MULTI-VIEW IDENTITY-CONSISTENT VIDEO GENERATION UNDER LARGE FACIAL-ANGLE VARIATIONS WITHOUT CROSS-PAIRED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-to-video generation still struggles to preserve a known identity when the generated face turns far from a single reference view. Adding multiple views supplies missing identity cues, but in-paired training introduces a complementary failure: the model can copy one reference angle for several frames and then switch abruptly to another. We call this failure *view-dependent copy-paste*. We present $\mathrm{Mv}^2\mathrm{ID}$, a multi-view conditioned framework that improves identity consistency while retaining natural facial motion using only in-paired supervision. The method concatenates video and reference latents, applies region-masking training to prevent direct copying and encourage complementary cue aggregation, and uses a reference-decoupled rotary positional encoding ($\mathrm{RD\text{-}RoPE}$) so reference images are not interpreted as future video frames. We also build a 22K-video large-angle dataset and introduce Multi-view Reference Consistency (MvRC) for identity evaluation across viewpoints. On a held-out set of 30 identities and 150 prompts, $\mathrm{Mv}^2\mathrm{ID}$obtains MvRC scores of 0.544/0.507 and a NaturalScore of 4.69, achieving the best identity consistency and text–video alignment among the compared methods while remaining competitive in naturalness.

1 INTRODUCTION

Diffusion models have become a strong foundation for image and video synthesis. Recent reference-to-video systems condition generation on one or more images to preserve a subject's appearance over time, yet identity often drifts when the face rotates far from the available view. This setting is especially demanding when the target identity is known in advance: a plausible face is insufficient if its structure, clothing, and fine details change across viewpoints.

The limitation is fundamental. A single face image is a projection of 3D shape and appearance under pose, illumination, and occlusion, so it does not uniquely specify the identity cues needed at unseen angles. Multi-view references provide those complementary cues, but naively concatenating them creates a new shortcut. With in-paired supervision, a target frame and its nearest reference view often share low-level appearance. The model can therefore copy that view instead of learning a view-invariant identity representation. The resulting *view-dependent copy-paste* locks the generated face to a discrete conditioning angle for a time interval and can cause an abrupt switch when another reference becomes more convenient (Fig. 2). Cross-paired data weakens this shortcut, but collecting reliable cross-paired examples at large pose gaps is expensive.

We address this problem with $\mathrm{Mv}^2\mathrm{ID}$, a multi-view conditioned video generator trained on in-paired data. We make two lightweight changes to a latent-token baseline. First, region-masking training removes random spatial regions from every reference image, so no single view can be copied in full and the model must combine complementary evidence. Second, reference-decoupled RoPE gives video and reference tokens separate spatial coordinates and a shared reference-time index, preventing reference images from being treated as subsequent video frames. The resulting model preserves identity without constraining the facial trajectory to the reference angles.

We further construct a large-angle human-centric dataset with 22K videos, curate an identity-disjoint test set, and define MvRC, a set-based identity metric that does not force every generated frame to

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.

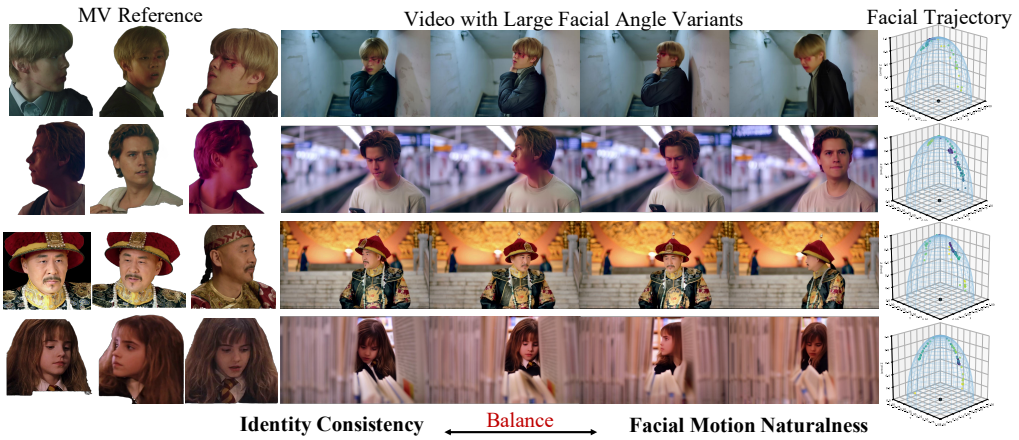

Figure 1: **Multi-view identity consistency.** Single-view conditioning loses identity at large facial angles, while direct multi-view conditioning can copy discrete reference poses. Mv^2IDcombines complementary multi-view cues and preserves a smooth facial trajectory.

match the same reference. Experiments show that Mv^2IDachieves the best identity consistency and text–video alignment among the compared methods while retaining competitive naturalness. Because reliable cross-paired data at large pose gaps is expensive, we treat cross-paired systems as external reference points rather than making a matched-budget claim about the two supervision regimes.

Our contributions are:

- We formulate multi-view, identity-known reference-to-video generation under large facial-angle changes and characterize the view-dependent copy-paste failure caused by in-paired supervision.
- We propose region-masking training and reference-decoupled RoPE, two components that add no inference-time modules and suppress shortcut copying while preserving cross-view identity cues.
- We release a 22K-video large-angle dataset and the MvRC evaluation protocol, and demonstrate improved identity–naturalness trade-offs on quantitative, qualitative, trajectory, and human evaluations.

2 Related Work

Latent video generation. Video generators model compressed visual latents with diffusion or flow-based Transformers. Early text-to-video systems (Ho et al., 2022; Singer et al., 2022; Villegas et al., 2022; Hong et al., 2022) were followed by latent diffusion and image-to-video scaling (Rombach et al., 2022; Blattmann et al., 2023; Bar-Tal et al., 2024). Wan, HunyuanVideo 1.5, and OpenSora scale these backbones (Wan et al., 2025; Wu et al., 2025; Zheng et al., 2024; Lin et al., 2024), while VACE unifies visual conditioning (Jiang et al., 2025). Their general-purpose objectives do not explicitly enforce facial identity across large viewpoint changes.

Personalized image and video generation. DreamBooth and Custom Diffusion adapt generators to subjects (Ruiz et al., 2023; Kumari et al., 2023); IP-Adapter, InstantID, and PhotoMaker inject reference features without per-subject tuning (Ye et al., 2023; Wang et al., 2024; Li et al., 2023). Tune-A-Video adapts image diffusion to video (Wu et al., 2023), while VideoBooth, DreamVideo, and Animate Anyone 2 develop appearance and motion conditioning (Jiang et al., 2024; Wei et al., 2024; Hu et al., 2025). ID-Animator uses a face adapter (He et al., 2024), ConsisID separates facial frequency signals (Yuan et al., 2025b), and MagicMirror uses facial features and conditioned normalization (Zhang et al., 2025a). Portrait-animation systems add facial, audio, or pose controls (Xu

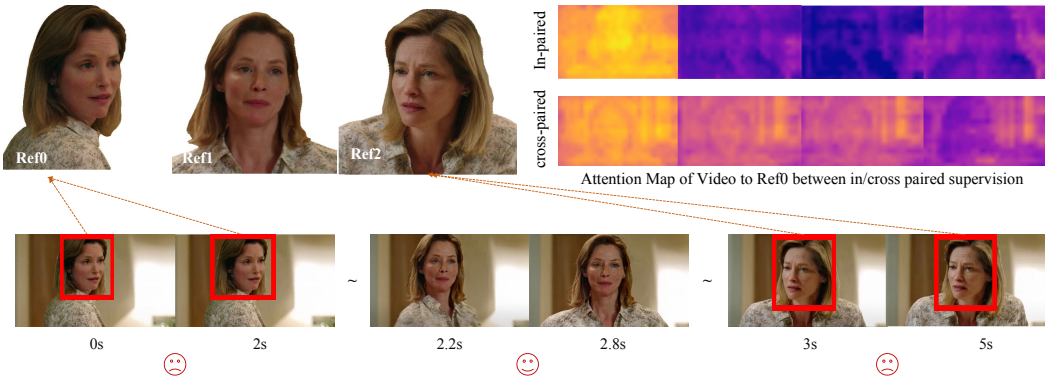

Figure 2: **View-dependent copy-paste.** The female example shows that in-paired training concentrates video-to-reference attention on one view and locks the generated face to its angle, whereas cross-paired supervision produces more balanced attention and smoother transitions.

et al., 2025; Qi et al., 2023; Jin et al., 2024; Hong et al., 2025). Our focus is multi-view facial evidence and motion shortcuts under in-paired supervision.

Subject-to-video and multi-view conditioning. Phantom, MAGREF, and HuMo support subject-consistent synthesis (Liu et al., 2025; Deng et al., 2025; Chen et al., 2025a). Concat-ID concatenates image and video latents and uses cross-video pairing (Zhong et al., 2025); Video Alchemist fuses references and text for open-set personalization (Chen et al., 2025b). Kaleido and MV-S2V explore multi-subject and multi-view interfaces (Zhang et al., 2025b; Song et al., 2026). ConsID-Gen combines auxiliary views with visual–geometric features for object-centric generation (Wu et al., 2026). OpenS2V-Nexus and Phantom-Data supply large-scale subject-centric resources (Yuan et al., 2025a; Chen et al., 2025c). We target large facial-angle changes *without cross-paired data*, combining masked references and decoupled positional coordinates under in-paired supervision.

Motion preservation, shortcut mitigation, and evaluation. Video Alchemist identifies same-video reference copy-and-paste and mitigates it with augmentation (Chen et al., 2025b). MagicID optimizes identity and dynamics through preference pairs, while PersonalVideo targets identity fidelity without dynamic or semantic degradation (Li et al., 2025a;b). Our analysis isolates multi-view angle locking and abrupt switching. VBench evaluates video quality and consistency; VBench-2.0 adds human fidelity, physics, and commonsense (Huang et al., 2023; Zheng et al., 2025). OpenS2V-Nexus supplies NaturalScore for perceptual naturalness (Yuan et al., 2025a). Abrupt angle switches may reduce naturalness despite high identity scores. We therefore combine NaturalScore with MvRC, trajectories, and human ratings, rather than treating it as a dedicated copy-paste detector.

3 Problem Setup and Failure Analysis

Let a video generator $\mathcal{M}$ produce a clean latent video z_0 conditioned on a text prompt c and n identity references $\mathcal{I} = \{I_1, \ldots, I_n\}$. The references show the same person at different facial angles. We use flow matching (Lipman et al., 2023; Liu et al., 2022); given a noisy latent z_t, the model predicts the velocity field that transports noise to z_0.

View-dependent copy-paste. We observed that direct multi-view conditioning improves identity but can collapse facial motion under in-paired supervision. Figure 2 shows a generated face remaining near one reference angle for several frames before switching to another. Attention maps concentrate on the nearest reference during the same interval, whereas cross-paired supervision distributes attention more evenly. This failure is also relevant to perceptual naturalness: a discrete-angle lock followed by an abrupt switch introduces temporal artifacts that NaturalScore is designed to penalize. This motivates training-time interventions that break direct copying without discarding multi-view information.

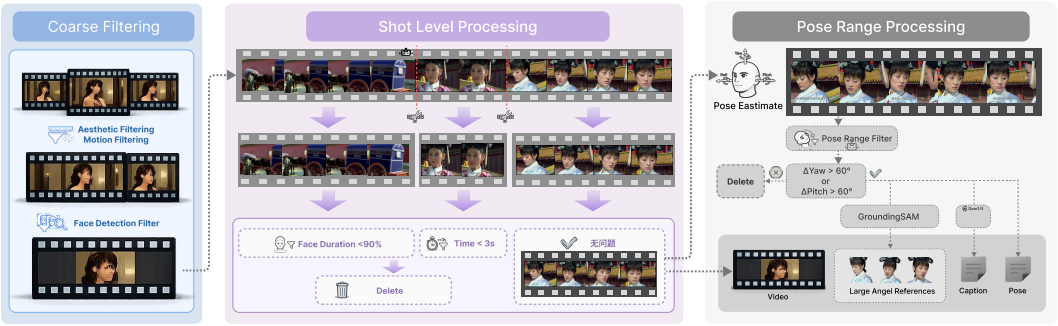

Figure 3: **Large-angle dataset construction.** We progressively apply quality and face filtering, shot-level processing, and pose-range filtering to obtain human-centric videos with large facial-angle variation, reference frames, pose metadata, and captions.

4 METHOD

4.1 LARGE-ANGLE DATASET

We build the training corpus with three stages (Fig. 3). We first filter raw videos using aesthetic and quality scores (Talebi & Milanfar, 2018; Xu et al., 2023) and face detection (Deng et al., 2019b). We then segment videos into coherent clips and keep single-subject tracks with long, stable face visibility. Finally, we estimate 3D head pose (yaw, pitch, and roll), retain clips with large pose ranges (e.g., yaw or pitch above $45°$), and generate or refine captions with Qwen2.5-VL (Bai et al., 2025). The resulting 22K videos provide diverse large-angle trajectories for in-paired training. The independent test set contains 30 identities, five prompts per identity, and ten additional multi-view images per identity.

4.2 MULTI-VIEW REFERENCE INJECTION

We encode the noisy video and clean reference images with the same video VAE. Video tokens and reference tokens are concatenated and processed by the DiT self-attention blocks:

$$\mathbf{X} = [\mathbf{X}_1^t, \ldots, \mathbf{X}_{m+1}^t, \mathbf{R}_1, \ldots, \mathbf{R}_n], \tag{1}$$

where $\mathbf{X}_i^t$ are video tokens at diffusion time t and $\mathbf{R}_j = \mathrm{VAE}(I_j)$ are reference tokens. Text features from a T5 encoder (Raffel et al., 2020) enter through the model's cross-attention layers. This token-level integration preserves the backbone and lets every video token aggregate evidence from all reference views.

4.3 REGION-MASKING TRAINING

A simple view-level mask (VM) blocks the reference whose angle is closest to a target frame. VM removes the most obvious shortcut, but indirect self-attention paths can still transmit the same information, and removing an entire view wastes useful identity cues.

Inspired by learning from partial observations in masked autoencoders (He et al., 2022), we randomly mask spatial regions of every reference. For reference I_j, we sample a binary pixel mask M_j with masking ratio ρ and feed

$$\mathbf{R}_j^m = \mathrm{VAE}(I_j \odot M_j) \tag{2}$$

to the model during training. Because every view is only partially observed, the generator cannot copy a complete face from one reference. It must combine complementary regions across views to predict the target video, which encourages view-invariant identity features. At inference we use the unmasked references, so the strategy adds no runtime cost.

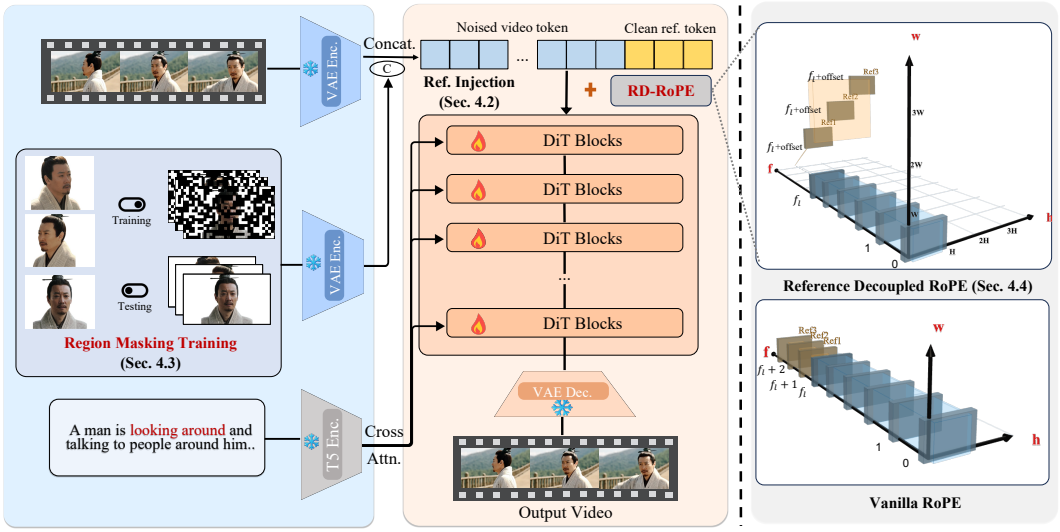

Figure 4: **Training pipeline.** Reference latents are concatenated with noisy video latents. Region masking removes spatial shortcuts during training, while RD-RoPE separates the positional coordinates of video and reference tokens during both training and inference. At inference, masking is removed and the full reference images are used.

4.4 Reference-decoupled RoPE

Rotary position embeddings encode relative token positions (Su et al., 2021). Vanilla 3D RoPE assigns each token a temporal coordinate f and spatial coordinates (h, w). If references are appended after video tokens, they receive consecutive temporal indices and may be interpreted as future frames. We introduce RD-RoPE with two changes. All reference images share one temporal coordinate offset, making them temporally equivalent, while each reference receives a disjoint spatial band:

$$f_j = f_{\text{last}} + \delta, \qquad h_j = h_v + jH, \qquad w_j = w_v + jW, \quad j = 1, \ldots, n, \tag{3}$$

where (h_v, w_v) are the base spatial coordinates and H, W are the video grid dimensions. Thus temporal RoPE models dynamics only within the video, while spatial offsets distinguish reference views and support cross-view interaction without imposing a false temporal order.

5 Experiments

5.1 Experimental Setup

We build the main model on Wan-2.1-T2V-14B (Wan et al., 2025) and use three reference images unless stated otherwise. Videos are sampled at 16 fps and resized to 480×832. Ablations use Wan-2.2-TI2V-5B (Wan et al., 2025). The 14B main model matches the scale of the compared systems, while the 5B backbone reduces the cost of repeated component experiments. Absolute scores across these two backbones should not be interpreted as a controlled model-scaling comparison. We compare with Phantom (Liu et al., 2025), MAGREF (Deng et al., 2025), and HuMo (Chen et al., 2025a); Phantom and MAGREF are evaluated with both their single-view and multi-view interfaces.

Metrics. AES, IQA, TVA, and HA are the VBench (Huang et al., 2023) aesthetic, image-quality, text–video alignment, and human-action scores. To measure identity across viewpoints, we crop each generated face, extract ArcFace (Deng et al., 2019a) or CurricularFace (Huang et al., 2020) embeddings, and compute the average cosine similarity to ten reference images:

$$\text{MvRC}(z_0^i) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \cos\big(M_f(z_0^i), M_f(r)\big), \tag{4}$$

where the video score averages frame scores. We report MvRC-Arc and MvRC-Cur. NaturalScore follows the evaluation protocol of OpenS2V-Nexus (Yuan et al., 2025a): we uniformly sample 16

Table 1: Comparison under single-view (SV) and multi-view (MV) conditioning. Best and second best are bold and underlined.

Method	AES↑	IQA↑	TVA↑	HA↑	MvRC-Arc↑	MvRC-Cur↑	NaturalScore↑
Phantom-SV (Liu et al., 2025)	0.523	0.621	0.091	<u>0.977</u>	0.408	0.365	3.98
Phantom-MV (Liu et al., 2025)	0.532	0.635	0.091	0.972	0.487	0.438	4.11
HuMo (Chen et al., 2025a)	0.531	**0.649**	<u>0.092</u>	**0.978**	<u>0.493</u>	<u>0.459</u>	**4.71**
MAGREF-SV (Deng et al., 2025)	0.567	0.602	0.085	0.947	0.270	0.223	4.40
MAGREF-MV (Deng et al., 2025)	**0.569**	0.602	0.086	0.952	0.338	0.286	4.43
Mv^2ID	<u>0.568</u>	<u>0.645</u>	**0.097**	0.960	**0.544**	**0.507**	<u>4.69</u>



Figure 5: Qualitative comparison. Mv^2IDbetter preserves facial structure and clothing details while maintaining continuous viewpoint changes.

frames and ask Seed2.0-Lite (Team, 2026) to rate realism, physical plausibility, visual naturalness, and visible temporal artifacts on a 1–5 scale. The prompt is adapted from OpenS2V-Nexus and is provided in the supplementary material. In our setting, NaturalScore is used as a complementary perceptual measure of motion naturalness; it is not interpreted as a standalone identity metric.

5.2 Main Comparison

Quantitative results. Table 1 shows that Mv^2IDachieves the best identity consistency, improving over HuMo by $+0.051$ and $+0.048$ with ArcFace and CurricularFace. It also obtains the highest text–video alignment while remaining competitive in naturalness. The gap between Phantom/MAGREF single-view and multi-view variants confirms that complementary views are valuable. The strong MvRC together with competitive NaturalScore suggests that the identity gain is not obtained solely by locking the output to one reference angle.

Qualitative and trajectory results. Figure 5 shows identity drift, lighting transfer, or abrupt motion in competing methods. Mv^2IDkeeps stable facial structure and clothing details across large

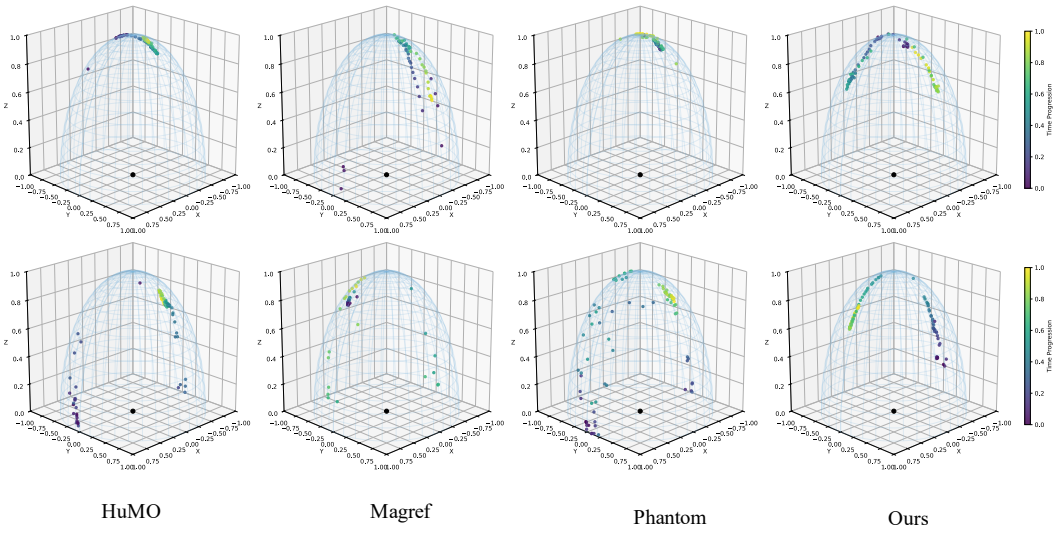

Figure 6: Facial-direction trajectories. Colors encode time; a smooth path indicates continuous viewpoint motion.

Table 2: Ablation of reference-decoupled RoPE (R) and region masking (M).

Setting	AES↑	IQA↑	TVA↑	HA↑	MvRC-Arc↑	MvRC-Cur↑	NaturalScore↑
B	0.568	0.664	0.089	0.960	<u>0.546</u>	<u>0.515</u>	3.70
B+R	<u>0.578</u>	<u>0.671</u>	<u>0.091</u>	<u>0.971</u>	**0.552**	**0.522**	<u>4.12</u>
B+M	0.569	**0.686**	**0.093**	0.966	0.537	0.504	3.93
B+R+M	**0.580**	<u>0.671</u>	0.087	**0.975**	0.535	0.502	**4.52**

B: base multi-view model; R: RD-RoPE; M: region masking.

rotations. In Fig. 6, each point is a normalized 3D facial direction and color denotes time. Baselines either cluster around a reference angle or jump between distant angles; our trajectory changes smoothly over time.

5.3 Ablation Studies

We start from direct multi-view conditioning (B) and add RD-RoPE (R) and region masking (M) independently and together. Results use the 5B ablation model.

Results. RD-RoPE improves NaturalScore from 3.70 to 4.12 and gives the strongest identity score in isolation. Region masking improves IQA and TVA while raising NaturalScore to 3.93. Combining the two yields the best AES, HA, and NaturalScore (4.52), with identity scores remaining close to the base. The small MvRC decrease in the full model is consistent with removing view locking: a trajectory that no longer copies a reference angle can be less similar to any one reference while receiving a higher naturalness score. This identity–naturalness trade-off is precisely the behavior targeted by our failure analysis.

Figure 7 supports this interpretation. The full model keeps clothing and accessory cues even when reference views disagree, and its facial path is smoother than the base, RD-RoPE-only, or masking-only variants.

Additional ablations. We further vary the number of references and the region-masking ratio (Tables 3 and 4). Increasing the number of views from one to two gives the largest identity gain, while a third view provides a smaller but consistent improvement. A 20% mask obtains a NaturalScore



Figure 7: Ablation visualization. Multi-view references improve identity; region masking and RD-RoPE remove discrete-angle copying while retaining fine-grained attributes.

Table 3: Effect of the number of reference images. Two views provide most of the identity gain; the third view gives a smaller additional improvement.

References	MvRC-Arc↑	MvRC-Cur↑
1	0.444	0.403
2	<u>0.502</u>	<u>0.463</u>
3	**0.504**	**0.464**

Table 4: Effect of the region-masking ratio. Higher ratios improve NaturalScore and suppress view-dependent copy-paste, yielding more natural transitions.

Mask ratio	NaturalScore↑
20%	3.82
40%	<u>4.13</u>
60%	**4.30**

of 3.82. At 40% and 60%, we do not detect obvious view-dependent copy-paste in the generated trajectories, and the NaturalScore increases from 4.13 to 4.30. We use three references and a 60% mask in the main experiments.

5.4 Human Evaluation

We sample 25 cases and recruit ten university students as independent evaluators. All participants have normal or corrected-to-normal vision and are not involved in model development or data collection. For each case, anonymized videos from six methods are presented in randomized order. Participants rate identity consistency, appearance naturalness, and motion/viewpoint naturalness using 1–5 Likert scales. Mv^{2}IDreceives the highest identity and appearance-naturalness ratings, while HuMo is slightly higher on motion/viewpoint naturalness. The method effect on identity is significant under both one-way ANOVA ($F = 81.70$, $p < 10^{-70}$) and Kruskal–Wallis ($H = 283.75$, $p < 10^{-50}$) (Fisher, 1925; Kruskal & Wallis, 1952). We report case-level means in Fig. 8; the evaluation protocol and NaturalScore prompt are provided in the supplementary material. The different identity and motion rankings highlight why both dimensions matter: stronger identity preservation does not automatically imply more natural dynamics. Together with the automatic scores, these ratings support the narrower conclusion that our method improves identity consistency while retaining competitive naturalness, rather than dominating every perceptual dimension.

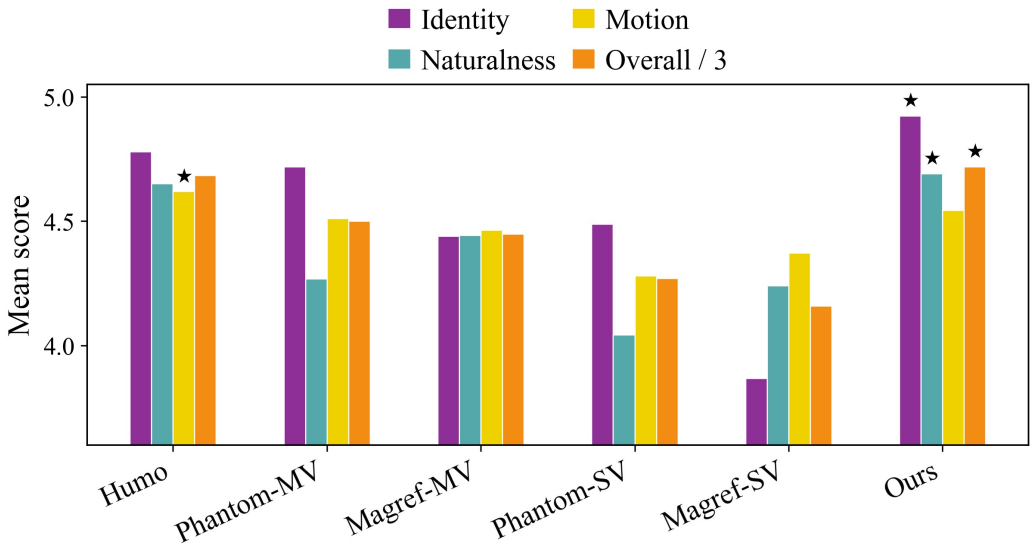

Figure 8: Human-study mean ratings across identity, naturalness, motion, and overall quality.

6 Limitations and Discussion

Our automatic naturalness score relies on a vision-language judge and can be sensitive to prompt wording, model bias, and judge randomness; we therefore use it as a complementary signal and retain human evaluation. The prompt is adapted from OpenS2V-Nexus rather than claimed to be an exact reproduction because our judge is Seed2.0-Lite. The dataset is human-centric and mined from videos with visible faces, so performance on severe occlusion, multiple people, or non-human subjects is not established. Finally, MvRC is a similarity measure rather than a complete perceptual identity test: because the references cover discrete angles, it should be interpreted together with NaturalScore, trajectory visualization, and human ratings. Future work should learn a calibrated identity–motion evaluator and extend training to cross-shot appearance changes.

The component ablations clarify the intended operating point. Relative to the direct multi-view baseline, the full 5B model increases NaturalScore by 0.82 while MvRC-Arc and MvRC-Cur decrease by 0.011 and 0.013. Thus the combined design does not maximize embedding similarity alone. Region masking disrupts complete reference-image copying during training, while positional decoupling distinguishes reference evidence from video time. Their joint benefit is supported by the naturalness results and visual trajectories. These observations motivate evaluating identity preservation alongside motion quality whenever reference images provide strong reconstruction shortcuts. In-paired supervision is a practical constraint that permits reference collection directly from training clips. Our results demonstrate useful cross-view identity learning within this constraint. The relative benefits of cross-paired supervision remain dependent on data coverage and training resources, and require a separate controlled study to quantify.

7 Conclusion

We studied identity-known video generation when facial angles move far beyond any single reference. Mv^{2}IDcombines multi-view latent conditioning, region-masking training, and reference-decoupled RoPE to reduce view-dependent copy-paste under inexpensive in-paired supervision. A 22K-video large-angle dataset and the MvRC protocol make the setting measurable. Across automatic metrics, trajectories, visual comparisons, and human ratings, Mv^{2}IDachieves the best identity consistency and text–video alignment among the compared methods while preserving competitive naturalness.

Reproducibility Statement

Data-processing, method, evaluation, and ablation details appear in the paper and appendix.

AI Use Statement

AI tools assisted with language editing and LaTeX organization; the authors verified the final content.

References

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, 2025. URL https://arxiv.org/abs/2502.13923.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. URL https://arxiv.org/abs/2401.12945.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning, 2025a. URL https://arxiv.org/abs/2509.08519.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject Open-set Personalization in Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6099–6110, June 2025b.

Zhuowei Chen, Bingchuan Li, Tianxiang Ma, Lijie Liu, Mingcong Liu, Yi Zhang, Gen Li, Xinghui Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom-data : Towards a general subject-consistent video generation dataset, 2025c. URL https://arxiv.org/abs/2506.18851.

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 (CVPR)*, pp. 4690–4699, 2019a. doi: 10.1109/CVPR.2019.00482. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild, 2019b. URL https://arxiv.org/abs/1905.00641.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, et al. Magref: Masked guidance for any-reference video generation with subject disentanglement. *arXiv preprint arXiv:2505.23742*, 2025.

Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh and London, 1925.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://openaccess.thecvf.

com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

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, 2024. URL https://arxiv.org/abs/2404.15275.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. URL https://arxiv.org/abs/2204.03458.

Fa-Ting Hong, Zunnan Xu, Zixiang Zhou, Jun Zhou, Xiu Li, Qin Lin, Qinglin Lu, and Dan Xu. Audio-visual controlled video diffusion with masked selective state spaces modeling for natural talking head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, 2022. URL https://arxiv.org/abs/2205.15868.

Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance, 2025. URL https://arxiv.org/abs/2502.06145.

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 (CVPR)*, 2020.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6689–6700, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing, 2025. URL https://arxiv.org/abs/2503.07598.

Xiaoyu Jin, Zunnan Xu, Mingwen Ou, and Wenming Yang. Alignment is all you need: A training-free augmentation strategy for pose-guided video generation. *arXiv preprint arXiv:2408.16506*, 2024.

William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. doi: 10.1080/01621459.1952.10483441.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. URL https://arxiv.org/abs/2212.04488.

Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. MagicID: Hybrid Preference Optimization for ID-Consistent and Dynamic-Preserved Video Customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12737–12746, October 2025a.

Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19406–19416, October 2025b.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023. URL https://arxiv.org/abs/2312.04461.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. Open-sora plan: Open-source large video generation model, 2024. URL https://arxiv.org/abs/2412.00131.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Zipeng Qi, Xulong Zhang, Ning Cheng, Jing Xiao, and Jianzong Wang. Difftalker: Co-driven audio-image diffusion for talking faces via intermediate landmarks. *arXiv preprint arXiv:2309.07509*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. URL https://arxiv.org/abs/2208.12242.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. URL https://arxiv.org/abs/2209.14792.

Ziyang Song, Xinyu Gong, Bangya Liu, and Zelin Zhao. Mv-s2v: Multi-view subject-consistent video generation, 2026. URL https://arxiv.org/abs/2601.17756.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. URL https://arxiv.org/abs/2104.09864.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

ByteDance Seed Team. Seed2.0 (model page and model card). https://seed.bytedance.com/zh/seed2, 2026. Accessed: 2026-03-04.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description, 2022. URL https://arxiv.org/abs/2210.02399.

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, 2025. URL https://arxiv.org/abs/2503.20314.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds, 2024. URL https://arxiv.org/abs/2401.07519.

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

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7623–7633, October 2023.

Mingyang Wu, Ashirbad Mishra, Soumik Dey, Shuo Xing, Naveen Ravipati, Hansi Wu, Binbin Li, and Zhengzhong Tu. ConsID-Gen: View-Consistent and Identity-Preserving Image-to-Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1853–1863, June 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2304.05977.

Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, et al. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15909–15919, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Chongyang Ma, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025a. URL https://openreview.net/forum?id=XKhLsRPMsw.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988, June 2025b.

Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. MagicMirror: ID-Preserved Video Generation in Video Diffusion Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14464–14474, October 2025a.

Zhenxing Zhang, Jiayan Teng, Zhuoyi Yang, Tiankun Cao, Cheng Wang, Xiaotao Gu, Jie Tang, Dan Guo, and Meng Wang. Kaleido: Open-sourced multi-subject reference video generation model. 2025b.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL https://arxiv.org/abs/2503.21755.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-ID: Towards Universal Identity-Preserving Video Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1927–1936, October 2025.

A APPENDIX

This appendix provides the full NaturalScore prompt, additional data-processing details, masking comparisons, a positional-encoding derivation, and results that complement the main paper.

A.1 NATURALSCORE PROMPT AND PROTOCOL

The following prompt is adapted from OpenS2V-Nexus (Yuan et al., 2025a) for Seed2.0-Lite. The judge receives 16 uniformly sampled frames and returns one video-level integer score. We use the same prompt and inference settings for all methods.

> You are asked to judge how realistic the given video clip appears based on 16 uniformly sampled frames. Evaluate common-sense consistency, physical plausibility, visual naturalness, and visible AI-generation artifacts. If humans are present, pay particular attention to facial anatomy, body proportions, and natural motion. Output one integer score from 1 to 5 only: 1 indicates definitely AI-generated with severe artifacts or implausible motion; 2 indicates likely AI-generated with noticeable distortions; 3 indicates borderline realism; 4 indicates mostly realistic with minor imperfections; and 5 indicates highly realistic, physically plausible, and free of visible artifacts.

A.2 ADDITIONAL DATA-PROCESSING DETAILS

Beyond the pipeline described in the main paper, we discard clips shorter than three seconds, clips with multiple people, and clips with intermittent face tracks. We optionally obtain foreground masks with Grounding DINO and SAM. Reference sampling selects three frames with pairwise pose gaps above $45°$. Table 5 summarizes the dataset characteristics. Experiments use NVIDIA A800 GPUs.

Table 5: Comparison with representative subject-to-video datasets.

Dataset	Human-centric	Large-angle	Pose metadata
OpenS2V-5M (Yuan et al., 2025a)	×	×	×
Phantom-Data (Chen et al., 2025c)	×	×	×
HuMoSet (Chen et al., 2025a)	✓	×	×
Ours	✓	✓	✓

A.3 REGION MASKING VERSUS VIEW MASKING

For view masking (VM), the VAE compresses four frames into one latent, and we use the first frame of each group as its pose anchor. We block attention from the corresponding video latent to the reference whose estimated pose is closest to that anchor. Figure 9 contrasts this view-level operation with region masking (RM); Fig. 10 adds examples with illumination and clothing conflicts.

A.3.1 MASKING RATIO

Figure 11 provides visual examples for the masking-ratio experiment reported in the main paper.

A.4 WHY SPATIAL OFFSETS MATTER IN RD-ROPE

RoPE attention depends on relative positions. For a two-dimensional subspace m,

$$q_i^{(m)\prime} = R(\omega_m p_i) q_i^{(m)}, \qquad k_j^{(m)\prime} = R(\omega_m p_j) k_j^{(m)}, \tag{5}$$

so the attention contribution is

$$(q_i^{(m)\prime})^\top k_j^{(m)\prime} = (q_i^{(m)})^\top R(\omega_m (p_j - p_i))\, k_j^{(m)}. \tag{6}$$

If only a temporal offset is applied to reference tokens, the relative coordinate changes from $(\Delta t, \Delta x, \Delta y)$ to $(\Delta t + o_t, \Delta x, \Delta y)$. This shifts temporal phases while leaving the spatial phases

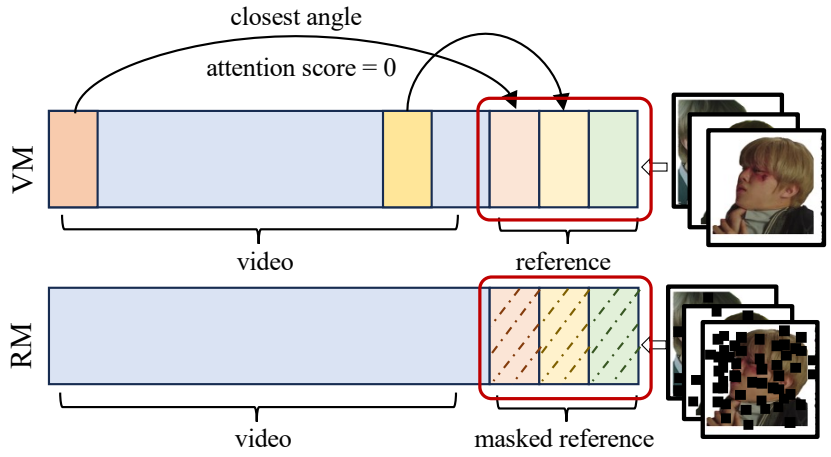

Figure 9: VM removes one complete view; RM removes partial regions from every view.



Figure 10: Visual comparison of view masking and region masking.

unchanged, producing a positional distribution that is unlike the pretrained video coordinate system. We therefore apply coherent offsets in all three axes:

$$(\Delta t, \Delta x, \Delta y) \mapsto (\Delta t + o_t, \Delta x + o_x, \Delta y + o_y). \tag{7}$$

The reference tokens remain temporally equivalent but occupy disjoint spatial bands. At an early training checkpoint (450 steps), the H/W-offset model already uses reference information, whereas the temporal-only model largely ignores it (Fig. 12).



Figure 11: Qualitative comparison of two region-masking ratios.

Table 6: Additional quality and alignment metrics for different reference counts. Best and second best are bold and underlined.

References	AES↑	IQA↑	TVA↑
1	0.561	<u>0.658</u>	0.084
2	**0.572**	**0.659**	<u>0.085</u>
3	<u>0.569</u>	0.654	**0.087**

A.5 Additional Reference-count Metrics

Table 6 complements the identity scores in the main paper with aesthetic quality, image quality, and text–video alignment. These metrics do not improve uniformly with additional views: two references obtain the highest AES and IQA, while three obtain the highest TVA.

A.6 More Qualitative Results

We provide additional visual comparisons and facial trajectories in Figs. 13 and 14. Dynamic reference counts and shuffled reference ranges can further improve individual samples, but we keep the fixed three-view protocol in the main paper so the contribution of RM and RD-RoPE remains isolated.



Figure 12: H/W offsets accelerate the use of reference information during early training.



Figure 13: Additional visual results.

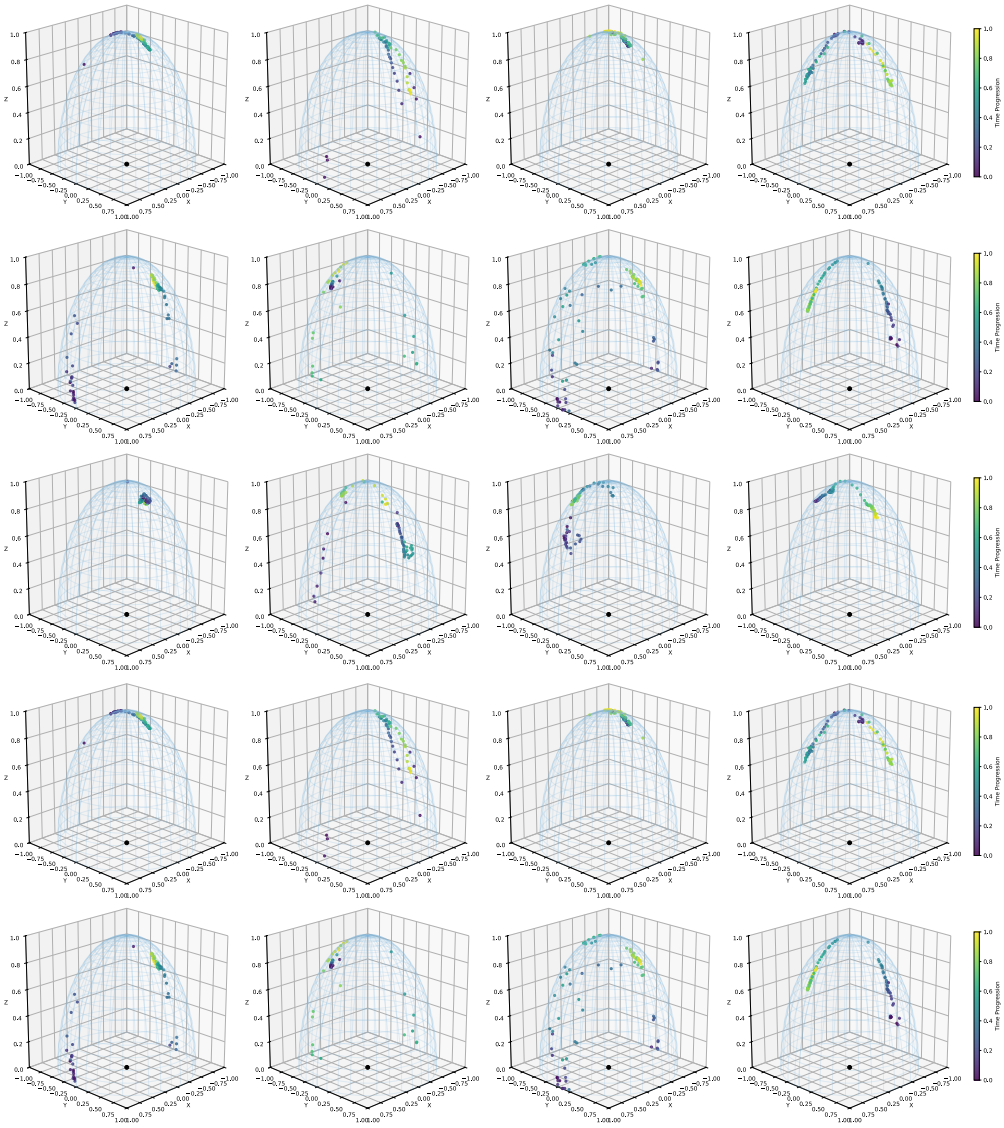

Figure 14: Additional facial trajectories for Mv^{2}IDand baselines.