naturally extracting global facial pose representations as holistic facial priors, enabling foundational generative models to better preserve identity in generated videos. As illustrated in Fig. 1, our method demonstrates superior consistency in maintaining facial identity throughout the generated video as the facial pose changes and occlusions occur. The contributions of this work are threefold:

- We systematically analyze the limitations and potential reasons of existing IPT2V methods in complex facial dynamic scenes, and propose a pose-faithful facial identity preservation learning paradigm, FaithfulFaces, to better preserve consistent identity in generated videos.
- We design a pose-shared identity aligner to encode global facial pose representation from the input single-view reference image via a pose-shared dictionary and a pose variation–identity invariance constraint with Euler angle embedding learning. Additionally, we develop a new

dataset pipeline to construct a task-oriented, high-quality video dataset with substantial facial pose diversity to ensure robust model training.

- We perform extensive experiments across diverse identity and dynamic scenarios. Both quantitative and qualitative results demonstrate the effectiveness of our FaithfulFaces, surpassing existing open-source and commercial methods.

2 Related Work

Thanks to the powerful data distribution modeling capability and stable training process of the continuous-time generative models [32, 23, 25, 39, 37, 18, 38], large-scale text-to-video generative models [30, 21, 45, 36, 6, 42, 19, 41, 40, 43] have been rapidly developed, further facilitating the Identity-preserving text-to-video generation (IPT2V) task. In the early stage, He *et al.* [8] proposed the ID-Animator method that uses a Unet-based lightweight text-to-video model AnimateDiff [7] and builds a face adapter for IPT2V.

The recent Diffusion Transformer (DiT) architecture [28] has shown promising generative capabilities and has become a mainstream backbone for video generation, such as open-source models HunyuanVideo [21], CogVideoX [45], and Wan [36]. Therefore, many recent IPT2V works are built upon and extend the DiT-based models [46, 48, 47, 44, 49, 5, 3]. For example, ConsisID [46] utilized CogVideoX as the basic generative model and designed a global and local facial extractor to capture global structure and local details as identity information. HunyuanCustom [14] was built upon the HunyuanVideo foundational model. VACE [17], Phantom [24], SkyReels-A2 [5], MAGREF [3], and Stand-In [44] used Wan as the foundational model.

Furthermore, due to the extremely broad range of real-world applications for IPT2V, numerous successful commercial models and tools have emerged, such as Vidu [35], Pika [29], Kling [20]. However, whether open-source methods or commercial tools, they are difficult to deal with complex facial dynamics, leading to distorted identity information in the generated videos. Therefore, we propose a new learning framework to mitigate this issue.

3 Method

3.1 Problem Formulation

Problem. Let I_{ref} and $\mathcal{P}$ denote a reference face image and a text prompt describing the semantics of the target video, respectively. The goal of identity-preserving text-to-video (IPT2V) generation is to create a video $\mathcal{V}$ under the condition of I_{ref} and $\mathcal{P}$. Thus, $\mathcal{V}$ should satisfy: **i)** the semantic information of $\mathcal{V}$ is aligned with $\mathcal{P}$ (i.e., textual alignment); and **ii)** most importantly, the facial identity information of the subject in $\mathcal{V}$ is consistent with I_{ref}. The generation process can be formalized as:

$$\mathcal{V} = \mathcal{G}\left(\mathbf{Z} \sim \mathcal{N}(\mu, \sigma^2), \phi(I_{\text{ref}}), \mathcal{P}\right), \tag{1}$$

where $\mathcal{G}$ is a text-to-video foundational generative model (e.g., Wan [36]). $\mathbf{Z}$ is a prior state sampled from the Gaussian prior distribution. ϕ denotes a function used to encode the identity information of I_{ref}. For the above equation, the foundational model $\mathcal{G}$ determines the degree of semantic alignment between $\mathcal{V}$ and $\mathcal{P}$. Therefore, researchers only need to select the strongest pretrained model and keep its original prior knowledge (e.g., LoRA Adapter [13]) during training, which is not the focus of the IPT2V task. For the function ϕ, which determines the fidelity of facial identity information, i.e., the consistency of facial structure and the fidelity of facial texture details in the generated video $\mathcal{V}$. Thus, this is a critical issue in the IPT2V task, and researchers are dedicated to constructing a robust ϕ that accurately represents the subject's identity information.

Recent state-of-the-art works have made various attempts and proposed diverse ϕ to improve the performance of IPT2V. For example, ConsisID [46] proposed a global facial extractor and a local facial extractor to extract low-frequency structures and high-frequency details of the reference image I_{ref}, respectively. Magic Mirror [47] designed a dual-branch facial feature extractor to capture both identity and structural features. However, they may struggle to handle situations involving complex facial dynamics, such as drastic changes in facial poses and emotions, or facial occlusions, resulting in distorted facial identity and facial structure in the generated videos. The reason behind this phenomenon is that the encoded identity information can only represent a single pose view of the input image, failing to capture global pose information.

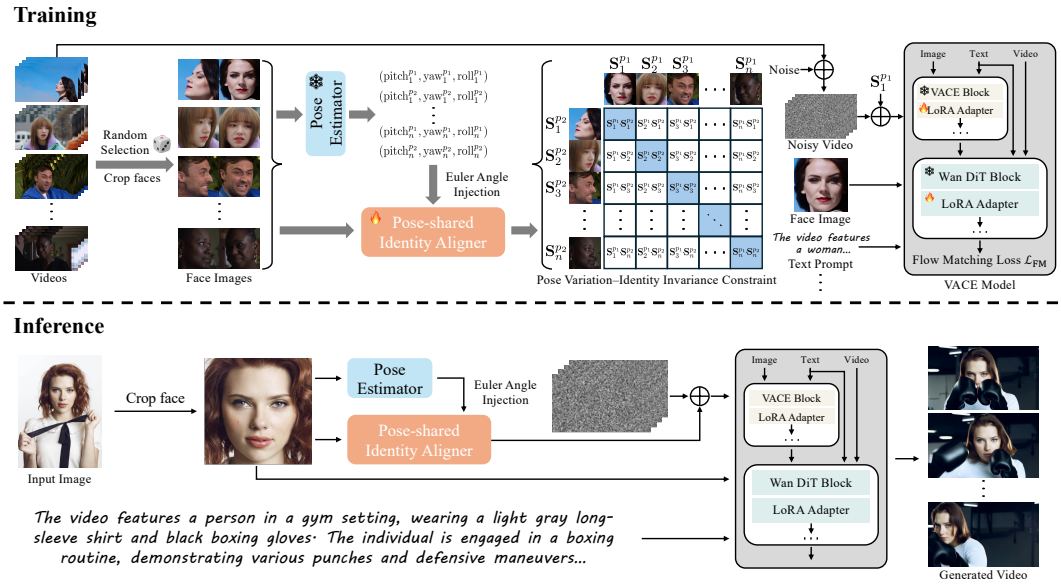


Figure 2: The framework of FaithfulFaces. During training, given n input videos per iteration, FaithfulFaces first randomly samples and crops two face images from each video. The cropped face images are then fed into a pose estimator to regress the three Euler angles, e.g., $(\text{pitch}_1^{p_1}, \text{yaw}_1^{p_1}, \text{roll}_1^{p_1})$, where p_1 and p_2 are simply used to mark two different poses. Next, the predicted Euler angles and the face images are then jointly fed into a pose-shared identity aligner, yielding $2n$ refined facial representations (e.g., $\mathbf{S}_1^{p_1}$), which are utilized to form a pose variation–identity invariance constraint. Finally, the refined representations are injected into the patch-embedded video tokens as identity conditioning for joint optimization. During inference, FaithfulFaces encodes a global facial pose feature from a single face image and incorporates it into the generative model to produce videos.

Main Idea. The identity information encoder could be partitioned into two parts: a basic facial identity encoder ϕ_{bas} and a global facial pose encoder ϕ_{gfp}. The former aims to encode the single-view facial structure information and facial texture details as existing methods do, and the latter aims to capture global facial pose representation. Formally, our generation process is defined as:

$$\mathcal{V} = \mathcal{G}\left(\mathbf{Z} \sim \mathcal{N}(\mu, \sigma^2), [\phi_{\text{bas}}(I_{\text{ref}}), \phi_{\text{gfp}}(I_{\text{ref}})], \mathcal{P}\right). \tag{2}$$

Accordingly, there are two questions that need to be solved:

- Global facial pose encoder ϕ_{gfp}. Representing faithful global facial pose from the input single-view reference image I_{ref} as introduced in Sec. 3.3.
- Automatic facial video dataset pipeline P_f. Collecting and preprocessing the video data with large changes in facial poses for training ϕ_{gfp} as introduced in Sec. 3.4.

3.2 Overview Framework

The overview framework of FaithfulFaces is illustrated in Fig .2, which is divided into the training stage and the inference stage. For the training stage, assuming there are n videos as input for each training iteration, we first randomly sample and crop two face images from each video. Subsequently, the cropped face images are fed into a pose estimator to regress the three Euler angles (i.e., pitch, yaw, roll) of the facial pose for each face image. These Euler angles, along with the face images, are then fed into our proposed pose-shared identity aligner to output $2n$ refined facial representations. Furthermore, the $2n$ facial representations from all video samples can be combined into two batches of facial data to form a pose variation–identity invariance constraint. In this constraint, face images from the same identity with different poses are paired as positive samples (diagonal pairs), while those of different identities are paired as negative samples. Finally, the output global facial pose representation is incorporated into the video generation backbone as identity conditioning. Specifically, the 5,120-dimensional identity representation is reshaped into a single token and zero-padded along the sequence

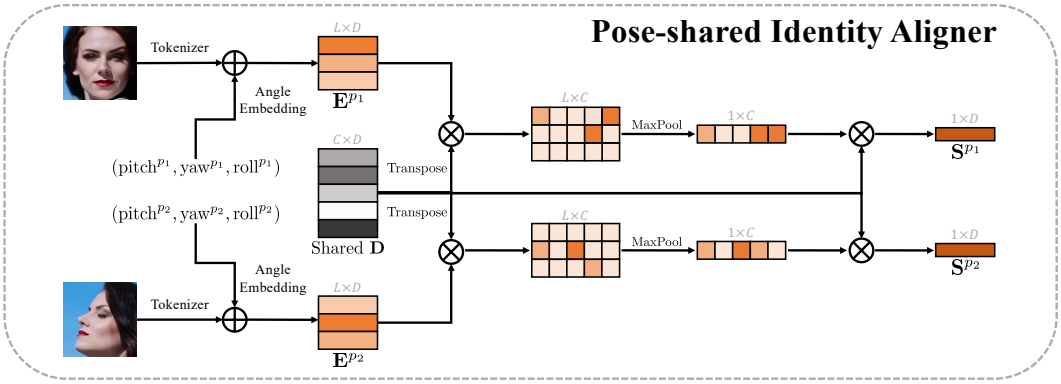


Figure 3: Architecture of the pose-shared identity aligner. Initially, the input face images are tokenized into sequential face embeddings. The corresponding Euler angles are encoded as Euler angle embeddings and injected into the face embeddings, resulting in the combined embeddings $\mathbf{E}^{p_1}$ and $\mathbf{E}^{p_2}$. Then, a pose-shared dictionary $\mathbf{D}$ is employed to refine and align $\mathbf{E}^{p_1}$ and $\mathbf{E}^{p_2}$, yielding the global facial pose representations $\mathbf{S}^{p_1}$ and $\mathbf{S}^{p_2}$. Finally, these representations serve as pose-faithful facial priors for the foundational generative model.

dimension to match the patch-embedded video tokens. The resulting identity conditioning is added to the video token sequence before the DiT Transformer blocks, rather than directly to the raw noisy latent. In practice, we utilize the VACE [17] as our foundational model and employ a LoRA training mode to fit these new data, where the VACE blocks are the basic facial identity encoder ϕ_{bas} in Eq. (2) to extract the single-view facial structure information and facial texture details.

During inference, users need only supply a single face image. The pose estimator regresses the Euler angles from this image, and both the angles and the image are passed to a well-trained identity aligner to generate the global facial pose representation. The resulting representation is then injected into the patch-embedded video tokens as identity conditioning, guiding the generative model to produce identity-consistent videos conditioned on the reference image and text prompt.

3.3 Pose-shared Identity Aligner for Global Facial Pose Representation

For the above framework, the most critical question is how to design and train the pose-shared identity aligner, i.e., encoder ϕ_{gfp} in Eq. (2), to represent robust global facial pose information.

Inspired by dictionary learning [34, 4], the key of our pose-shared identity aligner is to align the different facial poses into a refined dictionary space. Fig. 3 shows the architecture of the pose-shared identity aligner, which can receive face images with various poses and tokenize them into sequential face embeddings. These vanilla face embeddings contain only implicit pixel-level facial pose information, which hinders the model's ability to perceive facial pose. Thus, we aim to provide explicit pose information to guide the model's representation. Specifically, we utilize a pretrained facial pose estimator (6DRepNet [9] in practice) to regress three Euler angles: pitch, yaw, and roll. Notably, Euler angles possess a periodic property, which makes it natural to generate their embeddings using the timestep encoding method employed in diffusion models [12]. As shown in Fig. 3, we inject the Euler angle embeddings into the vanilla face embeddings to generate two new embeddings, marked as $\mathbf{E}^{p_1} \in \mathbb{R}^{L\times D}$ and $\mathbf{E}^{p_2} \in \mathbb{R}^{L\times D}$, where L and D denote the sequence length and dimensionality. p_1 and p_2 are simply used to mark two different poses. With these embeddings, we then define a learnable pose-shared dictionary matrix $\mathbf{D} \in \mathbb{R}^{C\times D}$, where C indicates the number of dictionary elements. Subsequently, $\mathbf{E}^{p_1} \in \mathbb{R}^{L\times D}$ and $\mathbf{E}^{p_2} \in \mathbb{R}^{L\times D}$ are projected into a dictionary space by calculating the correlation between each face embedding and $\mathbf{D}$ to obtain the correlation matrices, which can be further capsuled into two dictionary weights $\mathbf{W}^{p_1}$ and $\mathbf{W}^{p_2}$:

$$\mathbf{W}^{p_1} = \text{MaxPool}(\mathbf{E}^{p_1} \otimes \mathbf{D}^\top) \in \mathbb{R}^{1\times C}, \mathbf{W}^{p_2} = \text{MaxPool}(\mathbf{E}^{p_2} \otimes \mathbf{D}^\top) \in \mathbb{R}^{1\times C}, \tag{3}$$

where $\text{MaxPool}(\cdot)$ denotes a max pooling operation empirically determined in Appendix A.4. $\otimes$ means matrix multiplication. Finally, these dictionary weights can be used to obtain the global facial pose representations $\mathbf{S}^{p_1}$ and $\mathbf{S}^{p_2}$:

$$\mathbf{S}^{p_1} = (\mathbf{W}^{p_1} \otimes \mathbf{D}) \in \mathbb{R}^{1\times D}, \mathbf{S}^{p_2} = (\mathbf{W}^{p_2} \otimes \mathbf{D}) \in \mathbb{R}^{1\times D}. \tag{4}$$

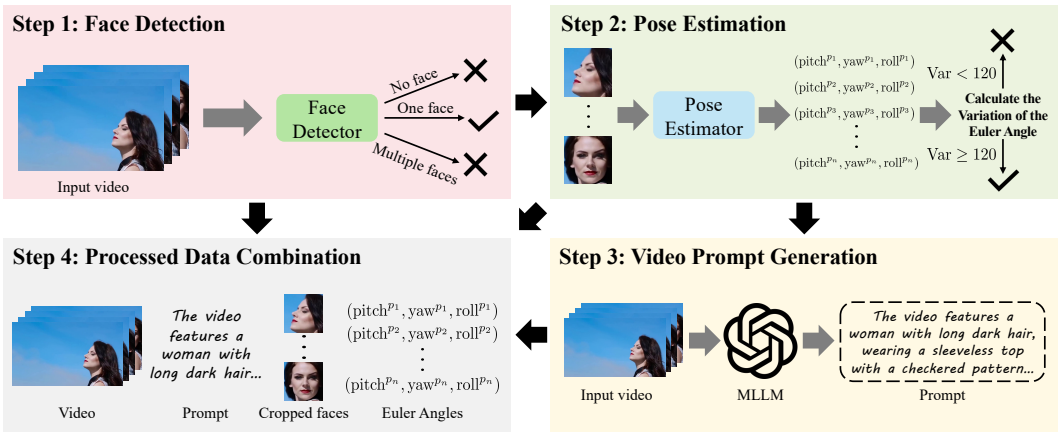


Figure 4: Dataset collection and processing pipeline. In Step 1, videos without face or with multiple faces are filtered out. Step 2 aims to select videos that exhibit significant variations in facial pose. For Step 3, we generate a text prompt for each selected video using MLLM. Step 4 ultimately integrates these fragmented data into a cohesive whole.

To optimize this aligner, we observe that the two batches of input facial data with different poses can exactly form a CLIP-like contrastive paradigm, as shown in the upper part of Fig. 2. Thus, we apply the most commonly used contrastive learning [31] to train our aligner:

$$\mathcal{L}_{\text{PIA}} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(\text{sim}(\mathbf{S}_i^{p_1}, \mathbf{S}_i^{p_2})/\tau)}{\sum_{j=1}^{n}\exp(\text{sim}(\mathbf{S}_i^{p_1}, \mathbf{S}_j^{p_2})/\tau)} - \frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(\text{sim}(\mathbf{S}_i^{p_2}, \mathbf{S}_i^{p_1})/\tau)}{\sum_{j=1}^{n}\exp(\text{sim}(\mathbf{S}_i^{p_2}, \mathbf{S}_j^{p_1})/\tau)}, \quad (5)$$

where n is the number of matched identity pairs in each training mini-batch, $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity function, and τ is a learnable temperature parameter with the default setting of [31]. During the whole training process, we integrate $\mathcal{L}_{\text{PIA}}$ with the objective of the generative model (i.e., flow matching [25]) $\mathcal{L}_{\text{FM}}$ to reach the full optimization objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PIA}} + \mathcal{L}_{\text{FM}}. \quad (6)$$

In practice, $\mathcal{L}_{\text{PIA}}$ and $\mathcal{L}_{\text{FM}}$ are responsible for their respective tasks during the training process. $\mathcal{L}_{\text{PIA}}$ is dedicated to constraining the alignment of different poses, while $\mathcal{L}_{\text{FM}}$ is dedicated to constraining the LoRA parameters to adapt to the input's global facial pose representation. This approach ensures that the different loss functions can focus on handling their specific tasks.

Remark 1 (**Deep insights and observations**) *Our design of the pose-shared identity aligner is not only intuitive but also admits a theoretical justification. Recall that $\mathcal{L}_{\text{PIA}}$ is equivalent to the InfoNCE loss [27], which provides a lower bound of the mutual information:*

$$I(\mathbf{S}^{p_1}; \mathbf{S}^{p_2}) \geq \log(n) - \mathcal{L}_{\text{PIA}}. \quad (7)$$

This inequality implies that minimizing $\mathcal{L}_{\text{PIA}}$ is not only aligning pose-variant embeddings but also maximizing the shared identity information across different poses. Hence, our aligner has an information-theoretic guarantee: the learned global representation cannot collapse unless $I(\mathbf{S}^{p_1}; \mathbf{S}^{p_2})$ vanishes. From the experimental observation, the visualization of the encoded facial identity in Fig. 6 confirms the above insights. Furthermore, the learned dictionary reveals meaningful activation patterns, wherein images with similar poses tend to frequently activate particular dictionary elements, as illustrated in Fig. 7. This indicates that the learned dictionary facilitates robust representation of faces across a wide range of poses.

3.4 Dataset Construction

Beyond framework design, a critical challenge persists: constructing a video dataset with significant variations in facial poses for training our proposed pose-shared identity aligner. This is because ordinary facial micro-movements or static videos are insufficient to satisfy our training requirements.

To address this issue, we construct a new dataset collection and processing pipeline. Note that this part omits standard data collection and preprocessing procedures that have been widely adopted in previous works [17, 14, 24], such as video clip segmentation, resolution standardization, OCR filter, aesthetic filter, clarity filter, etc. The original videos are sourced from OpenHumanVid [22], a publicly documented human-centric video dataset obtained through its official application-and-review process. We further curate and process these videos to construct our task-oriented training dataset, without independently collecting additional human-subject videos. The resolution of each video is standardized to 832×480 pixels. Fig. 4 illustrates our dataset collection and processing pipeline, which consists of four steps: face detection, pose estimation, video prompt generation, and processed data combination.

Face Detection. Since our work only focuses on the single-subject video generation task, we first need to filter out two types of videos: videos without face and videos with multiple faces. Specifically, we utilize InsightFace [16] for face detection on each video frame. Videos are filtered out if more than one face is detected in any single frame. Additionally, videos in which no faces are detected throughout the entire sequence are also excluded.

Pose Estimation. This step constitutes the core of the entire dataset pipeline, aiming to select videos that exhibit significant variations in facial pose. Taking a video $\mathcal{V}$ as an example, we first use the facial bounding boxes obtained in the previous step to crop the face regions from each video frame. These cropped face regions are then fed into the pose estimator 6DRepNet to predict three Euler angles for each detected face. Note that in practice, we enlarge the bounding boxes by a factor of 1.5 to predict Euler angles more accurately. Next, the three Euler angles for each face are stored separately in three lists, denoted as $\mathcal{X}_{\text{pitch}}$, $\mathcal{X}_{\text{yaw}}$, and $\mathcal{X}_{\text{roll}}$, and we can calculate the variation of Euler angles throughout the entire video:

$$\text{Var} = [\max(\mathcal{X}_{\text{pitch}}) - \min(\mathcal{X}_{\text{pitch}})] + [\max(\mathcal{X}_{\text{yaw}}) - \min(\mathcal{X}_{\text{yaw}})] + [\max(\mathcal{X}_{\text{roll}}) - \min(\mathcal{X}_{\text{roll}})], \quad (8)$$

where $\max(\cdot)$ and $\min(\cdot)$ represent the maximum and minimum values in the list, respectively. Furthermore, it is necessary to determine a reliable variation threshold to filter out qualified videos. To determine this threshold, we first randomly sample 2000 videos from the output of step 1 and manually annotate them. Our criterion for qualified videos is that the facial pose in the video must show at least a transition from frontal to profile (or vice versa), or exhibit significant up-and-down movement. Videos meeting these criteria are labeled as qualified, and we finally determine that the threshold is 120. With this threshold, we can retain videos with large facial pose changes; that is, $\text{Var} > 120$ is qualified, while $\text{Var} < 120$ is discarded.

Video Prompt Generation. After collecting qualified videos, we need to generate a text prompt for each video. Here, we use Qwen2.5-VL [1] to generate information-rich text prompts for qualified videos, focusing on describing the subjects' appearance, actions, and background. We then perform extensive manual calibration and refinement to improve the accuracy of text prompts.

Processed Data Combination. After the above three steps of data screening and preprocessing, we ultimately integrate these fragmented data into a cohesive whole. As shown in step 4 of Fig. 4, each sample in our well-constructed dataset contains four elements: video, text prompt, cropped face images, and Euler angles. We manually check all processed data to ensure that all videos are qualified enough, and ultimately generate 51,624 samples for model training.

4 Experiments

4.1 Implementation Details

Our FaithfulFaces framework utilizes the DiT-based generative model VACE-14B [17] as our foundational model. For the pose-shared identity aligner, the number of dictionary elements of $\mathbf{D}$ is set to 4096, empirically determined in Appendix A.2. We set the resolution of each video to 832×480 pixels and extract 81 consecutive frames for training. In the training phase, we use the LoRA training mode with rank 128 to fit new data. The whole framework is trained on 32 NVIDIA H20 GPUs with a batch size of 32. In addition, we set an independent batch size of 1024 for the pose-shared identity aligner to perform adequate pose alignment, and the total number of training steps is set to 5000.

Evaluation details. We conduct experiments and evaluations on several face images used in ConsisID [46], which consist of 30 persons, and we randomly sample one image for each identity. We then

Table 1: Quantitative comparison with evaluated baselines.

Methods	Open-Source	FaceSim-Cur ↑	FaceSim-Arc ↑	FID ↓	CLIPScore ↑
Vidu [35]	×	0.293	0.278	234.65	30.08
Kling [20]	×	0.447	0.416	194.80	33.06
ConsisID [46]	✓	0.365	0.350	205.03	30.29
VACE-14B [17]	✓	0.403	0.382	191.02	31.83
HunyuanCustom [14]	✓	0.453	0.432	187.32	31.36
Phantom-14B [24]	✓	0.484	0.456	214.99	29.67
Concat-ID-Wan [49]	✓	0.408	0.387	189.55	31.49
SkyReels-A2 [5]	✓	0.410	0.384	237.29	28.10
Stand-In [44]	✓	0.415	0.395	196.21	30.38
MAGREF [3]	✓	0.417	0.392	207.69	31.13
FaithfulFaces (Ours)	✓	**0.568**	**0.542**	**164.24**	**33.93**

construct 20 challenging text prompts that drive the models to generate videos featuring significant facial pose variations, expression changes, and facial occlusions across diverse scenarios. The details can be found in Appendix A.14. We consider four standard evaluation metrics that are used in prior works [30, 46] to measure the quality of generated videos. **FaceSim-Arc** and **FaceSim-Cur** are employed to assess identity preservation by measuring feature discrepancies between face regions in the generated videos and those in real face images within the ArcFace [2] and CurricularFace [15] feature spaces. For visual quality, we utilize the commonly used **FID** [11] by calculating feature differences in the face regions between the generated frames and real face images within the InceptionV3 [33] feature space. For textual alignment, we utilize the **CLIPScore** [10] to measure the similarity between videos and text prompts.

4.2 Baseline Comparisons

We compare our FaithfulFaces with the current state-of-the-art methods, including two commercial products (Vidu [35], Kling [20]) and eight open-source models (ConsisID [46], VACE [17], HunyuanCustom [14], Phantom [24], Concat-ID-Wan [49], SkyReels-A2 [5], Stand-In [44], MAGREF [3]). For these open-source methods, ConsisID is based on CogVideoX-5B [45], HunyuanCustom is based on HunyuanVideo [21], VACE, Phantom, Concat-ID-Wan, SkyReels-A2, Stand-In, and MAGREF are based on Wan [36], providing diversity for evaluation and comparison. For each method, we generate 600 videos (30 persons $\times$ 20 prompts) for evaluation, which is larger than the current top-tier community standard (e.g., MAGREF [3] generates 120 videos for evaluation).

Quantitative results. Tab. 1 lists the quantitative results of different methods. From these results, we can observe that FaithfulFaces achieves the best IPT2V performance under four evaluation metrics. In particular, FaithfulFaces gains considerable performance improvements in the FaceSim-Cur and FaceSim-Arc metrics used to measure identity preservation of generated videos. This improvement can be attributed to FaithfulFaces's ability to provide a robust global facial pose prior for foundational generative models, enabling more effective identity preservation.

Qualitative results. Fig. 5 provides some visual comparisons of our FaithfulFaces against seven baselines, including a case of generating a video of a person engaging in a boxing routine. We can first observe that the subjects in the videos generated by different methods have various facial pose changes and even facial occlusion due to intense movements. Furthermore, we discover that these open-source and commercial methods exhibit varying degrees of identity distortion and facial collapse as the subject moves. In contrast, our FaithfulFaces can output a high-quality video with clear facial structures and consistent identity details. Similar observations are made in Appendix A.18, where more visual results are provided. Video demos are included in the Supplementary Material.

4.3 Ablation Studies

We evaluate the effects of the key components in FaithfulFaces, including the pose-shared identity aligner (Aligner) and the injection of Euler angle embeddings (Euler). For a fair comparison, all ablation variants are fine-tuned from the same VACE-14B backbone using identical 51,624 training samples, LoRA configurations, optimization settings, and training steps. In particular, the "w/o (Aligner, Euler)" variant undergoes the same fine-tuning procedure as FaithfulFaces, thereby

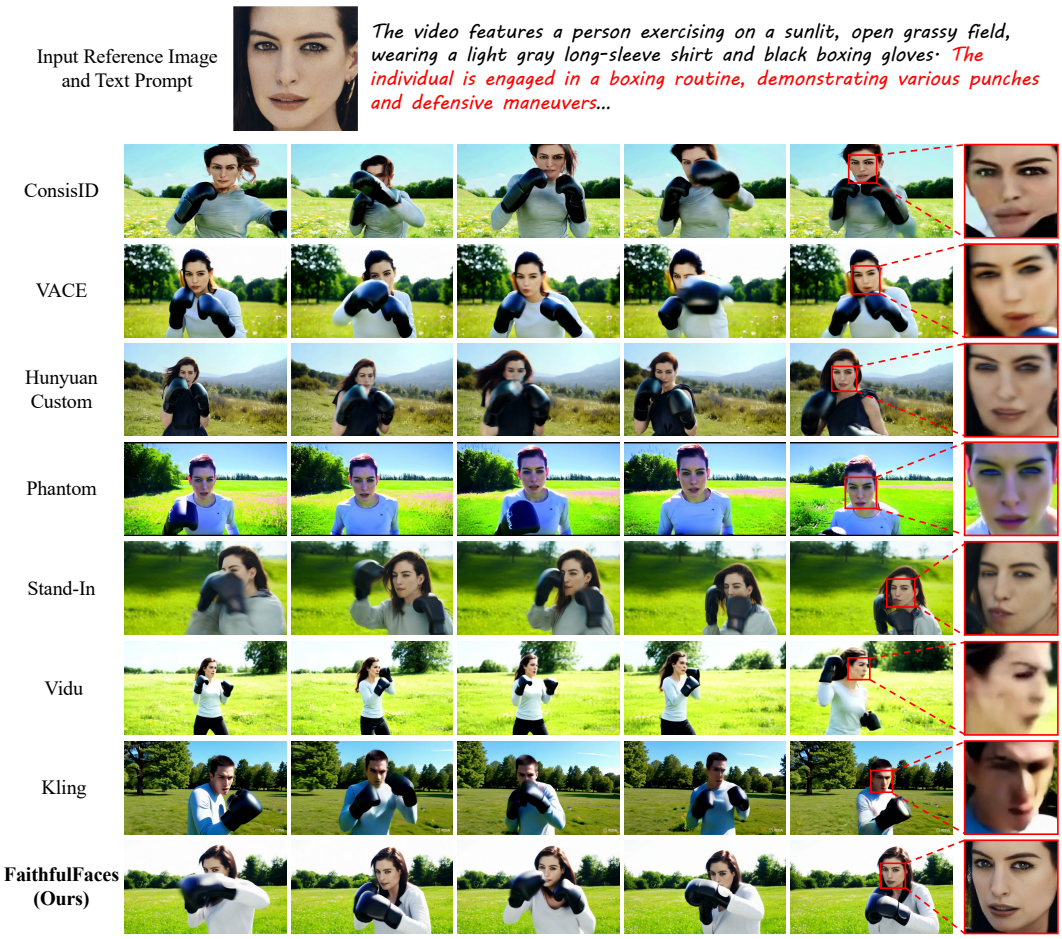


Figure 5: Visual comparisons of different methods. The goal is to generate a video of a person engaging in a boxing routine, characterized by diverse facial poses and instances of facial occlusion. In contrast to these state-of-the-art methods, FaithfulFaces produces a video of superior quality, exhibiting clear facial structures and consistent identity preservation.

isolating the improvements introduced by the proposed components from those arising from additional training data. The results are presented in Tab. 2, from which we draw the following conclusions: **i)** Identity Aligner is effective and yields substantial performance improvements, as it represents global facial pose information from the input single-view reference image, thereby enhancing the identity consistency of the generated videos. **ii)** The inclusion of Euler Embedding yields further improvements, confirming the feasibility and effectiveness of explicitly injecting pose information. More ablation studies are provided in Appendix A.2, A.3, A.4, and A.5, including qualitative analysis, dictionary elements, pooling operation type, and identity features. Additionally, in Appendix A.7 and A.8, we investigate the robustness to non-frontal views and the identity aligner's sensitivity to Euler angle variations.

Table 2: Ablation study of the key components in FaithfulFaces.

Aligner	Euler	FaceSim-Cur ↑	FaceSim-Arc ↑	FID ↓	CLIPScore ↑
✓	✓	**0.568**	**0.542**	**164.24**	**33.93**
✓	×	0.522	0.497	173.57	33.31
×	×	0.437	0.414	186.82	32.05

Visualization of the Encoded Facial Identity. Fig. 6 visualizes the distribution of encoded facial identities under different settings. Specifically, we randomly select 7 videos with different identities that are not included in the training data, and sample 8 face images with different facial poses from

each video. All sampled face images are then encoded by the different methods, and the encoded features are projected into a 2D space by t-SNE [26]. We can observe that *FaithfulFaces w/o (Identity Aligner, Euler Embedding)* suffer from the collapse of facial identity due to the absence of global facial pose awareness. At the same time, *FaithfulFaces w/o (Euler Embedding)* alleviates the collapse of facial identity, but the discriminability of its facial identity representation remains limited. In contrast, *FaithfulFaces* shows promising identity separability, demonstrating its faithful facial identity and naturally enhancing the performance of IPT2V.

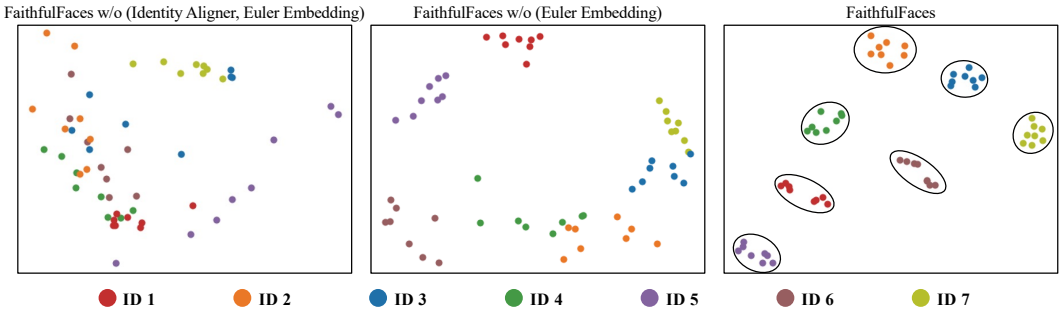


Figure 6: Visualization of the encoded facial identity (ID). FaithfulFaces demonstrates promising ID separability, indicating that its encoded identity representation exhibits high faithfulness and fidelity.

5 Conclusion

In this paper, we have proposed FaithfulFaces, a pose-faithful facial identity preservation learning framework for IPT2V. FaithfulFaces is motivated by the observation that existing methods often struggle to handle some intricate facial dynamic scenarios, largely due to their insufficient awareness of global facial pose. To encode the global facial pose representation from the input single-view face image, we propose a pose-shared identity aligner that refines and aligns distinct facial poses by a pose-shared dictionary and a pose variation–identity invariance constraint with Euler angle embedding learning. In particular, we construct a task-oriented, high-quality dataset with substantial facial pose diversity for robust training. Extensive experiments validate the effectiveness of FaithfulFaces.

Limitations. Despite its effectiveness, FaithfulFaces still has several limitations. First, our framework primarily focuses on single-subject IPT2V generation, and its applicability to multi-subject scenarios remains unexplored. Second, although the proposed pose-shared identity aligner improves identity preservation under substantial pose variations, extreme head rotations and severe facial occlusions may still lead to identity degradation. Finally, since our framework builds upon a pretrained video generation backbone, certain visual artifacts and motion inconsistencies inherited from the backbone may persist. Addressing these limitations, particularly multi-subject identity preservation and robustness under challenging facial dynamics, will be explored in future work.

Acknowledgements

This work was supported by the 2025 Tencent Rhino-bird Research Elite Program, the National Natural Science Foundation of China under Grant 62476133, and the Fundamental Research Funds for the Central Universities under Grant 11300-312200502507.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

[3] Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement. In *The Fourteenth International Conference on Learning Representations*, 2026.

[4] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660, 2022.

[5] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

[6] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

[7] 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. In *International Conference on Learning Representations (ICLR)*, 2024.

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

[9] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500. IEEE, 2022.

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

[11] 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. *Advances in neural information processing systems*, 30, 2017.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[13] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[14] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

[15] 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 *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[16] InsightFace. Insightface. `https://github.com/deepinsight/insightface`, 2025.

[17] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

[18] Jian Jin, Zhenbo Yu, Yang Shen, Zhenyong Fu, and Jian Yang. Latexblend: Scaling multi-concept customized generation with latent textual blending. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23585–23594. IEEE, 2025.

[19] Jian Jin, Zhenbo Yu, Yang Shen, Zhenyong Fu, and Jian Yang. Latent textual space: Pivot for scaling customized text-to-image and video generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[20] Kling. Kling. `https://klingai.com/`, 2026.

[21] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[22] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7752–7762. IEEE, 2025.

[23] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[24] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025.

[25] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[29] Pika. Pika. `https://pika.art/`, 2025.

[30] 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.

[31] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[34] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[35] Vidu. Vidu. `https://www.vidu.com`, 2026.

[36] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[37] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21968–21977. IEEE, 2023.

[38] Yuanzhi Wang, Zhen Cui, Mengyi Liu, Xiaoya Zhang, and Zechao Li. Incomplete multimodal probability flow recovery for emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[39] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

[40] Yuanzhi Wang, Yong Li, Mengyi Liu, Xiaoya Zhang, Zhen Cui, and Jian Yang. Training-free controllable text-guided video editing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[41] Yuanzhi Wang, Yong Li, Mengyi Liu, Xiaoya Zhang, Xin Liu, Zhen Cui, and Antoni B Chan. Re-attentional controllable video diffusion editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8123–8131, 2025.

[42] Yuanzhi Wang, Yong Li, Xiaoya Zhang, Xin Liu, Anbo Dai, Antoni B Chan, and Zhen Cui. Edit temporal-consistent videos with image diffusion model. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(12):1–16, 2024.

[43] Yuanzhi Wang, Xuhua Ren, Jiaxiang Cheng, Bing Ma, Kai Yu, Tianxiang Zheng, Qinglin Lu, and Zhen Cui. Implicit preference alignment for human image animation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 129632–129651. PMLR, 2026.

[44] Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. Stand-in: A lightweight and plug-and-play identity control for video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2026.

[45] 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. In *The Thirteenth International Conference on Learning Representations*, 2025.

[46] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025.

[47] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magic mirror: Id-preserved video generation in video diffusion transformers. *arXiv preprint arXiv:2501.03931*, 2025.

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

[49] Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-id: Towards universal identity-preserving video synthesis. *arXiv preprint arXiv:2503.14151*, 2025.

A Appendix

A.1 Observations in Learned Dictionary

To explicitly observe the learned pose-shared dictionary, we visualize the activations of the dictionary for five representative facial poses in Fig. 7. Specifically, we screen all face images of five facial poses in our dataset based on Euler angles and feed them into the pose-shared identity aligner to calculate the dictionary weight corresponding to each face image. We then record the indices of the top-10 elements in each weight vector, representing the most prominently activated dictionary elements for each face image. From the results presented in Fig. 7, we observe that particular dictionary elements are consistently activated by face images with similar poses. For example, the *frontal pose* tends to activate dictionary elements with indices 3, 562, and 2806, whereas the *upward-looking pose* frequently activates those with indices 2, 704, and 1856. This observation demonstrates that the learned dictionary captures meaningful patterns, potentially enabling robust representations of faces exhibiting a wide range of poses.

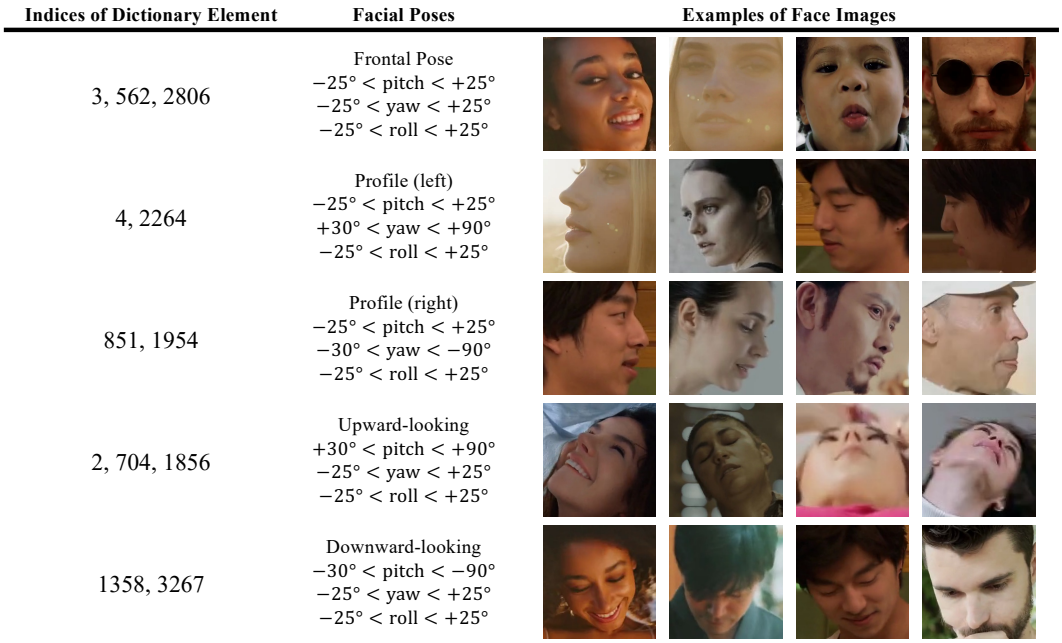


Figure 7: Visualization of learned pose-shared dictionary. We visualize five representative facial poses and observe that similar poses tend to frequently activate particular dictionary elements.

A.2 Exploring the Effects of Different Dictionary Elements

In this section, we conduct the ablation study to explore the effects of different dictionary elements in $\mathbf{D}$. Specifically, we conduct six sets of experiments in which the number of dictionary elements is set to 1024, 2048, 4096, 8192, 16384, and 32768, respectively. Fig. 8 illustrates the performance of our FaithfulFaces with various dictionary elements under two FaceSim metrics, and we can observe that the best performance is reached when the number of dictionary elements is set to 4096. Subsequently, the performance exhibits only slight variations as the number of dictionary elements increases further. In addition, the quantitative results of different dictionary elements are listed in Tab. 3. Hence, we use 4096 dictionary elements in our experiments.

A.3 Qualitative Analysis of Ablation Study for Key Components

In addition to the quantitative results of the ablation study presented in Tab. 2, we provide a qualitative analysis with visualizations in Fig. 9. We can observe that *Ours w/o (Identity Aligner, Euler Embedding)* shows obvious distortion of facial structures and facial details due to the lack of global facial pose awareness. *Ours w/o Euler Embedding* mitigates facial distortions owing to the global facial

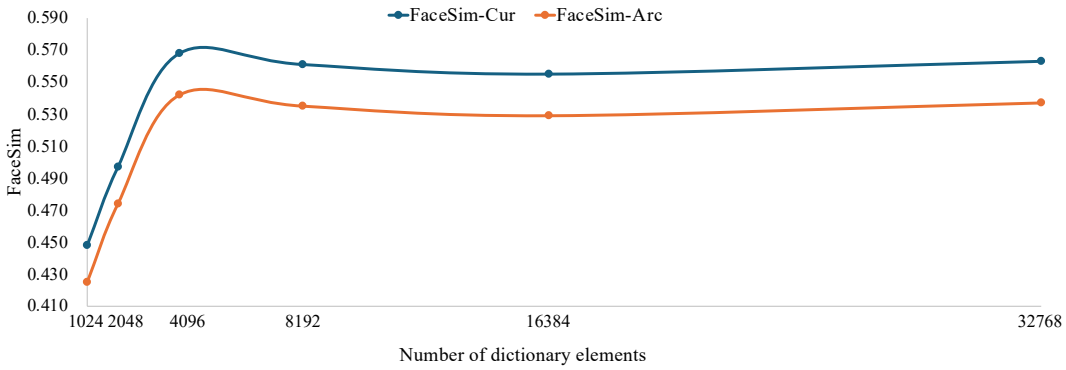


Figure 8: Ablation study on different dictionary elements.

Table 3: Quantitative results of different dictionary elements.

Dictionary Elements	FaceSim-Cur ↑	FaceSim-Arc ↑	FID ↓	CLIPScore ↑
1024	0.448	0.425	183.39	32.28
2048	0.497	0.474	176.80	32.89
4096	**0.568**	**0.542**	**164.24**	**33.93**
8192	0.561	0.535	165.94	33.82
16384	0.555	0.529	166.64	33.83
32768	0.563	0.537	165.96	33.82

pose representation provided by our pose-shared identity aligner; however, its identity consistency and facial stability remain suboptimal. In contrast, *Ours* generates high-quality results with clear facial structures and consistent identity details.



Figure 9: Visual comparisons of the ablation study for key components in FaithfulFaces.

A.4 Ablation Study of Pooling Operation Type

We evaluate the effects of different pooling operation types in the pose-shared identity aligner, and the results are listed in Tab. 4. From these results, we can observe that performance is optimal when using the max pooling operation. We also analyzed the potential underlying reason: the majority of information in face images of the same identity across different poses is similar or redundant. Therefore, using max pooling can alleviate a large amount of redundant information and extract highly abstract pose variations.

Table 4: Quantitative results of different pooling operation types.

Pooling Type	FaceSim-Cur ↑	FaceSim-Arc ↑	FID ↓	CLIPScore ↑
Sum Pooling	0.444	0.421	185.40	32.17
Mean Pooling	0.559	0.533	165.64	33.84
Max Pooling	**0.568**	**0.542**	**164.24**	**33.93**

A.5 Ablation Study of Different Identity Features

We conduct the ablation study by replacing the aligner features with the ArcFace [2] feature and the CLIP [31] feature to measure the aligner's specific effect. The experimental results are listed in Tab. 5. We can observe that performance significantly deteriorates when using the ArcFace feature or the CLIP feature. The underlying reason is **i)** ArcFace feature is unable to represent global facial pose information; and **ii)** CLIP features inherently encode the alignment between text and image modalities, rather than being dedicated to representing facial identity. This ablation study further demonstrates the effectiveness of the pose-shared identity aligner.

Table 5: Quantitative results of different identity features.

Feature Type	FaceSim-Cur ↑	FaceSim-Arc ↑	FID ↓	CLIPScore ↑
CLIP	0.447	0.422	183.57	32.13
ArcFace	0.475	0.453	177.20	32.70
Aligner (Ours)	**0.568**	**0.542**	**164.24**	**33.93**

A.6 Stability of Contrastive Loss in Optimization Procedure

Fig. 10 illustrates the value of the contrastive loss in the pose-shared identity aligner from different training steps. We can observe that the loss value gradually decreases during training and eventually converges to approximately 0.2. These results demonstrate the stability and convergence of contrastive learning for the pose-shared identity aligner during training.

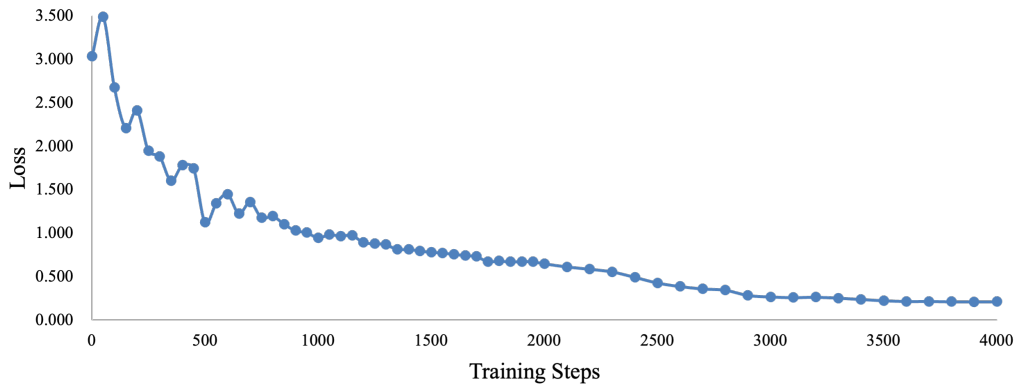


Figure 10: The value of the contrastive loss in the pose-shared identity aligner (named aligner loss) during the training process.

A.7 Discussion on Non-frontal View Robustness

In this section, we discuss the robustness of the method when the input reference image is a non-frontal view. Specifically, we collect 10 identities with both frontal and non-frontal face images for ablation and comparative experiments (the strongest baseline Phantom [24] as the representative). The results are shown in the Tab. 6, we can observe that both Ours w/o Aligner and the strongest baseline Phantom suffer a severe performance decrease exceeding 50% when non-frontal face images are used as input. In contrast, our method can control the performance decrease within 25%. These results

provide strong evidence that the pose-shared identity aligner can mitigate performance degradation in non-frontal face scenarios.

Additionally, we further provide visualization results of the frontal view and the non-frontal view in Fig. 11. We can further observe that when using a non-frontal image as input, the faces generated by the Phantom and Ours w/o Aligner completely collapse. In contrast, our method still maintains identity consistency. These visual results once again demonstrate that our method is capable of improving identity consistency in non-frontal face scenarios.

Table 6: Quantitative results of frontal view and non-frontal view. The values reported in each cell denote FaceSim-Cur/ FaceSim-Arc

Methods	Frontal	Non-frontal
Phantom [24]	0.470 / 0.435	0.231 (↓ 50.85%) / 0.216 (↓ 50.34%)
Ours w/o Aligner	0.448 / 0.423	0.202 (↓ 54.91%) / 0.197 (↓ 53.43%)
Ours	0.544 / 0.522	0.409 (↓ 24.82%) / 0.392 (↓ 24.90%)

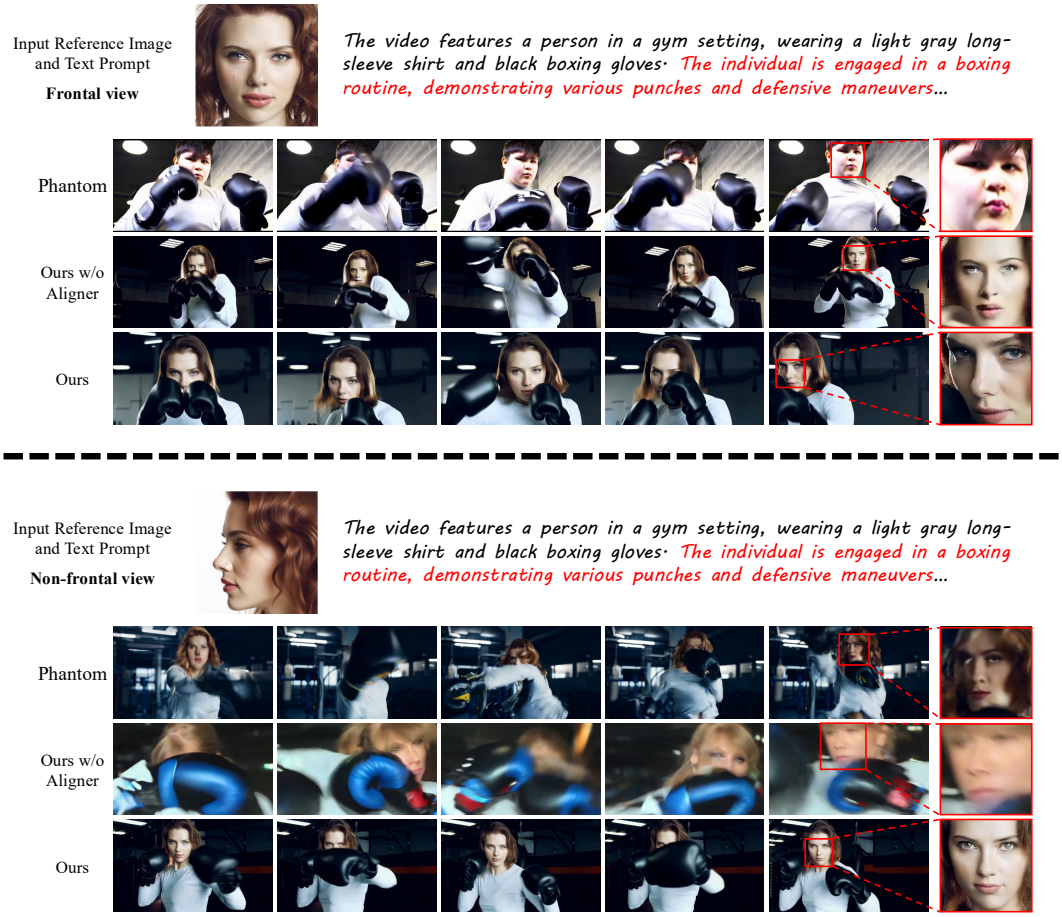


Figure 11: Visual comparisons of frontal view and non-frontal view. We can observe that when using a non-frontal image as input, the faces generated by the SOTA method Phantom and the baseline method (Ours w/o Aligner) completely collapse. In contrast, our method still maintains identity consistency.

A.8 Discussion on the Robustness of Identity Aligner for Euler Angles

In our pose-shared identity aligner, the sparsity design of the dictionary representation mechanism inherently possesses a certain degree of noise tolerance. To demonstrate this, we conduct the ablation experiments involving Euler angle perturbations. Specifically, given four perturbation ranges for

Euler angles: $-5^\circ \sim +5^\circ$, $-10^\circ \sim +10^\circ$, $-15^\circ \sim +15^\circ$, $-20^\circ \sim +20^\circ$. We apply random perturbations within these ranges to the predicted Euler angles to observe the impact on performance. The experimental results are shown in the Tab. 7, we can observe that performance exhibits only minor variations within the perturbation range of $-15^\circ \sim +15^\circ$. Significant performance degradation occurs only when perturbations exceed 15° (representing substantial errors). These results demonstrate that our method exhibits high robustness to Euler angle errors within a certain range.

Table 7: Quantitative results under different Euler angle perturbations.

Perturbation Range	FaceSim-Cur ↑	FaceSim-Arc ↑	FID ↓	CLIPScore ↑
No Perturbation	0.568	0.542	164.24	33.93
$-5^\circ \sim +5^\circ$	0.566	0.540	164.54	33.89
$-10^\circ \sim +10^\circ$	0.557	0.531	166.06	33.77
$-15^\circ \sim +15^\circ$	0.552	0.526	167.07	33.74
$-20^\circ \sim +20^\circ$	0.523	0.499	172.54	33.56

A.9 Pose-Conditioned Identity Preservation Analysis

To further evaluate the effectiveness of FaithfulFaces in preserving facial identity under large pose variations, we conduct a pose-conditioned identity analysis. Specifically, we estimate the yaw angle of each generated frame with a valid face detection and group the frames into six 30° intervals spanning -90° to 90°. For each interval, we compute the average FaceSim-Cur and FaceSim-Arc scores between the generated faces and their corresponding reference identities. This evaluation directly measures identity preservation under different generated head poses without requiring ground-truth pose trajectories. As shown in Table 8, identity similarity generally decreases for all methods as the head pose approaches extreme profile views. Nevertheless, FaithfulFaces consistently achieves the highest FaceSim-Cur and FaceSim-Arc scores across all pose intervals. Notably, in the challenging $[-90^\circ, -60^\circ)$ interval, FaithfulFaces achieves scores of 0.466 and 0.448, substantially outperforming the strongest competing baseline, Phantom, which obtains 0.253 and 0.244, respectively. Similar improvements are observed for large positive yaw angles. These results demonstrate that the proposed pose-shared identity alignment effectively mitigates identity degradation under substantial head rotations, supporting more robust identity preservation across diverse viewing angles.

Table 8: Pose-conditioned identity preservation across different yaw intervals. Each entry reports FaceSim-Cur / FaceSim-Arc (↑). The best results are highlighted in bold.

Yaw Interval	VACE-14B	Phantom	MAGREF	**FaithfulFaces**
$[-90^\circ, -60^\circ)$	0.232/0.213	0.253/0.244	0.245/0.226	**0.466/0.448**
$[-60^\circ, -30^\circ)$	0.355/0.332	0.382/0.363	0.363/0.340	**0.571/0.546**
$[-30^\circ, 0^\circ)$	0.471/0.446	0.502/0.476	0.484/0.458	**0.644/0.623**
$[0^\circ, 30^\circ)$	0.476/0.452	0.509/0.485	0.489/0.461	**0.651/0.624**
$[30^\circ, 60^\circ)$	0.363/0.341	0.392/0.374	0.371/0.350	**0.575/0.554**
$[60^\circ, 90^\circ]$	0.231/0.216	0.263/0.251	0.242/0.229	**0.473/0.452**

A.10 Effectiveness of the Pose-shared Dictionary

To investigate the necessity of the proposed pose-shared dictionary, we consider a straightforward alternative that directly concatenates facial tokens with Euler angle embeddings and injects them into the video generation backbone, without dictionary-based aggregation. Both variants adopt the same backbone, training data, and optimization settings for a controlled comparison. As shown in Tab. 9, removing the dictionary substantially degrades performance across all metrics. Specifically, the full model improves FaceSim-Cur and FaceSim-Arc from 0.445 and 0.426 to 0.568 and 0.542, respectively, while reducing FID from 182.95 to 164.24. These improvements suggest that directly injecting pose-conditioned facial features is insufficient for effective identity preservation, as the features may still contain pose-specific and redundant information. In contrast, the proposed dictionary provides learnable shared bases to aggregate identity-relevant features across different poses, thereby facilitating more discriminative and pose-robust identity representations.

Table 9: Ablation study on the effectiveness of the pose-shared dictionary.

Method	FaceSim-Cur ↑	FaceSim-Arc ↑	FID ↓	CLIPScore ↑
Ours w/o Dictionary	0.445	0.426	182.95	32.14
Ours	**0.568**	**0.542**	**164.24**	**33.93**

A.11 Video-Level Quality Evaluation

To further assess the temporal consistency and motion quality of generated videos, we conduct additional evaluations using three VBench metrics: Subject Consistency, Motion Smoothness, and Dynamic Degree. As shown in Tab. 10, FaithfulFaces consistently achieves the best performance across all three metrics. These results indicate that FaithfulFaces not only enhances facial identity preservation, as demonstrated by the identity similarity metrics, but also maintains strong video-level consistency and smooth motion dynamics.

Table 10: Video-level quality evaluation using VBench metrics. Higher values indicate better performance.

Method	Subject Consistency ↑	Motion Smoothness ↑	Dynamic Degree ↑
VACE-14B	95.80%	96.69%	97.60%
Phantom	96.34%	97.07%	97.22%
MAGREF	95.92%	96.92%	98.33%
FaithfulFaces	**98.33%**	**97.99%**	**98.61%**

A.12 Human Preference Study

To further assess the perceptual quality of generated videos beyond automated metrics, we conduct a human preference study involving 10 participants and 30 sets of generated videos. Each set includes videos generated by FaithfulFaces and three representative baselines, i.e., VACE, Phantom, and MAGREF. For each pairwise comparison, participants are presented with the text prompt, reference image, and two generated videos (ours and a baseline) displayed side by side in randomized order. Participants are asked to select the preferred video by considering three aspects: (1) *Identity Preservation*, assessing the consistency of the generated facial identity with the reference image; (2) *Visual Quality*, evaluating the overall perceptual quality of the generated video; and (3) *Text Alignment*, measuring how well the generated content follows the text prompt. The overall pairwise preference results are reported in Tab. 11.

Table 11: Human preference study.

Comparison	Ours Win (%)	Ours Lose (%)
Ours vs. VACE	**81.3**	18.7
Ours vs. Phantom	**84.7**	15.3
Ours vs. MAGREF	**78.3**	21.7

A.13 Implementation Details of the Pose-shared Identity Aligner

Face Feature Encoding. We first detect the facial region using InsightFace and resize the cropped face while preserving its aspect ratio, followed by padding to a resolution of 480×832. Given a reference face tensor of shape $B \times 3 \times 1 \times 480 \times 832$, the pretrained Wan VAE encodes it into a latent representation of shape $B \times 16 \times 1 \times 60 \times 104$. We then reuse the DiT patch-embedding layer, implemented as a 3D convolution with kernel size and stride $(1, 2, 2)$, to obtain a feature tensor of shape $B \times 5120 \times 1 \times 30 \times 52$. Flattening the spatiotemporal dimensions produces 1,560 facial tokens, each with 5,120 dimensions, resulting in a tensor of shape $B \times 1560 \times 5120$.

Euler Pose Embeddings. We estimate the pitch, yaw, and roll angles using 6DRepNet. Each Euler angle is independently encoded into a 5,120-dimensional sinusoidal embedding and processed by a shared two-layer MLP. The resulting three embeddings are concatenated into a 15,360-dimensional vector and subsequently projected back to 5,120 dimensions. This pose embedding is broadcast along the token dimension and added to the 1,560 facial tokens before dictionary-based aggregation.

Pose-shared Dictionary Aggregation. The pose-conditioned facial tokens interact with a learnable dictionary containing 4,096 vectors, each of dimension 5,120. This produces a similarity score tensor of shape $B \times 1560 \times 4096$. We apply max pooling over the 1,560 facial tokens to obtain dictionary-level aggregation weights. A weighted combination of the dictionary vectors then yields a single identity representation of shape $B \times 5120$. Since its feature dimension matches the hidden dimension of VACE-14B, no additional output projection is required.

Identity Conditioning Interface. Importantly, the resulting identity representation is not directly added to the raw noisy video latent. Instead, the noisy latent is first processed by the DiT patch-embedding layer and flattened into a video token sequence $\mathbf{X}_t \in \mathbb{R}^{B \times L_t \times 5120}$, where L_t denotes the number of spatiotemporal video tokens. The identity representation $\mathbf{e}_{id} \in \mathbb{R}^{B \times 5120}$ is first reshaped into a single identity token of shape $B \times 1 \times 5120$. We then zero-pad this token along the sequence dimension to match the video token length L_t, obtaining $\mathbf{E}_{id} \in \mathbb{R}^{B \times L_t \times 5120}$. The identity-conditioned video tokens are computed as

$$\widetilde{\mathbf{X}}_t = \mathbf{X}_t + \mathbf{E}_{id}. \tag{9}$$

Consequently, the identity representation is injected into the first video token before the DiT Transformer blocks, while subsequent self-attention operations propagate identity information across the spatiotemporal sequence. This reshape-and-padding operation ensures dimensional compatibility without modifying the original DiT architecture.

A.14 Prompt Construction

We now elaborate on the construction of challenging test text prompts designed to drive models to generate videos exhibiting significant facial pose variations, expression changes, and facial occlusions across diverse scenarios. For character movement, we select several representative scenes: 1) **Boxing** with facial pose variations and facial occlusions; 2) **Head shaking and dancing** with facial pose variations; 3) The character **transitions from having their back to the camera to facing the camera**; 4) **Ballet** with facial pose variations; 5) **Speech** with facial pose variations and expression changes; 6) Some descriptions used to generate **dramatic changes in facial expressions and poses**, as shown in the second case in Fig. 16.

With these basic movement scenes, we use GPT-4.1[3] to generate information-rich text prompts. Taking the boxing scene as an example, the generated text prompt is: *"The video features a person in a gym setting, wearing a light gray long-sleeve shirt and black boxing gloves. The individual is engaged in a boxing routine, demonstrating various punches and defensive maneuvers. The camera closely follows the person's movements, keeping their face and gloves prominent in the frame, and capturing detailed facial expressions and dynamic action. The background is dimly lit, with overhead lights providing illumination. The gym environment is evident from the visible equipment and the industrial setting, which adds to the intensity of the scene."* Subsequently, based on this, we employ GPT-4.1 to generate text prompts for different background scenarios, such as *"open grassy field"* in Fig. 5, *"urban street setting"* in Fig. 18. Ultimately, we curate 20 high-quality text prompts that span various actions and background scenarios.

A.15 Ethics Statement and Broader Impact

Advancements in identity-preserving text-to-video generation technology are poised to support and empower the creative processes of artists and designers. FaithfulFaces is capable of generating high-quality, realistic human videos. However, it also raises concerns regarding misinformation, potentially undermining the reliability of video content. Additionally, this technology could be misused to generate deceptive content for fraudulent purposes. It is important to recognize that any technology is susceptible to misuse. Nevertheless, it is feasible to train a classifier that can distinguish between real and FaithfulFaces-generated videos based on their texture features.

Regarding data ethics, our training data are curated from OpenHumanVid, obtained through its official application-and-review process. To reduce potential privacy and redistribution risks, we do not release the original training videos, extracted facial images, or the processed training dataset.

Responsible Release. To facilitate reproducibility, we plan to release the implementation code of FaithfulFaces. However, we will not redistribute the original training videos, extracted facial images,

[3]https://openai.com/index/gpt-4-1/

or processed training dataset due to potential privacy and licensing concerns. The public release of trained model weights will be subject to further review of data licensing, privacy implications, and potential misuse risks. We do not currently commit to providing a public inference API.

A.16 Reproducibility Statement

First, we have explained the implementation of FaithfulFaces in detail in Sec. 4.1. Second, we have explained the details of training and inference in Fig. 2 and Sec. 3.2. Third, we have explained the details of the dataset construction in Sec. 3.4. Finally, we plan to release the implementation code of FaithfulFaces to facilitate reproducibility, while the training data will not be redistributed due to privacy and licensing considerations.

A.17 The Use of Large Language Models

This submission utilizes a large language model for grammar checking.

A.18 More Visualization Results

In this section, we provide more visual comparisons of different methods in Figs. 12, 13, 14, and 15 to demonstrate the effectiveness of our method. Additionally, we provide more showcases of identity-preserving videos generated by our FaithfulFaces in Figs. 16, 17, 18, and 19, covering a variety of identities, actions, and scenes.

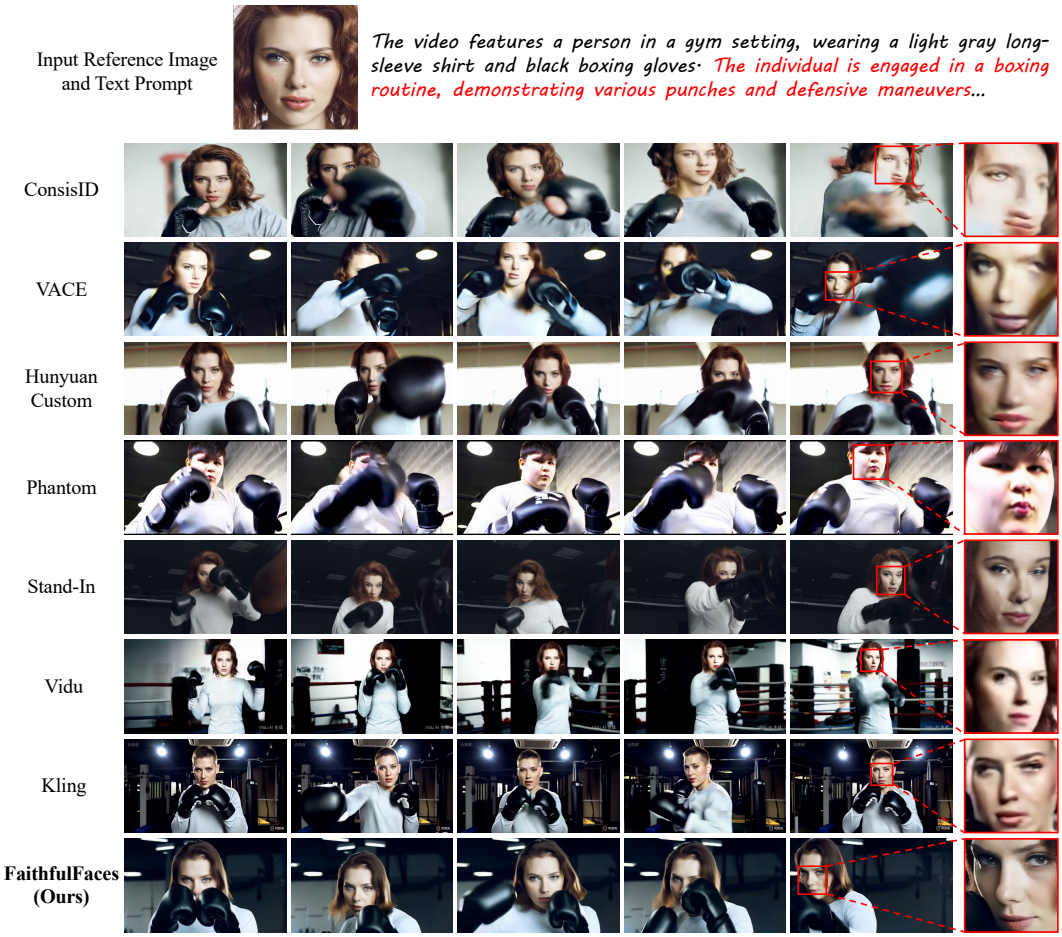


Figure 12: Complete visual comparisons of different methods for the case of Fig. 1.

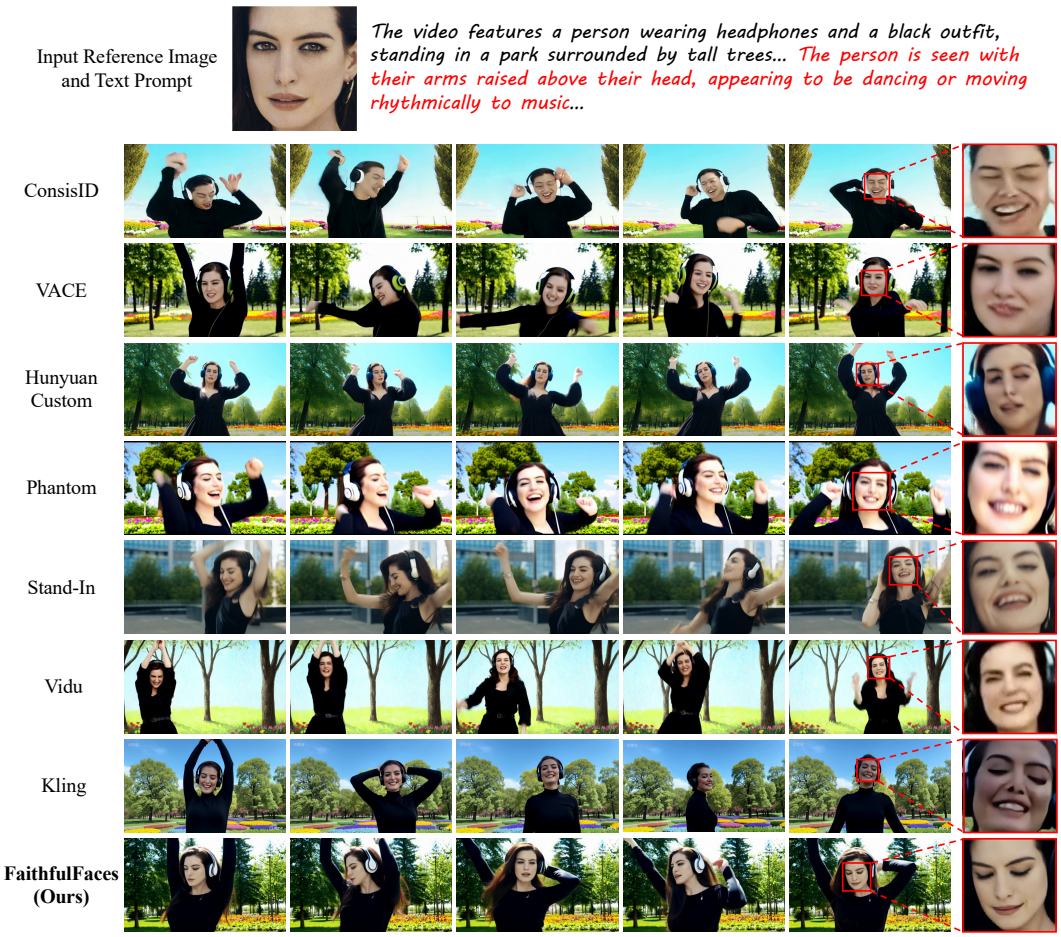


Figure 13: More visual comparisons of different methods.

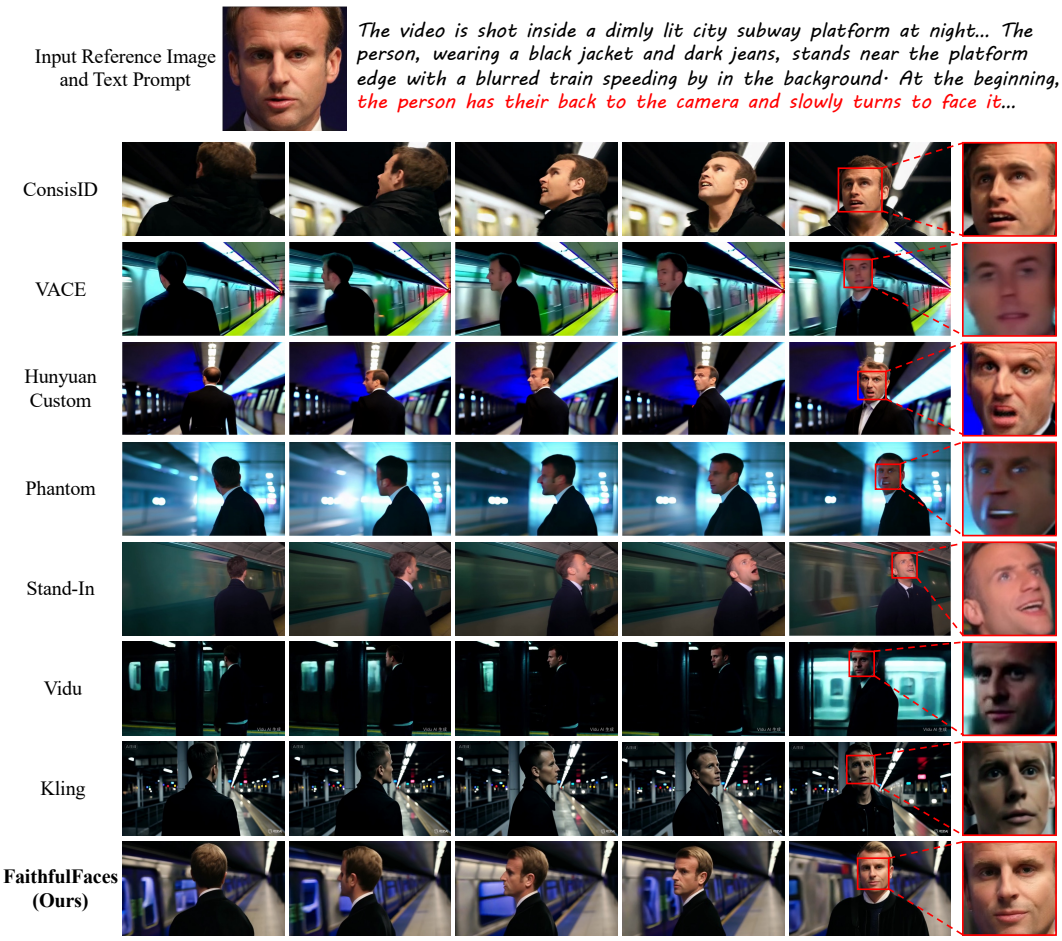


Figure 14: More visual comparisons of different methods.



Figure 15: More visual comparisons of different methods.



Figure 16: More showcases of identity-preserving videos generated by our FaithfulFaces.



Figure 17: More showcases of identity-preserving videos generated by our FaithfulFaces.



Figure 18: More showcases of identity-preserving videos generated by our FaithfulFaces.



Figure 19: More showcases of identity-preserving videos generated by our FaithfulFaces.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The key contributions claimed in the abstract and introduction are elaborated in Sec. 3.3 and Sec. 3.4. We provide extensive quantitative and qualitative experimental results to verify the effectiveness of these contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Sec. 5, Conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: First, we have explained the implementation of FaithfulFaces in detail in Sec. 4.1. Second, we have explained the details of training and inference in Fig. 2 and Sec. 3.2. Third, we have explained the details of the dataset construction in Sec. 3.4. In addition, the codes will be publicly available when the paper is accepted.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: No code is included in this submission.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the implementation details in Sec. 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: While standard error bars are computationally prohibitive for large-scale video generation tasks, we ensure statistical significance by evaluating our method across a comprehensive test set of 600 diverse video generations. We use robust, standardized metrics (FaceSim, FID, CLIPScore) over this entire set to reliably demonstrate our performance gains.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4.1 provides the type of compute workers GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the Broader Impact in Appendix A.15.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not publicly release the training dataset or trained model weights, nor do we provide a public inference API. Potential misuse risks are discussed in Appendix A.15.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used in this paper are properly credited in Sec. 4.1.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This submission only utilizes LLMs for grammar checking.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.