EXPRESSION-DIVERSE REFERENCES FOR IDENTITY-PRESERVING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Identity-preserving video generation aims to maintain a subject's identity while synthesizing realistic videos. Yet a single reference portrait captures the subject's appearance under only one facial configuration. As expressions change, facial appearance can vary in highly identity-specific ways, leaving the subject's appearance under unseen expressions underdetermined by the reference alone. This expression-dependent variation also complicates evaluation: similarity to a neutral reference may decrease under strong expressions even for real images of the same person. We investigate this limitation from both generation and evaluation perspectives. First, we quantify how face-recognition similarity varies with expression intensity using controlled photographs and MEAD videos. We then construct a compact yet expressive reference gallery that captures diverse expression-dependent facial configurations. Matching against this gallery provides a more robust measure of identity similarity under expressive motion. To further expose performance degradation with expression intensity, we report identity similarity separately for mild, intense, and extreme expressions. For generation, we extend Stand-In to condition on our expression-diverse reference sets and develop a data-curation pipeline that extracts consistent yet diverse face crops from training videos. In practical settings where only a single portrait is available, we construct the reference set by synthesizing additional expressions with a pretrained facial reenactment model. On our controlled benchmark of 10 subjects and 48 prompts per subject, both real and synthesized reference sets outperform the evaluated baselines in identity similarity across all three expression-intensity regimes, with the largest improvements for extreme expressions. In this regime, nearest-gallery CurricularFace similarity increases from 60.95 for the strongest baseline to 74.80 with real reference sets and 68.36 with synthesized reference sets. These results show that representing an identity with an expression-diverse reference gallery substantially improves identity preservation under expressive motion.

1 INTRODUCTION

Recent text-to-video models (Wan et al., 2025; Yang et al., 2024; Gao et al., 2025) can synthesize realistic and diverse videos from natural-language prompts. Identity-preserving video generation extends this capability by introducing visual references to control the identity of the generated subject while following the text prompt. Recent methods build on pretrained video generators and incorporate reference-conditioning mechanisms to preserve the target identity throughout generation (Wang et al., 2026; Xue et al., 2026). Strong facial expressions pose a particular challenge: a person should remain recognizable even as their eyes, mouth, cheeks, and skin undergo substantial changes. A single portrait, as adopted by Zhong et al. (2025); Yuan et al. (2025b), captures only one facial configuration and provides limited evidence of these person-specific appearance changes (Figure 1).

This limitation affects both generation and evaluation. In generation, single-reference methods can lose identity fidelity as expression intensity (the degree to which a facial expression deviates from neutral) increases (Figure 6). In evaluation, comparing each generated face only with a neutral reference can conflate expression variation with identity error. We observe this effect even in real images: across 70 non-neutral photographs of 10 subjects, CurricularFace (Huang et al., 2020) similarity to each subject's neutral reference decreases with expression intensity (Spearman $\rho = -0.701$) (Figure 3a). We observe a similar trend on a larger-scale video dataset, MEAD (Wang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2020) (Table 1). Thus, single-reference similarity can be a suboptimal metric for identity consistency under strong expression changes.



Subject 1 Neutral | Subject 1 Expression 2 | Subject 2 Neutral | Subject 2 Expression 2

Figure 1: Expression-induced appearance changes are person-specific. Given the same expression instructions, the two subjects exhibit different changes in gaze, brow, and mouth geometry, illustrating information absent from a neutral reference.

We study expression-diverse references as a shared basis for conditioning and evaluation. For evaluation, we collect a standardized gallery of eight facial configurations per subject, comprising one neutral portrait and seven expressions (Figure 2). We score each generated face by its maximum embedding cosine similarity to the subject's gallery, which provides a more expression-robust face similarity metric. We additionally report identity similarity separately for mild, intense, and extreme expressions, together with the fraction of generated frames in each regime. This makes failures at high expression intensity visible and helps interpret scores when methods produce different expression distributions.

For generation, we extend Stand-In (Xue et al., 2026) to condition on multiple references of the same person. The model is an efficient 315M-parameter LoRA adapter with only 25k training steps at batch size 16. Expression-diverse references provide complementary observations of the subject's facial appearance that are absent from a single portrait. To train the model, we curate consistent and diverse face crops from OpenHumanVid (Li et al., 2024) and vary the number of conditioning references during training. When multiple captures are unavailable, we use LivePortrait (Guo et al., 2024) to synthesize expressive references from a single neutral portrait, using driving expressions from a subject outside the evaluation set. This allows the same video generator to operate from a single portrait. Although identity similarity is lower than with real references, it still exceeds that of all evaluated baselines.

We evaluate on a controlled benchmark containing 10 subjects and 48 expression-focused prompts per subject. Both real and synthetic reference galleries achieve higher nearest-gallery identity similarity than all evaluated baselines in each of the three intensity regimes. Under extreme expressions, CurricularFace similarity, is 74.80 with real galleries and 68.36 with synthetic galleries, compared with 60.95 for the strongest baseline (Table 2). Ablations compare expression-diverse references with repeated neutral portraits and pose-diverse neutral references, examining the role of expression coverage beyond reference count (Table 5). Across both recognizers and all evaluated baselines, our real-gallery model wins on at least 97.9% of matched subject–prompt pairs. The win rate remains at least 96.1% on the subset where our model produces greater peak expression intensity, providing additional evidence that identity gains can persist under stronger facial deformation (Table 3).

Our contributions are threefold:

- We identify and quantify the dependence of single-reference face similarity on expression intensity in real imagery and introduce a standardized gallery-based multi-reference evaluation protocol stratified by expression intensity.
- We develop expression-diverse multi-reference conditioning and its data-curation pipeline, which enables the model to better capture person-specific facial deformations under expressive motion.

- We demonstrate improvements in identity similarity with both real expression reference galleries and reference galleries synthesized from a single portrait, both surpassing baselines in all expression regimes and especially the intense and extreme ones, even when multiple-expression captures are not available.

2 Related Work

Video Generation Models provide the generative backbone for human video synthesis. CogVideoX (Yang et al., 2024) combines a temporally compressive 3D VAE with an expert transformer to produce long, coherent text-conditioned clips. Wan2.2 (Wan Team, 2025) introduces a two-expert architecture specialized for high- and low-noise denoising stages, providing a backbone for multiple downstream tasks. LTX-2 (HaCohen et al., 2026) jointly models video and audio, while MiniMax H3 (MiniMax, 2026) accepts multimodal character and motion references for video generation and editing. These general-purpose models provide strong visual priors, but text conditioning lacks precise control over the subject's identity in generated videos, whereas image-to-video generation generally requires a complete initial frame rather than just a reference portrait.

Identity-Preserving Video Generation conditions pretrained video generation backbones on reference images of one or more identities. The most common approach is to use a single reference image per identity and inject its features into the video generator. ConsisID (Yuan et al., 2025b) decomposes a reference face into low-frequency global features and high-frequency local details and injects them at different depths of a diffusion transformer. Concat-ID (Zhong et al., 2025) instead concatenates VAE-encoded reference features with video latents, while Stand-In (Xue et al., 2026) adds a lightweight conditional image branch with restricted self-attention and conditional position mapping. Phantom (Liu et al., 2025) learns joint text–image injection from text–image–video triplets.

Benchmarking Video Generation Models for facial identity similarity typically relies on a single reference. VBench (Huang et al., 2024) evaluates video generation along 16 disentangled dimensions. Its subject-consistency metric compares DINO features across frames of a generated video, measuring temporal stability but not facial identity relative to a reference person. OpenS2V-Eval (Yuan et al., 2025a) more directly evaluates subject-to-video generation. For human identity, its FaceSim protocol detects faces and computes CurricularFace (Huang et al., 2020) cosine similarity to the input face. However, as we show below, this single-reference comparison can penalize genuine identity-preserving expression changes when the generated face departs from the reference expression. Our benchmark instead compares each frame with the nearest member of a standardized expression gallery and reports results stratified by expression intensity.

Facial Reenactment and Related Tasks. Facial reenactment animates a source portrait using head pose and expressions extracted from a driving video. LivePortrait (Guo et al., 2024) extracts appearance features from a single source image, estimates implicit 3D keypoints for the source and driving frames, warps the source features according to their motion, and decodes the result into animated frames. Person-specific 3D approaches such as GaussianAvatars (Qian et al., 2024) instead rig optimized 3D Gaussians to a parametric face model, and video face-swapping systems such as OmniFace (Guo et al., 2026) transfer a source identity onto an existing target video while preserving its motion and scene. These tasks receive an explicit driving signal and are distinct from our text-driven generation setting.

3 Evaluating Identity Fidelity at High Expression Intensity

Previous work evaluates facial identity by comparing a generated face with a single reference image of the same subject in the embedding space of a pretrained face-recognition model. However, we show that the face similarity score is entangled with the expression variance. We collect a gallery of reference images $\mathcal{G} = \{I_{i,e}\}$ for 10 subjects $i \in \{1, \dots, 10\}$ and 8 expressions $e \in \{0, \dots, 7\}$, where $e = 0$ is the neutral expression. The images were taken in a light stage with a fixed frontal camera position and flat lighting setup, as shown in Figure 2. For each subject i, we compare each non-neutral image ($e \neq 0$) with the neutral image ($e = 0$), computing cosine similarity separately with ArcFace (Deng et al., 2019) (s_{Arc}) and CurricularFace (Huang et al., 2020) (s_{Cur}). We measure expression intensity as $\Delta\Psi_{i,e} = \|\Psi_{i,e} - \Psi_{i,0}\|_2$, where $\Psi_{i,e}$ is estimated using SMIRK (Retsinas

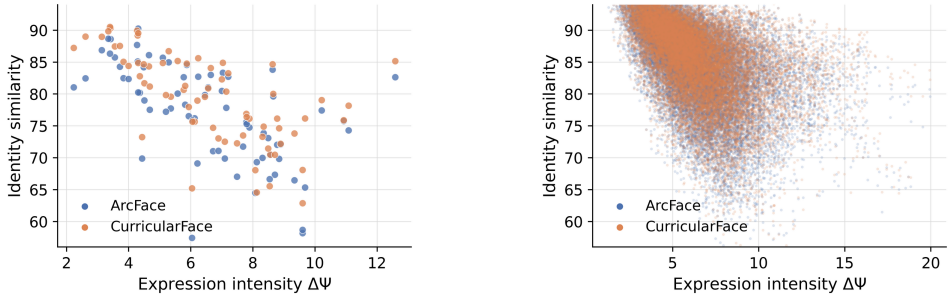
et al., 2024) (Appendix A.1). As shown in Figure 3, both identity metrics decrease as expression intensity increases, both in our collected dataset and in the publicly available MEAD dataset (Wang et al., 2020). To further quantify this, we report both Pearson's r and Spearman's ρ in Table 1.



Figure 2: Example gallery for one of the 10 subjects with 8 expressions including neutral. All subjects were asked to perform the same set of expressions, whose instructions are listed in Appendix C. The gallery captures diverse expression-dependent facial configurations to support more robust identity evaluation and provide richer information for video generation models.

Table 1: Correlation between expression intensity $\Delta\Psi$ and identity similarity on both datasets.

		ArcFace		CurricularFace	
Dataset	n	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
Lightstage	70	-0.587	-0.662	-0.636	-0.701
MEAD	24,190	-0.521	-0.590	-0.496	-0.567



(a) Lightstage (b) MEAD

Figure 3: ArcFace and CurricularFace similarity versus expression intensity on real images. Each point represents a non-neutral photograph (Lightstage) or video frame (MEAD), measured relative to the subject's neutral reference. MEAD preprocessing is described in Appendix A.2.

Therefore, for recognizer $R \in \{\mathrm{Arc}, \mathrm{Cur}\}$, we define the nearest-gallery identity similarity as $s_R^*(\hat{I}_i) = \max_{e\in\{0,\ldots,7\}} s_R(I_{i,e}, \hat{I}_i)$[1], where $\hat{I}_i$ is a frame generated for subject i. This provides a more robust measure of identity fidelity when the expression intensity is high.

Furthermore, we report the nearest-gallery identity similarities for three ranges of $\Delta\Psi$: mild ($\Delta\Psi < 4.5$), intense ($4.5 \leq \Delta\Psi < 8.5$), and extreme ($\Delta\Psi \geq 8.5$). As shown in Figure 4, the thresholds approximately correspond to the first and third quartiles of $\Delta\Psi$ across the 70 non-neutral real photographs, which already depict deliberately intense expressions. This places the majority of photographs in the intense regime while retaining similarly sized mild and extreme tails.

We combine the six basic emotions proposed by Ekman (1992) with four environments and two triggers to create 48 prompts per subject (Appendix E). The prompted expressions differ from the reference configurations to discourage direct copying.

[1]We report similarities on a 100-point scale, following the convention of Wang et al. (2026).

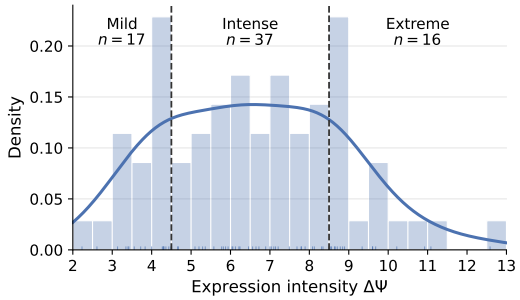
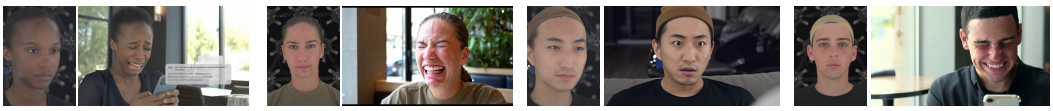


Figure 4: Distribution of expression intensity $\Delta\Psi$ across the 70 non-neutral light-stage photographs. The solid curve is a kernel density estimate of the density histogram, and marks along the horizontal axis denote individual photographs.

4 Training Multi-Reference Subject-to-Video Models

Baselines. We provide the neutral portrait to the single-reference subject-to-video baselines: ConsisID (Yuan et al., 2025b), Concat-ID (Zhong et al., 2025), Stand-In (Xue et al., 2026), and Phantom (Liu et al., 2025). For Gen-4.5 (Runway Research, 2025), we use image-to-video mode through the official API, with the text prompt and the neutral portrait as the initial frame. For AnyID (Wang et al., 2026), which accepts up to five references, we provide the neutral portrait and four other expression images (1, 3, 5, 7). We use official checkpoints for the locally run baselines. For Stand-In, we use its adapter for Wan 2.2.

Failure Mode of Existing Methods. We analyze identity preservation across all generated frames at different expression intensities. Figure 6 shows that the single-reference models exhibit a substantial drop in identity similarity when the expression intensity $\Delta\Psi$ is large despite high similarity at low $\Delta\Psi$. Some common artifacts for these models are presented in Figure 5. The multi-reference model AnyID (Wang et al., 2026) shows a more stable curve, benefiting from our expression-diverse references, but still has low identity similarity across all intensity regimes.



(a) Distorted mouth (b) Distorted glabella (c) Incorrect facial features

Figure 5: Representative facial artifacts produced by existing identity-preserving video generation methods under expressive motion. Each generated frame is shown beside the neutral reference portrait provided to the method.

Data Curation. We use the OpenHumanVid (Li et al., 2024) dataset for training our multi-reference subject-to-video models. For each video, we sample 10 evenly spaced frames and use SCRFD (Guo et al., 2021) to detect the face. We retain the largest detected face with a detection score above 0.5 and crop it using a bounding box expanded by a factor of 1.6. We compute ArcFace embeddings for the retained crops and cluster them by linking together any two crops with cosine similarity above 0.35. Videos with more than 2 clusters or more than 40% of the crops outside the dominant cluster are treated as multi-subject videos and are discarded, and we keep the largest cluster for the remaining videos. If more than eight crops remain, we select a diverse subset of eight. Starting from the highest-quality crop $i_0 = \arg\max_j Q(I_j)$, we greedily add the farthest remaining crop i_{k+1} until eight crops have been selected:

$$i_{k+1} = \arg\max_{j} \min_{j' \in \{0,1,\ldots,k\}} 0.75 \cdot (1 - s_{Arc}(I_j, I_{i_{j'}})) + 0.25 \cdot \Delta t(j, i_{j'})$$

where $Q(I_j)$ is the quality score, $s_{Arc}(I_j, I_{i_{j'}})$ is the ArcFace cosine similarity, and $\Delta t(j, i_{j'})$ is the temporal distance between the two frames, normalized to $[0, 1]$.

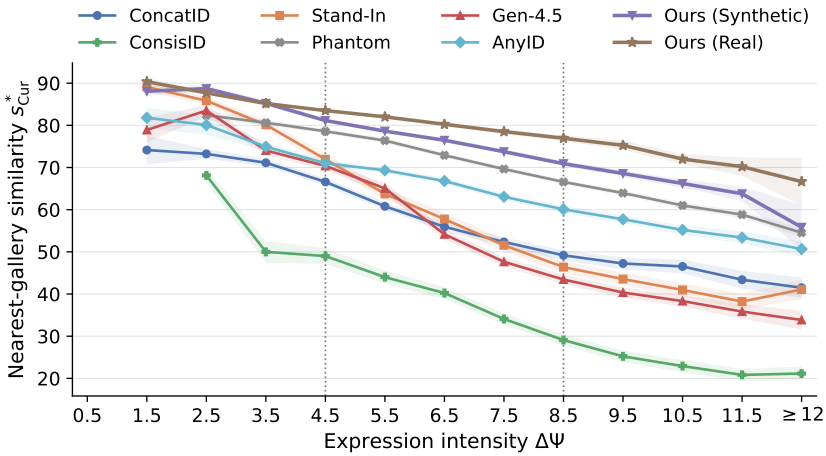


Figure 6: Distribution of nearest-gallery CurricularFace identity similarity in different bins of expression intensity $\Delta\Psi$. We use all generated frames from the baselines and our method, run SMIRK (Retsinas et al., 2024) to obtain the expression vector Ψ, and measure $\Delta\Psi = \|\Psi - \Psi_{\text{neutral}}\|_2$ against the neutral reference (see Appendix A.1 for details). Each point gives the mean identity similarity within a unit-width $\Delta\Psi$ bin, connected across bins; the final point aggregates the open-ended $\Delta\Psi \geq 12$ bucket. Shaded regions denote 95% confidence intervals computed using clip-clustered standard errors. Dotted lines mark the thresholds between mild, intense, and extreme expressions.

Architecture and Training. The pipeline for our method is shown in Figure 7. We build on Stand-In (Xue et al., 2026), a single-reference identity-preserving adapter for Wan2.2-T2V-A14B (Wan et al., 2025). It adds rank-128 LoRA adapters to the query, key, and value projections of each self-attention block, applied only to reference-image tokens, while the backbone parameters, including those used to process video tokens, remain frozen. We extend this architecture to accept multiple reference images. We independently resize and VAE-encode each of the N references at 512×512 resolution, yielding 1024 tokens per reference. All references share diffusion timestep 0 and the same grid of 3D RoPE (Su et al., 2021) coordinates at temporal index -1 and spatial indices starting beyond the video grid, namely $\{-1\} \times [H_V, H_V + H_I) \times [W_V, W_V + W_I)$, whereas the video tokens are placed on $[0, F) \times [0, H_V) \times [0, W_V)$, where $F = 21$ is the number of latent frames, compressed by the VAE from 81 frames. Reference tokens attend only within their own image, while video tokens attend to all tokens. The attention mask over the concatenated tokens is

$$M_{Q,K}(i,j) = \begin{cases} 1 & \text{if } i \text{ is a video token} \\ 1 & \text{if } i, j \text{ are image tokens from the same reference image} \\ 0 & \text{otherwise.} \end{cases}$$

The reference tokens do not attend to the text prompt in the cross-attention layers, and are discarded before untokenization and VAE decoding. We train with 400k randomly selected unique clips. During training, we shuffle the references. At each step, with probability 0.5, we use all N available images; otherwise, we draw k uniformly from $\{1, \ldots, N\}$ and randomly select a subset of size k. This improves robustness to different reference counts. We use an effective batch size of 16.

Inference Pipeline with a Single Reference. Our model provides the best results when multiple references are provided, but when only a single reference is available, we use LivePortrait (Guo et al., 2024) to synthesize seven expressive references, driven by expressions from a held-out subject. The portrait and synthesized images form an eight-image gallery (see example in Figure 9, Appendix D).

5 Results

5.1 Qualitative Results

In Figure 8, we show videos generated by our model using a real reference gallery, *Ours (Real)*, or a LivePortrait-generated gallery, *Ours (Synthetic)*, alongside the representative baselines Stand-In and

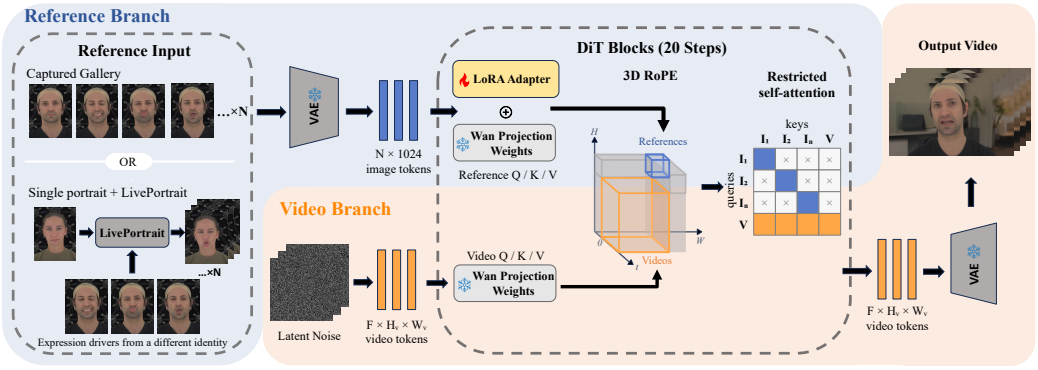


Figure 7: Overview of our multi-reference subject-to-video generation pipeline.

AnyID. With information from a diverse set of expressions, our model produces frames that better preserve identity under strong expressions.

5.2 Results on Our Evaluation Dataset

We evaluate the baselines and our models on the evaluation dataset introduced above. Table 2 shows the nearest-gallery identity similarities s^*_{Arc} and s^*_{Cur}, stratified by expression intensity. Our model, whether using synthetic or real reference galleries, outperforms all baselines in all three expression-intensity regimes. The performance gap is especially pronounced in the intense and extreme regimes, demonstrating the effectiveness of expression-diverse multi-reference conditioning in preserving identity under expressive motion.

Table 2: Identity fidelity on our evaluation dataset. The percentage columns report the fraction of each model's frames in each expression-intensity regime.

	All frames		Mild			Intense			Extreme		
Model	$s^*_{\mathrm{Arc}}\uparrow$	$s^*_{\mathrm{Cur}}\uparrow$	%	$s^*_{\mathrm{Arc}}\uparrow$	$s^*_{\mathrm{Cur}}\uparrow$	%	$s^*_{\mathrm{Arc}}\uparrow$	$s^*_{\mathrm{Cur}}\uparrow$	%	$s^*_{\mathrm{Arc}}\uparrow$	$s^*_{\mathrm{Cur}}\uparrow$
Concat-ID [ICCVW2025]	59.31	58.19	24.6	71.64	70.44	54.7	58.47	57.17	20.7	46.82	46.30
Phantom [ICCV2025]	65.67	66.85	7.0	80.40	79.77	40.8	71.84	72.20	52.2	58.88	60.95
ConsisID [CVPR2025]	33.02	31.81	3.0	50.88	50.53	50.3	39.19	38.31	46.8	25.24	23.63
Stand-In [CVPR2026]	54.37	53.97	15.9	81.04	81.24	39.7	56.87	56.28	44.5	42.63	42.18
AnyID [CVPR2026]	62.83	62.74	7.8	73.86	73.77	54.4	66.36	66.02	37.8	55.48	55.77
Gen-4.5	55.02	54.52	22.1	76.41	76.99	44.2	55.66	55.07	33.7	40.17	39.07
Ours (Synthetic)	76.35	76.54	14.9	83.16	84.31	71.0	76.48	76.53	14.1	68.48	68.36
Ours (Real)	**80.78**	**81.68**	29.0	**83.68**	**85.03**	63.5	**80.23**	**80.95**	7.4	**74.27**	**74.80**

Furthermore, Table 3 shows that our model achieves higher identity fidelity on matched subject–prompt pairs, including the subset where our real-gallery model produces greater peak expression intensity than the competing method.

6 Ablation Studies

6.1 Number of Reference Images

Increasing the reference count improves overall CurricularFace similarity by approximately 1.3 points per added reference, with the largest gains under intense and extreme expressions (Table 4). A ninth reference provides no further overall gain, which justifies our choice of using 8 references in the gallery. For context, we also report AnyID (Wang et al., 2026), which accepts 5 references.

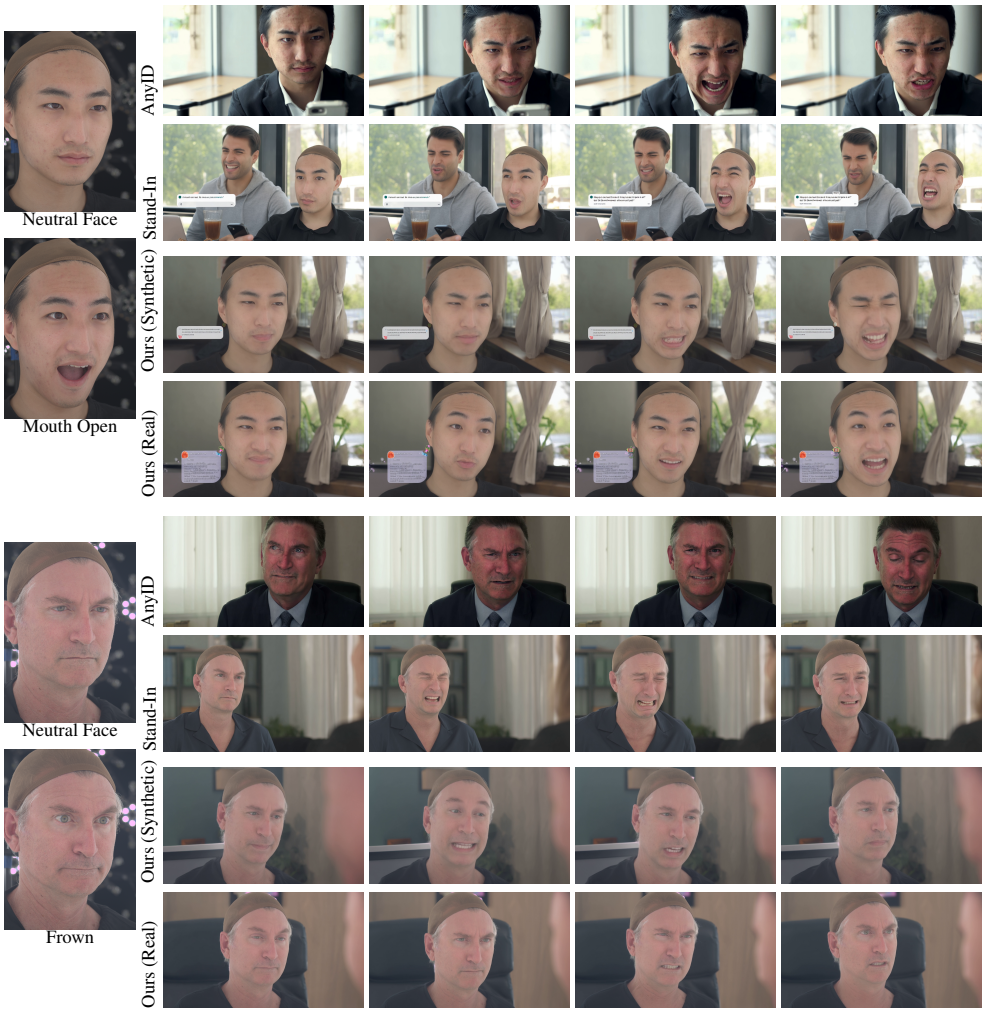


Figure 8: Qualitative comparison. The leftmost images are the subject's neutral portrait and one expressive reference; each row shows four temporally ordered frames from the same generated video.

Table 3: Win rates for Ours (Real) on matched subject–prompt pairs, using per-video mean nearest-gallery similarity. The right-hand columns include only pairs where ours has greater peak expression intensity.

	All pairs		Ours has higher peak intensity	
Baseline	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$
Concat-ID	97.9%	99.0%	96.1%	98.3%
Phantom	98.8%	98.5%	96.9%	96.9%
ConsisID	100.0%	100.0%	100.0%	100.0%
Stand-In	97.9%	98.3%	96.4%	96.4%
AnyID	99.0%	99.2%	96.9%	97.7%
Gen-4.5	100.0%	100.0%	100.0%	100.0%

6.2 Expression Diversity vs. Pose Diversity

Table 5 isolates the effect of expression diversity by comparing three sets of eight references: repeated neutral images, pose-diverse neutral images, and expression-diverse images. The results show that the expression-diverse set achieves higher fidelity than the pose-diverse set in all three

Table 4: Effect of reference count on identity fidelity. Each additional reference adds an expression to the preceding reference set; the ninth reference uses an held-out expression outside the standard gallery. Scores within 0.5 points of the best value in each similarity column are bolded.

		All frames		Mild			Intense			Extreme		
Model	Refs.	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$
AnyID	5	62.83	62.74	7.8	73.86	73.77	54.4	66.36	66.02	37.8	55.48	55.77
Ours	5	77.22	77.77	24.9	**83.92**	**85.05**	60.6	77.03	77.39	14.5	66.51	66.82
Ours	6	78.36	78.94	25.3	**83.62**	**84.78**	63.8	77.83	78.24	10.8	69.17	69.39
Ours	7	79.59	80.34	27.4	**83.72**	**85.03**	63.5	78.93	79.49	9.1	71.80	72.18
Ours	8	**80.78**	**81.68**	29.0	**83.68**	**85.03**	63.5	**80.23**	**80.95**	7.4	**74.27**	**74.80**
Ours	9	**80.71**	**81.45**	28.4	**83.45**	84.51	64.6	**80.33**	**80.97**	7.1	73.16	73.50

regimes. The repeated neutral set achieves the highest fidelity in the mild regime, but shows a similar score in the intense regime and a lower score in the extreme regime compared to the expression-diverse set. The repeated neutral set also produces a much lower share of intense and extreme frames. This suggests that its high overall similarity comes partly from suppressing expression intensity, highlighting the importance of evaluation stratified by intensity.

Table 5: Effect of reference-image diversity on identity similarity. Scores within 0.5 points of the best value in each similarity column are bolded.

	All frames		Mild			Intense			Extreme		
Reference set	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$
Repeated neutral	**85.85**	**86.36**	71.5	**88.21**	**88.71**	27.3	**80.55**	**81.06**	1.1	64.93	66.31
Pose-diverse neutral	57.27	57.41	8.7	67.75	67.71	60.0	59.46	59.52	31.2	50.11	50.47
Expression-diverse	80.78	81.68	29.0	83.68	85.03	63.5	**80.23**	**80.95**	7.4	**74.27**	**74.80**

6.3 Temporal Positional Embedding

We design our network to treat the reference images as an unordered set, so all the references are assigned the same temporal embedding to avoid introducing an ordering bias during training. For comparison, we also train a variant that assigns each reference a distinct temporal embedding. During training, we shuffle the reference images, and for each image $i = 1, 2, \cdots, N$ in the shuffled order, we assign it a temporal embedding $-(i+1)$ on the video-to-image attention path. Reference-image self-attention retains the shared coordinate -1. This negative-temporal variant is trained and evaluated with the same eight expression-diverse references and generation settings as ours. The shared-coordinate model achieves higher identity fidelity in all regimes, shown in Table 6.

Table 6: Effect of temporal positional embeddings for multiple references.

	All frames		Mild			Intense			Extreme		
Temporal coordinates	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$	%	$s^*_{\text{Arc}} \uparrow$	$s^*_{\text{Cur}} \uparrow$
Distinct negative	79.98	80.84	31.9	82.75	84.20	61.3	79.35	79.96	6.7	72.56	72.94
Shared $t = -1$ (ours)	**80.78**	**81.68**	29.0	**83.68**	**85.03**	63.5	**80.23**	**80.95**	7.4	**74.27**	**74.80**

7 Conclusion

We show the limitations of a single neutral portrait for capturing person-specific facial changes and evaluating identity under strong expressions. To address this, we introduce an expression-diverse gallery for evaluation and propose a multi-reference conditioning model, which improves identity similarity across all expression-intensity regimes, especially under extreme expressions. These results highlight the importance of expression diversity in identity-preserving video generation. We encourage future work to provide larger expression-diverse datasets for benchmarking and training.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments and to help implement methods. We have not used generative AI tools for proposing or refining hypotheses, cleaning and reformatting datasets, or interpreting results. Generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for proposing the title and editing the paper to improve readability. We have reviewed all AI-assisted work. AI-assisted edits are manually reviewed before incorporation into the submitted work. We evaluate feedback from generative AI and decide which suggestions to adopt. AI-written code is manually audited. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work includes photographs of human faces and expressions. All subjects have signed consent forms, granting permission for the use of their images in our research and for publication in academic venues. We ensure that all data is handled in accordance with relevant privacy laws and regulations, and that any sensitive information is properly anonymized when necessary.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing detailed descriptions of our evaluation pipelines and training and inference settings, along with the expressions and prompts used in our evaluation set, in the paper and supplementary materials. We will make code, data, and model weights publicly available upon acceptance, before the camera-ready version is submitted.

REFERENCES

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*, pp. 4690–4699, 2019.

Paul Ekman. Facial expressions of emotion: New findings, new questions. *Psychological Science*, 3, Issue 1:34–38, 1992. doi: 10.1111/j.1467-9280.1992.tb00253.x.

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.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

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

Xu Guo, Fulong Ye, Xinghui Li, Pengqi Tu, Pengze Zhang, Qichao Sun, Songtao Zhao, Xiangwang Hou, and Qian He. OmniFace: Bridging the image-to-video gap for high-fidelity face swapping via diffusion transformer. In *European Conference on Computer Vision*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

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 *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5900–5909. IEEE, 2020.

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

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. *arXiv preprint arXiv:2412.00115*, 2024.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, October 2025.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. MiniMax Research, 2026. URL https://www.minimax.io/blog/minimax-h3. Accessed September 23, 2026.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20299–20309, 2024.

George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Runway Research. Runway gen-4.5: State-of-the-art ai video generation. https://runway.com/research/introducing-runway-gen-4.5, 2025. Accessed: 2026-09-18.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

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.

Wan Team. Wan2.2: Open and advanced large-scale video generative models. https://github.com/Wan-Video/Wan2.2, 2025. Accessed September 24, 2026.

Jiahao Wang, Hualian Sheng, Sijia Cai, Yuxiao Yang, Weizhan Zhang, Caixia Yan, Bing Deng, and Jieping Ye. Anyid: Ultra-fidelity universal identity-preserving video generation from any visual references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12808–12817, June 2026.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, August 2020.

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.

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

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025a.

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*, pp. 12978–12988, 2025b.

Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-id: Towards universal identity-preserving video synthesis. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1927–1936. IEEE, 2025.