Identity-Preserving Text-to-Video Generation via Agentic Enhancement and Semantic Repair

Jiayi Gao
gaojiayi26@stu.pku.edu.cn
Wangxuan Institute of Computer Technology, Peking University
Beijing, China

Changcheng Hua
hcc@stu.pku.edu.cn
Wangxuan Institute of Computer Technology, Peking University
Beijing, China

Jiaqi Tang
tangjiaqi@stu.pku.edu.cn
Wangxuan Institute of Computer Technology, Peking University
Beijing, China

Yuxin Peng
pengyuxin@pku.edu.cn
Wangxuan Institute of Computer Technology, Peking University
Beijing, China

Yang Liu*
yangliu@pku.edu.cn
Wangxuan Institute of Computer Technology, Peking University
Beijing, China

Abstract

Identity-preserving video generation aims to synthesize videos that follow natural-language instructions while maintaining the visual identity of a given subject. Recent commercial video generation models have achieved strong visual quality and motion realism, but they still suffer from identity drift, incomplete instruction following, and missing visual details under complex prompts. Since these models are usually closed-source black boxes, directly improving them through parameter optimization is often infeasible. We therefore propose Agentic Enhancement and Semantic Repair (AESR), a lightweight enhancement framework for identity-preserving video generation. To improve prompt construction before generation and mitigate the above failures, AESR introduces a global agentic prompt enhancement module. This module learns model-specific prompting formats from official documentation, acquiring human-centered video generation priors from human-interaction data, and accumulates test-domain identity-preserving generation experience into a reusable playbook through an agentic loop. To further repair errors of videos generated with enhanced prompts, AESR further introduces a sample-level visual semantic repair module, which uses a VLM to locate erroneous video segments and design repair instructions, edits selected frames into explicit visual references, and guides a video editing model to fix local semantic or identity-related errors. We also adopt a lightweight Mixture-of-Experts selection strategy to choose reliable outputs from different generation and refinement paths. Under the official evaluation protocol of the ACM MM 2026 Identity-Preserving Video Generation Challenge, our system MIPL_Video ranked first in Track 1, demonstrating the effectiveness of AESR for practical identity-preserving video generation. The code is available at https://github.com/oceanflowlab/AESR.

*Corresponding author.

MM '26, Rio de Janeiro, Brazil.

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3837688

CCS Concepts

• **Computing methodologies** → **Computer vision tasks**; *Image and video acquisition.*

Keywords

text-to-video generation, identity preserved video generation, agentic based video generation

ACM Reference Format:
Jiayi Gao, Changcheng Hua, Jiaqi Tang, Yuxin Peng, and Yang Liu. 2026. Identity-Preserving Text-to-Video Generation via Agentic Enhancement and Semantic Repair. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3767308.3837688

1 Introduction

Identity-preserving video generation aims to synthesize a video that follows a natural-language instruction while maintaining the visual identity of a given source subject. Recent commercial video generation models, such as Seedance 2.0 [22], have achieved impressive visual quality and motion realism. However, they still struggle in identity-sensitive scenarios: the generated subject may drift from the source identity, and complex instructions involving multiple visual elements or motion events may lead to missing objects, incomplete actions, or underspecified target states. Since these models are typically closed-source black boxes, directly improving them through parameter optimization is often infeasible. We therefore focus on what remains controllable at the interface level: prompt construction before generation and draft repair through external editing models afterward.

A straightforward way to mitigate these failures is to sample multiple candidates or manually rewrite the prompt. However, both solutions are costly and brittle. Repeated sampling requires multiple calls to expensive commercial APIs, while manual rewriting heavily depends on rewriter's experience. More importantly, effective prompts are shaped by both model-specific preferences and test-domain experience. Different video models may prefer different instruction formats due to their training data, such as how action decomposition, subject identity and camera motions are arranged. Meanwhile, test samples with similar instruction demands, such

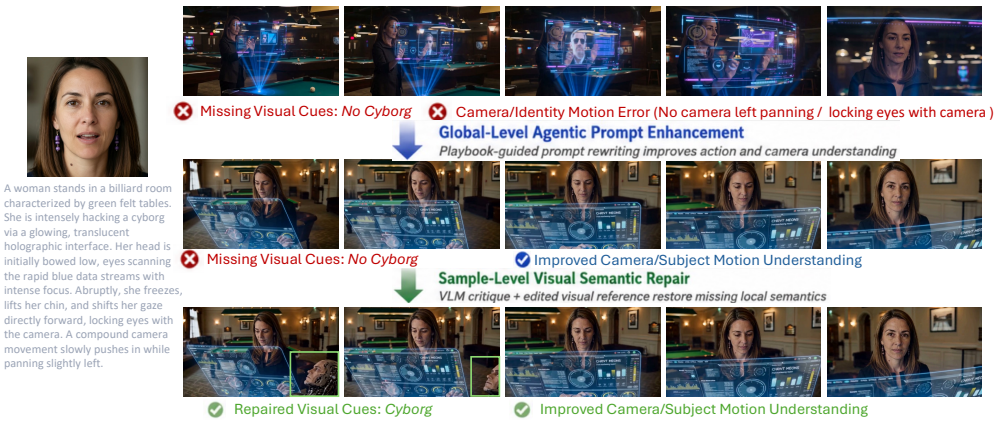

Figure 1: AESR progressively improves identity-preserving video generation. Global-level agentic prompt enhancement generally improves prompt alignment by improved motion understanding, while sample-level visual semantic repair restores missing visual elements.

as complex actions or camera-motion control, often share reusable domain-level experience. Moreover, some errors are difficult to fix by simply expanding the text prompt. Even when the instruction explicitly specifies certain visual elements, such as the cyborg in Fig. 1, the generated video may still omit them. Since video models must coordinate appearance, actions, and camera motion across multiple frames, adding more text can introduce ambiguity when correcting local visual omissions; explicit visual references instead provide concrete guidance for repairing.

Therefore, we propose **AESR**, a lightweight enhancement framework for IPVG via **Agentic Enhancement and Semantic Repair**, consists of two parts: **Global-level Prompt Enhancement** via an agentic loop, and **Sample-level Visual Semantic Repair**. At the global level, inspired by [38], we introduce an agentic loop mechanism that converts model prompt preferences and test-domain experience into reusable generation knowledge and store them into an ever-evolving playbook. The initial knowledge of the Playbook comes from two sources: the official prompting documentation of the target video model, and an external Playbook learned from the Human-object Interaction (HOI) Editing dataset [6]. The former provides model-specific prompting preferences, allowing the VLM to learn the instruction organization favored by the target model. The latter provides experience from HOI video generation. Since more than 95% of the identity-preserving videos in the test set involve human interactions, as estimated using GPT-4o [12], we draw prior knowledge from relevant external datasets to enhance interaction descriptions. In the loop, the agent analyzes failed test videos and incrementally updates the Playbook with generalizable experience, allowing it to adapt from initialized knowledge to the test domain. The updated Playbook is then used to enhance prompts for challenging videos that require stronger instruction following as in Fig 1. At the sample level, we introduce visual semantic repair to address errors that text alone cannot resolve. For visual errors in the initial video, such as identity drift or missing elements, edited image references provide concrete target states and serve as visual anchors for the motion-aware video editor to repair action and camera-motion errors. Specifically, we use a VLM [20, 26] to locate the approximate erroneous segment and design a repair instruction, and we use an image editor to correct specific frames within that segment as explicit visual references showing "what should appear".These repaired reference frames, together with the original video, are then fed into a video editing model to supplement or fix the missing or incorrect visual elements in the generated result.

Putting these components together, we provide a practical enhancement solution for identity-preserving video generation with closed-source video models. In addition, we follow a lightweight Mixture-of-Experts (MoE) selection strategy to compare and integrate candidate videos produced by different generation or refinement paths, selecting the most reliable result for final submission. According to the official evaluation protocol of the ACM MM 2026 Identity-Preserving Video Generation Challenge, our system ranked first on the Facial Identity-Preserving Video Generation Track, validating the effectiveness of the proposed framework.

To sum up, our contributions are threefold:

- We propose AESR, a lightweight interface-level enhancement framework for identity-preserving video generation with closed-source video models.
- We introduce two complementary modules in AESR: a global agentic prompt enhancement module that converts model-specific prompt preferences and test-domain experience into reusable generation knowledge for enhancing input prompts, and a sample-level visual semantic repair module that localizes errors, builds edited references, and uses video editing models for targeted refinement.
- Under the comprehensive official evaluation protocol of the ACM MM 2026 Identity-Preserving Video Generation Challenge, our system ranked first in the Facial Identity-Preserving Video Generation Track.

2 Related Work

Diffusion models [2, 5, 8, 14, 30, 32, 33] have propelled significant progress in many downstream tasks [24, 25, 34–37] including

identity-preserving generation [27, 28, 31]. Early approaches primarily relied on per-ID fine-tuning methods, such as MotionBooth [29] and DreamVideo [28], which incorporated reference content by fine-tuning model parameters or introducing additional modules. However, these methods required retraining for each new identity, greatly limiting scalability and practical deployment. To address these challenges, tuning-free strategies emerged, ACE++ [18] and PhotoMaker [15] developed subject-preserved image generation models based on this approach. More recently, advanced models like Phantom [17], VACE [13] and LTX-2 [9] have demonstrated the capability to generate consistent multi-subject videos in open-domain scenarios [3, 16], steadily closing the performance gap with commercial solutions like Hailuo [19], Vidu [1] and Seedance 2.0[22]. Nevertheless, these methods require collecting large amounts of data and time-consuming post-training. Instead, we treat identity-conditioning mechanism as an implementation choice and study how to improve identity-preserving video generation through interface-level prompt enhancement and post-generation repair, which is more compatible with closed-source video models.

3 Method

3.1 Overview

Given a source identity image I and a user instruction P, AESR aims to generate a video that faithfully follows the instruction while preserving the visual identity of the source subject. Since modern video generation and editing models are typically black-box systems, AESR improves controllability through two interface-level interventions. As illustrated in Fig. 2, to acquire and exploit generalizable knowledge for enhancing the instruction of samples, we propose **Global-Level Agentic Playbook Enhancement**, which aims to convert model-specific prompting preferences, human-interaction video generation experience, and failure experience observed from test samples into reusable generation knowledge, thereby improving the quality of generated videos. To address sample-specific errors that still remain in the videos produced by Stage I, we further propose **Sample-Level Visual Semantic Repair**, which focuses on diagnosing and correcting semantic gaps, missing visual elements, and identity-related errors in individual generated videos.

3.2 Global-Level Agentic Playbook Enhancement

To continuously acquire and reuse generalizable generation knowledge from the test domain, we propose Global-Level Agentic Playbook Enhancement. The key motivation is that many failures in identity-preserving video generation are not isolated, but instead exhibit transferable patterns across different test samples. For example, complex actions often require clearer temporal decomposition and identity preservation requires more fine-grained appearance descriptions. Therefore, rather than manually rewriting the prompt for each individual sample, AESR aims to extract reusable experience from test-time generation feedback and convert it into general knowledge that can benefit subsequent samples.

Specifically, the Playbook stores reusable knowledge for prompt enhancement, including high-level strategies, general templates, and common failure modes. Strategies describe general generation principles, such as staged action descriptions, strengthened identity details, and the separation of subject motion from camera motion. Templates define how different prompt components should be organized, such as the ordering of human actions, human-object interactions, and target-state descriptions. Failure modes record common errors of the target model and their remedies, such as missing objects, weak identity constraints, or unclear final states. With this structured organization, the Playbook remains compact while providing clear guidance for subsequent prompt enhancement.

AESR initializes the Playbook from two complementary sources. The first source is the official prompting documentation of the target video model, which provides model-specific priors about how to organize identity, action, scene, style, and camera descriptions. By analyzing these documents, a VLM automatically learns the instruction organization favored by the target model and assigns the extracted experience into the three types of Playbook knowledge, helping the Playbook adapt to the input preferences of the target model. The second source is an external Playbook with the same format, learned from the HOI-Edit human-object interaction dataset and directly merged into the initial Playbook. Since identity-preserving video generation contains many human interaction events, such as contact, manipulation, movement, and state changes between humans and objects, we aim to draw reusable experience from relevant external human-object interaction data.

To enable the Playbook to accumulate test-domain-specific guidance from cross-sample feedback, AESR updates the initialized Playbook through an agentic loop using the generation results of the test set. The loop consists of four steps: generation, analysis, reflection, and consolidation. First, the generator enhances the input instruction with the current Playbook and produces a draft video. Second, the analyzer inspects the draft video and identifies its failures. Third, the reflector abstracts these concrete failures into reusable experience, such as stronger object constraints, clearer temporal staging, or more detailed identity descriptions. Finally, the curator reviews the newly extracted experience together with existing Playbook knowledge and incrementally writes only useful, non-redundant guidance back into the Playbook. In this way, prompt enhancement can automatically accumulate and reuse knowledge from feedback across test samples.

Finally, the updated Playbook is used to enhance prompts for challenging samples. The agentic prompt generator incorporates relevant strategies, templates, and pitfalls to produce an enhanced prompt better suited to the target model, which is then used to generate the initial draft video.

3.3 Sample-Level Visual Semantic Repair

Although global prompt enhancement improves the initial draft video, some errors are highly sample-specific and cannot be reliably fixed by adding more textual descriptions. For example, the generated video may miss visual elements required by the instruction, drift from the source identity, or misunderstand camera motions. To address these errors, AESR introduces Sample-Level Visual Semantic Repair, which uses visual diagnosis and explicit edited reference frames to guide video editing for repairing.

First, AESR uses a vision-language critic to compare the draft video with the original instruction and identify their mismatches. The critic produces a structured diagnosis, including the error type,

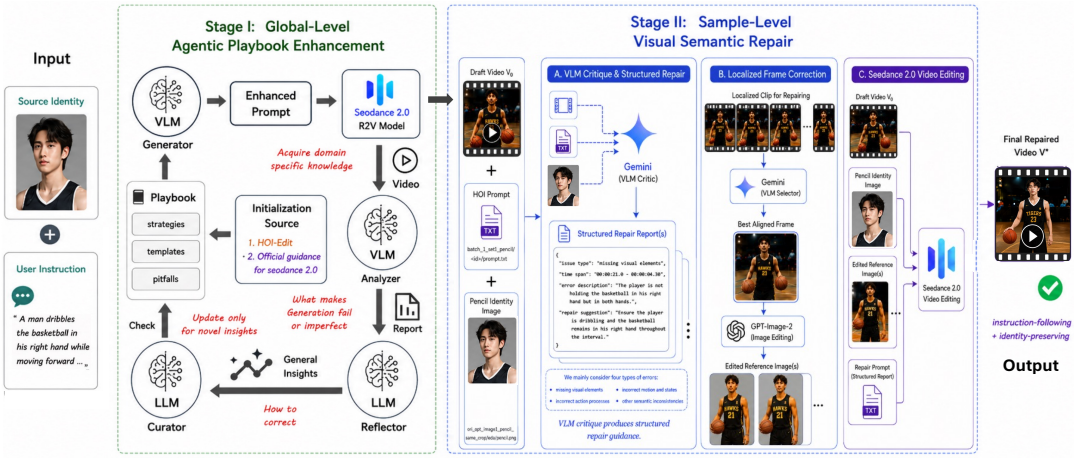

Figure 2: Two-stage AESR pipeline: Stage I shows global-level Agentic Playbook Enhancement, Stage II shows sample-level Visual Semantic Repair with pencil-style reference images.

problematic temporal segment, error description and repair suggestion. We mainly consider four types of errors: missing visual elements, incorrect motion end states, incorrect motion processes, and other semantic inconsistencies. Errors related to missing elements or incorrect end states are usually repaired with edited keyframes, while process-level errors, such as camera moving speed, or transition patterns, are repaired with textual editing instructions.

Second, for errors that require visual guidance, AESR selects a representative keyframe from the problematic segment and edits it according to the repair suggestion. The edited frame serves as an explicit visual target for the repaired video. This is useful when text alone is insufficient to describe spatial relations, object appearance, human pose, or identity-related details.

Finally, AESR constructs the repair inputs for Seedance 2.0 video editing, including a comprehensive editing instruction organized by the VLM [26], the edited keyframes corresponding to the target temporal segments, and the original identity reference image. The comprehensive instruction is generated based on the structured diagnosis, the time-annotated editing requirements, and the original video prompt. Seedance 2.0 then edits the draft video according to these inputs and produces the final repaired result.

4 Experiments

4.1 Challenge Setting and Evaluation Metrics

Track 1 of the Identity-Preserving Video Generation (IPVG) Challenge, namely Facial IPVG, aims to generate videos from textual prompts while consistently preserving the facial identity of a given reference subject throughout the generated video. The Facial IPVG test set contains 200 evaluation samples, each consisting of one reference image and one textual prompt for video generation. The official evaluation assesses generated videos from two primary aspects: identity preservation and video quality. Identity preservation measures whether the generated video maintains the facial characteristics of the reference identity across the entire video, while video quality evaluates perceptual aspects including visual fidelity, motion smoothness, and text alignment through a combination of automatic metrics and human evaluation. Detailed information about the challenge can be found on the official website.[1]

To obtain a more comprehensive assessment of generated videos and support the subsequent MoE-based selection strategy, we further evaluate each video from three perspectives: text alignment, identity consistency, and video quality. 1. For text alignment, previous methods based on CLIP [21] suffer from limitations such as the maximum token length of 77. In our evaluation, 89% of the textual prompts are truncated when computing CLIPScore. Therefore, we do not adopt CLIPScore as a primary evaluation metric. Instead, we use StoryScore [23] to measure story-level semantic consistency between the generated video and the prompt. Specifically, dense video captions are first generated and then compared with the original prompt based on semantic similarity. We further adopt GMEScore [39] as a complementary text-alignment metric. 2. For identity consistency, we calculate the similarity between generated video frames and the reference image using two facial recognition models, namely ArcFace [4] and CurricularFace [10], resulting in ArcScore and CurScore, respectively. 3. For video quality, we adopt two metrics from VBench [11], including Motion Smoothness and Imaging Quality, to evaluate motion quality and visual quality. All adopted metrics are normalized to the range of $[0, 1]$, where higher values indicate better performance.

4.2 MoE Selection Strategy

Different methods exhibit complementary strengths in text alignment, identity preservation, and video quality. To fully leverage these advantages and achieve the best challenge performance, we adopt a Mixture-of-Experts (MoE) strategy to select the optimal result for each sample. For each generated video, we first compute an overall score based on the evaluation metrics described above:

$$\text{Overall Score} = \sum_{i \in \mathcal{M}} w_i \cdot M_i, \tag{1}$$

[1] https://hidream-ai.github.io/ipvg-challenge-2026.github.io/

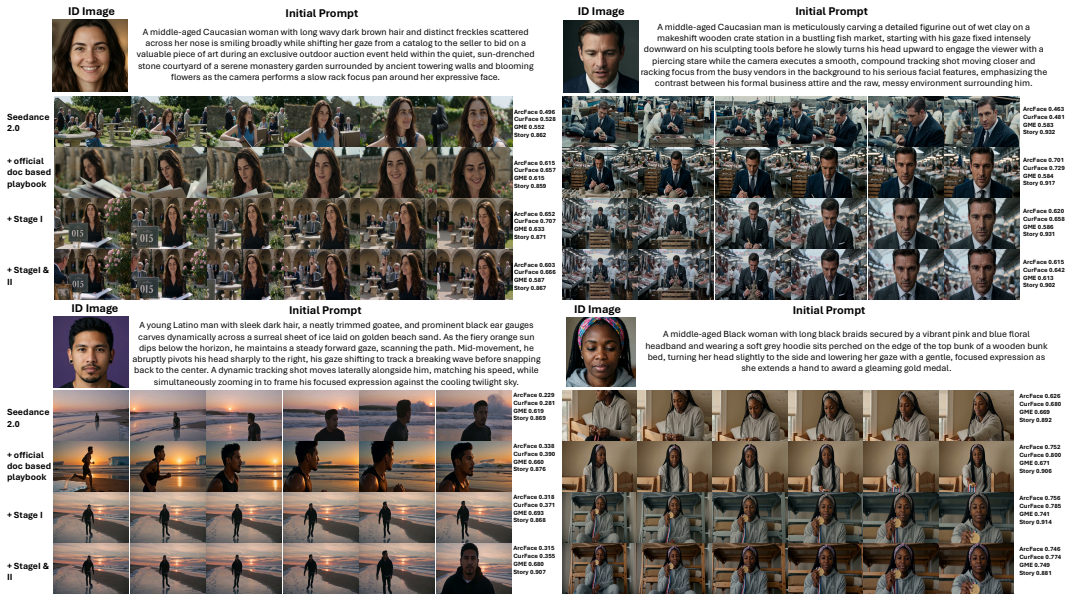

Figure 3: Qualitative Results.

where $\mathcal{M}$ denotes the set of evaluation metrics, M_i is the score of the corresponding metric, and w_i is its associated weight. For each test sample, we generate candidate videos using five different methods and calculate the overall score for each candidate. The video with the highest overall score is the final submission result.

4.3 Implementation Details

We use Seedance 2.0 as the main closed-source video generation and editing backbone. For each sample, the input consists of a source identity image and a natural-language instruction. Before calling the video APIs, we convert all identity-bearing images into pencil-style identity proxies. This representation preserves identity-relevant facial structure and shape while reducing the influence of irrelevant appearance factors such as background color and illumination, supporting stable batch API calls. This format is used consistently in both Stage I reference-to-video generation and Stage II visual semantic repair with edited image references. We generate 14-second videos for all samples, since the prompts often contain multiple visual elements and motion events; the longer duration gives the model more temporal capacity to cover the requested content. For Stage I, we first generate an initial round of videos using the original prompts from the test set, and analyze these results with the initialized Playbook to extract reusable experience. We then further include failure cases produced by different variants of our method in the Playbook update process, allowing it to accumulate test-domain knowledge from richer cross-sample feedback.

4.4 Qualitative Results

Figure 3 shows qualitative results of different AESR variants, including vanilla Seedance generation, prompt enhancement with playbook initialized with only official documentation, Stage-I global enhancement, and the final result with Stage-II visual semantic repair. The results show that introducing playbook knowledge improves alignment with both textual content and motion descriptions in the prompt. For example, in the bottom-left case, the prompt requires the subject to maintain a steady forward gaze while scanning the path, which is better captured after playbook enhancement. It also reduces generation artifacts, like the duplicated identity subject in the top-right case. The comparison between the second and third rows further shows the importance of incorporating test-domain experience into GPE. Compared with using only official prompting formats, domain experience better preserves complex actions and visual details. For example, the bottom-left case better reflects the action and scene relation described by "carves dynamically across a surreal sheet of ice"; the top-right case more accurately realizes the behavior of staring at the viewer; and the bottom-right case more reliably generates the required bunk-bed scene. This indicates that GPE not only learns model-preferred prompt formats, but also accumulates domain-level knowledge useful for complex instruction following. In Stage II, SVR further corrects local visual semantic errors that remain in the draft video. Although the draft may preserve the overall scene and motion structure, it can still miss key visual cues, subject/camera motion, or identity details. For example, the top-left case may miss the gaze transition required by "shifting her gaze from a catalog to the seller to bid"; the bottom-left case may not clearly reach the final camera state of "zooming in to frame his focused expression"; and the bottom-right case may drift on identity details such as "a vibrant pink and blue floral headband". SVR uses a VLM to locate problematic segments and design repair instructions, edits selected frames as explicit visual references, and guides the video editing model to repair incorrect local elements. These qualitative results show that global prompt enhancement and sample-wise visual repair are complementary: the former improves draft generation at the instruction level, while the latter provides direct visual guidance for fine-grained local errors.

Table 1: Average metrics on the selected 50 videos from Testset of IPVG2026.

Method	Text Alignment		Identity Consistency		Video Quality			OverallScore↑	HumanScore↑
	StoryScore↑	GMEScore↑	CurScore↑	ArcScore↑	Motion↑	Imaging↑	FID↓		
Vanilla generation	0.8965	0.6563	0.5165	0.4755	0.9933	0.6662	220.11	0.6803	3.42
Official prompt enhancement	0.8916	0.6608	**0.6253**	**0.5868**	0.9938	0.6814	**166.35**	**0.7266**	3.68
+ GPE	0.8972	0.6688	0.5952	0.5618	**0.9939**	0.6836	187.76	0.7179	3.82
+ text-only repair	0.8952	0.6704	0.6082	0.5732	0.9929	**0.6961**	175.25	0.7245	3.74
Full AESR	**0.8994**	**0.6776**	0.5762	0.5441	0.9934	0.6885	184.89	0.7129	**4.01**

Table 2: Comparison with existing identity-preserving video generation methods.

Methods	Text Alignment		Identity Consistency		Video Quality			OverallScore↑
	CLIPScore↑	GMEScore↑	CurScore↑	ArcScore↑	Motion↑	Imaging↑	FID↓	
Hailuo [19]	30.15	0.6087	0.0866	0.0705	**0.9880**	**0.6772**	243.54	0.4638
Phantom-14B [17]	30.17	0.6231	0.2459	0.2391	0.9810	0.6367	268.61	0.5266
VACE-14B [13]	30.00	0.6187	0.2034	0.1927	0.9750	0.6349	250.35	0.5063
TPIGE [7]	29.04	0.6040	0.2564	0.2424	0.9701	0.6265	252.88	0.5204
Seedance2 [22]	30.59	0.6560	0.2189	0.1997	0.9789	0.6348	265.95	0.5226
Seedance2 + HOI-Edit based playbook [6, 22]	**31.26**	**0.6598**	**0.2834**	**0.2662**	0.9824	0.6548	**239.81**	**0.5534**

Table 3: Official leaderboard of Track 1 – Facial Identity-Preserving Video Generation in the ACM MM 2026 Identity-Preserving Video Generation Challenge.

Rank	Team	Final Score↓
1	MIPL_Video	2.5
1	USTC-CMI	2.5
3	xuanyuan_fuxi	3.0

4.5 Quantitative Results

Before the release of the official IPVG 2026 test set, we first compared different base video generation models on an earlier IPVG 2025 evaluation set, and also tested the enhancement strategy based on the HOI-Edit Playbook. This evaluation was used for model selection before this year's official test results became available. As shown in Table 2, Seedance2 achieves the strongest text alignment and a competitive overall score among existing identity-preserving video generation methods. Since the new test set contains more complex textual descriptions, we adopt Seedance2 as the base generation and editing model for AESR. Furthermore, the HOI-Edit-based Playbook initialization strategy further improves the overall performance, demonstrating that human-object interaction experience can effectively enhance the quality of identity-preserving video generation on this benchmark. We then analyze AESR components on 50 selected videos from the IPVG 2026 test set in Table 1. Different paths show complementary strengths. Official prompt enhancement achieves the best CurScore and ArcScore, suggesting that model-provided prompting priors help identity-related metrics. GPE further improves StoryScore and GMEScore, showing that test-domain Playbook knowledge helps complex instruction following. Text-only repair improves Imaging Quality, while full AESR obtains the best text-alignment scores.

The component results also reveal limitations of current automatic metrics. Face-recognition metrics such as ArcScore can fluctuate noticeably even when identities look well preserved to humans. As shown in the top-right example of Fig. 3, the second and third rows look similar in identity preservation but have clearly different face scores. In contrast, text-alignment metrics more directly reflect whether complex instructions are followed, but their score range is smaller and thus less visible in the overall score. Therefore, following the Track 1 evaluation protocol of the ACM MM 2026 Identity-Preserving Video Generation Challenge, we further conduct human scoring that jointly considers identity preservation and video quality. Based on this human evaluation, full AESR achieves the best performance among the compared variants. The final leaderboard further supports our findings. As shown in Table 3, under the official protocol combining automatic metrics and human evaluation, our team MIPL_Video with AESR ranked first in Track 1 Facial Identity-Preserving Video Generation with a final score of 2.5, demonstrating the effectiveness of the full system.

5 Conclusion

We presented AESR, a lightweight interface-level framework for identity-preserving video generation with closed-source video models. AESR improves controllability without accessing model parameters by combining global agentic prompt enhancement with sample-level visual semantic repair. The former accumulates reusable generation knowledge from prompting priors and test-sample feedback, while the latter uses visual diagnosis and edited reference frames to correct local semantic and identity-related errors. A lightweight MoE strategy further selects reliable outputs across different generation and refinement paths. Extensive experiments demonstrate that AESR consistently improves identity preservation, semantic fidelity, and visual quality across diverse editing scenarios. Furthermore, our solution achieved first place in Track 1 of the official ACM MM 2026 Identity-Preserving Video Generation Challenge, validating the effectiveness and practical applicability of AESR.

Acknowledgements. This work was supported by the grants from the National Natural Science Foundation of China (62372014, 62525201, 62132001, 62432001), Beijing Nova Program and Beijing Natural Science Foundation (4252040, L247006).

References

[1] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. 2024. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233* (2024).

[2] Xinhao Cai, Yixuan Sun, Minghang Zheng, Qingchao Chen, Xin Jin, Song-Chun Zhu, and Yang Liu. 2026. Music-to-Dance Generation via Atomic Movements. In *European Conference on Computer Vision (ECCV)*. arXiv:2607.13978 [cs.CV]

[3] 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. 2025. Multi-subject Open-set Personalization in Video Generation. *arXiv preprint arXiv:2501.06187* (2025).

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

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

[6] Jiayi Gao, Qingchao Chen, Yuxin Peng, and Yang Liu. 2026. Taming I2V Models for Image HOI Editing: A Cognitive Benchmark and Agentic Self-Correcting Framework. In *Proceedings of the International Conference on Machine Learning*. To appear.

[7] Jiayi Gao, Changcheng Hua, Qingchao Chen, Yuxin Peng, and Yang Liu. 2025. Identity-Preserving Text-to-Video Generation via Training-Free Prompt, Image, and Guidance Enhancement. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 13751–13757.

[8] Jiayi Gao, Zijin Yin, Changcheng Hua, Yuxin Peng, Kongming Liang, Zhanyu Ma, Jun Guo, and Yang Liu. 2025. ConMo: Controllable Motion Disentanglement and Recomposition for Zero-Shot Motion Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7191–7200. doi:10.1109/CVPR52734.2025.00674

[9] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. 2026. LTX-2: Efficient Joint Audio-Visual Foundation Model. *arXiv preprint arXiv:2601.03233* (2026).

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

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

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

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

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

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

[16] Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, et al. 2025. Movie Weaver: Tuning-Free Multi-Concept Video Personalization with Anchored Prompts. *arXiv preprint arXiv:2502.07802* (2025).

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

[18] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. 2025. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487* (2025).

[19] MiniMax. 2024. Hailuo s2v-01. https://www.minimaxi.com/en/news/s2v-01-release/.

[20] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. 2026. A Survey on Fine-Grained Multimodal Large Language Models. *Chinese Journal of Electronics* 35, 2 (2026), 771–803. doi:10.23919/cje.2025.00.336

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

[22] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. 2026. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148* (2026).

[23] Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. 2026. Msvbench: Towards human-level evaluation of multi-shot video generation. *arXiv preprint arXiv:2602.23969* (2026).

[24] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent YF Tan, and Jiashi Feng. 2024. InstaDrag: Lightning Fast and Accurate Drag-based Image Editing Emerging from Videos. *arXiv preprint arXiv:2405.13722* (2024).

[25] Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Wangbo Yu, Chaoran Feng, Yatian Pang, Bin Lin, and Li Yuan. 2024. Cycle3D: High-quality and Consistent Image-to-3D Generation via Generation-Reconstruction Cycle. *arXiv preprint arXiv:2407.19548* (2024).

[26] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

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

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

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

[30] Fan Xie, Dan Zeng, Qiaomu Shen, and Bo Tang. 2025. A Comprehensive Survey on Text-to-Video Generation. *Chinese Journal of Electronics* 34, 4 (2025), 1009–1036. doi:10.23919/cje.2024.00.151

[31] Huiyu Xu, Shuaifan Jin, Zhibo Wang, Zhongjie Ba, and Tao Wei. 2025. PSA-NeRF: Personalized Spatial Attention Neural Rendering for Audio-Driven Talking Portraits Generation. *Chinese Journal of Electronics* 34, 6 (2025), 1821–1831. doi:10.23919/cje.2024.00.095

[32] Zhu Xu, Ting Lei, Zhimin Li, Guan Wang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2025. Weakly Supervised Dynamic Scene Graph Generation with Temporal-enhanced In-domain Knowledge Transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. arXiv:2508.04943 [cs.CV]

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

[34] Wangbo Yu, Chaoran Feng, Jiye Tang, Xu Jia, Li Yuan, and Yonghong Tian. 2024. EvaGaussians: Event Stream Assisted Gaussian Splatting from Blurry Images. *arXiv preprint arXiv:2405.20224* (2024).

[35] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. *arXiv preprint arXiv:2409.02048* (2024).

[36] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. 2024. MagicTime: Time-lapse Video Generation Models as Metamorphic Simulators. *arXiv preprint arXiv:2404.05014* (2024).

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

[38] Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. 2025. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618* (2025).

[39] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv preprint arXiv:2412.16855* (2024).