

000 001 002 003 004 005 006 007 008 009 010 IDENTITY-PRESERVING IMAGE-TO-VIDEO 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 GENERATION VIA REWARD-GUIDED OPTIMIZATION

031
032
033
Anonymous authors
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
Paper under double-blind review

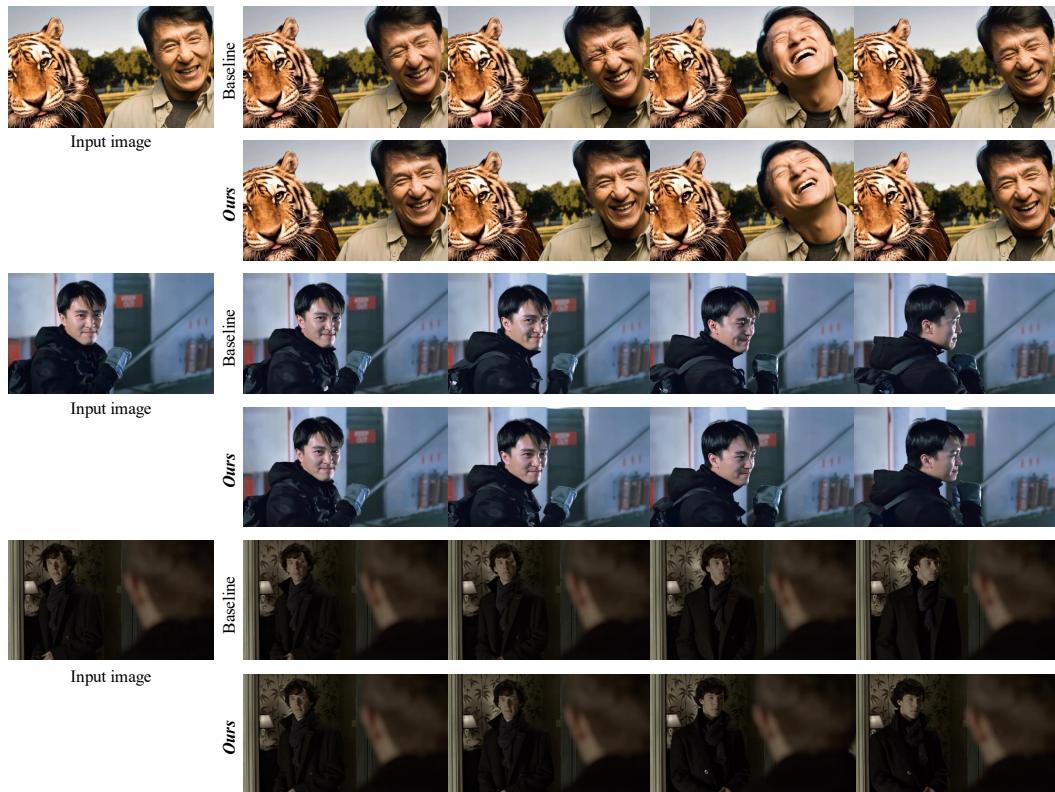
ABSTRACT

Recent advances in image-to-video (I2V) generation have achieved remarkable progress in synthesizing high-quality, temporally coherent videos from static images. Among all the applications of I2V, human-centric video generation includes a large portion. However, existing I2V models encounter difficulties in maintaining identity consistency between the input human image and the generated video, especially when the person in the video exhibits significant expression changes and movements. This issue becomes critical when the human face occupies merely a small fraction of the image. Since humans are highly sensitive to identity variations, this poses a critical yet under-explored challenge in I2V generation. In this paper, we propose Identity-Preserving Reward-guided Optimization (IPRO), a novel video diffusion framework based on reinforcement learning to enhance identity preservation. Instead of introducing auxiliary modules or altering model architectures, our approach introduces a direct and effective tuning algorithm that optimizes diffusion models using a face identity scorer. To improve performance and accelerate convergence, our method backpropagates the reward signal through the last steps of the sampling chain, enabling richer gradient feedback. We also propose a novel facial scoring mechanism that treats faces in ground-truth videos as facial feature pools, providing multi-angle facial information to enhance generalization. A KL-divergence regularization is further incorporated to stabilize training and prevent overfitting to the reward signal. Extensive experiments on Wan 2.2 I2V model and our in-house I2V model demonstrate the effectiveness of our method. Code will be released to advance the related research.

1 INTRODUCTION

Identity preservation is a widely popular and explored topic in text-to-video (T2V) generation. Numerous recent methods (He et al., 2024; Liu et al., 2025c; Yuan et al., 2025; Wang et al., 2025; Li et al., 2024; 2025; Song et al., 2025; Xie et al., 2025; Zhang et al., 2025; Zhong et al., 2025; Jiang et al., 2025) attempt to generate high-fidelity videos that maintain consistent identity from user-provided reference subjects, often by incorporating auxiliary identity modules and carefully curated datasets. However, it remains a persistent yet underexplored challenge in image-to-video (I2V) generation. Despite diffusion transformer-based models such as CogVideoX (Yang et al., 2024), HunyuanVideo (Kong et al., 2024), and Wan (Wan et al., 2025) have significantly improved image-to-video (I2V) generation quality, they still struggle to faithfully preserve the identity of the input image, especially human faces. A significant challenge arises when the reference image contains a low-resolution human face and the motion prompt involves large-scale movements (e.g., jumping, turning). Under these conditions, as illustrated in Fig. 1, the propagation of errors across frames results in progressive identity degradation, causing the generated subject to drift from the appearance in the initial frame.

One might consider explicitly injecting identity features into the model—a practice commonly used in subject-driven generation. However, since the identity is already fully encoded in the first frame, the challenge lies not in the absence of information but in its preservation. Merely adding more identity cues is a redundant exercise that does not tackle the root cause of temporal degradation. Moreover, these supervised finetuning (SFT) methods suffer from exposure bias, a discrepancy between training and inference dynamics caused by training on ground-truth inputs but self-generated inputs at inference, leading to error accumulation and identity drift. In addition, such architecturally-



080
081 Figure 1: Given a input image with human faces, our method can achieve superior identity consis-
082 tency in generated. **(Zoom in for best view)**

083
084 intrusive methods are inherently single-subject by design and struggle to scale to scenes with multi-
085 ple individuals, making them impractical for real-world applications. This raises a natural question:
086 Can we enhance the identity preservation capabilities of a general-purpose foundation I2V model
087 without altering its architecture or compromising its original abilities?

088 In this paper, we present an image-to-video generation framework designed to maintain identity
089 fidelity without extra identity networks or facial attribute injection but from the perspective of rein-
090 forcement learning. Prior works such as ReFL (Xu et al., 2023), DRaFT (Clark et al., 2023), Align-
091 Prop (Prabhudesai et al., 2023) and VADER (Prabhudesai et al., 2024) directly supervise the final
092 output of a diffusion model using differentiable reward functions. These reward-guided approaches
093 enable effective adaptation of diffusion models to generate videos that better align with task-specific
094 objectives, thereby achieving better alignment with desired outcomes. Our approach performs direct
095 reward optimization by leveraging a face identity model (Deng et al., 2019) as the reward model.
096 To eliminate exposure bias, we thus train on-policy by initializing trajectories from pure Gaussian
097 noise and backpropagates the identity reward through the last denoising steps of the sampling chain.
098 This trajectory-aware optimization yields robust identity improvements, faster convergence, and bet-
099 ter stability. To prevent reward hacking, we propose a robust multi-view facial scoring mechanism
100 that leverages facial feature pools constructed from videos, substantially improving generalization
101 to pose, expression, and lighting variations while mitigating the copy-paste phenomenon. We also
102 incorporate a KL-divergence regularization that constrains the deviation of the tuned model from
103 the original diffusion model at each gradient step. This multi-step regularization better preserves the
104 model’s original capabilities while enhancing identity consistency.

105 We evaluate our methodology on both Wan 2.2 I2V model (Wan et al., 2025) and our in-house 15B
106 I2V model using comprehensive quality and identity preservation metrics, comparing against state-
107 of-the-art methods for identity-preserving video generation. Extensive experiments demonstrate that
our approach generates high-quality videos with significantly improved identity consistency.

108 **2 RELATED WORK**

110 **Identity preservation in video generation.** Current identity-preserving video generation methods
 111 mainly focus on T2V generation, relying on additional modules to inject identity information. ID-
 112 Animator (He et al., 2024) and ConsisID (Yuan et al., 2025) encode ID-relevant information with
 113 a face adapter. FantasyID (Zhang et al., 2025) introduces 3D facial geometry to preserve structure.
 114 Concat-ID (Zhong et al., 2025) and Stand-In (Xue et al., 2025a) concatenate reference face tokens at
 115 the end of video token and employ 3D self-attention to fuse them. Some methods (Wang et al., 2025;
 116 Xie et al., 2025; Song et al., 2025) propose mixture of facial experts to dynamically fuse identity,
 117 semantic, and detail via a adaptive gate. PersonalVideo (Li et al., 2024) utilizes per-scene LoRA
 118 adapters for video personalization, but it requires finetuning the pretrained model for each new per-
 119 son during inference. Unlike T2V generation, I2V models naturally support first-frame conditioning
 120 for better controllability. Adding extra modules to improve identity preservation would disrupt the
 121 original structure of foundation models and increase complexity. To address this, our method di-
 122 rectly learns feedback from an identity reward function, producing videos that more faithfully meet
 123 identity-specific goals and maintain identity consistency.

124 **Reinforcement learning for diffusion models.** To tailor models to user needs, models are often
 125 fine-tuned using carefully curated datasets or perform reinforcement learning (Prabhudesai et al.,
 126 2024; Liu et al., 2025a;b; Xue et al., 2025b). Compared to supervised finetuning, which relies on the
 127 standard diffusion training loss, reinforcement learning offers a more direct and efficient optimiza-
 128 tion pathway by leveraging a reward model or curated preference data. While the former approach
 129 applies a general, pixel-level reconstruction signal, RL provides targeted feedback that directly ad-
 130 dresses high-level perceptual attributes, such as identity consistency. This allows for lightweight,
 131 precise adjustments to the model’s output distribution, correcting specific failure modes without ex-
 132 tensive retraining. Direct preference optimization (DPO) (Rafailov et al., 2023; Wallace et al., 2024)
 133 skips a reward model and learns directly from human preferences, simplifying the pipeline but only
 134 optimizes relative preference ranking, with gradients derived from “positive negative” differences
 135 and susceptible to confounding factors such as aesthetics and clarity, lacking absolute calibration.
 136 MagicID (Li et al., 2025) leverages DPO to improve identity consistency for T2V generation, but
 137 still requires per-scene LoRA adaptation and per-person fine-tuning of the pretrained model, which
 138 substantially limits its practicality. While Group-based Reward Policy Optimization (GRPO) (Shao
 139 et al., 2024) stabilizes RL via intra-group advantage normalization, its efficiency in I2V is limited
 140 by low response diversity within a single prompt, resulting in highly similar video samples that un-
 141 dermine the utility of group-wise advantage estimation. For our goal of enforcing facial identity
 142 consistency, a metric that can be reliably quantified by a reward model, a more direct reward feed-
 143 back learning paradigm is both better suited and more efficient. Therefore, we introduce the first
 144 facial reward feedback framework for I2V that directly optimizes gradients for identity preservation.

145 **3 PRELIMINARY**

147 **Diffusion Models.** Diffusion models are a class of generative models that learn to generate data by
 148 gradually denoising latents that are initialized from pure noise. The process consists of two phases:
 149 a forward diffusion process and a reverse denoising process. In the forward process, the data x_0
 150 is gradually corrupted by adding Gaussian noise over T time steps to obtain $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This
 151 process is modeled as a Markov chain with fixed variance schedule β_t :

152
$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}), t = 1, \dots, T. \quad (1)$$

154 The reverse process is learned by a neural network that approximates the reverse conditional distri-
 155 butions $p_\theta(x_{t-1}|x_t)$. Diffusion models learn the denoising network ϵ_θ by minimizing the following
 156 re-weighted variational lower bound of the marginal likelihood (Ho et al., 2020):
 157

158
$$\mathcal{L}_\theta = \mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2], \quad (2)$$

160 where $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. Once trained,
 161 samples are generated by reversing the diffusion process using the learned network ϵ_θ .

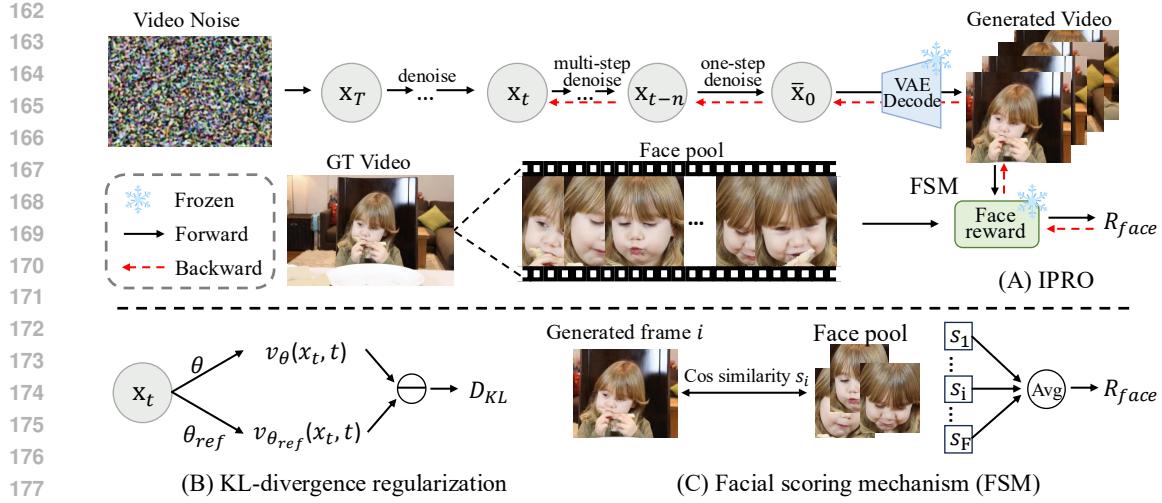


Figure 2: **Overview of our method.** (A) IPRO predicts \bar{x}_0 from the noise input x_T , and the prediction is visualized through a frozen VAE decoder and scored by a face reward model with our facial scoring mechanism (C). This reward signal is used to update the trainable parts of the model, thereby steering the generation process to produce videos with consistent identity. (B) We further incorporate a KL-divergence regularization to alleviate reward hacking.

4 METHODOLOGY

This work introduces a novel video diffusion model based on reinforcement learning to enhance identity preservation in I2V generation. The overall pipeline is illustrated in Fig. 2. Section 4.1 details reward-model strategies for maintaining facial identity consistency. Section 4.2 then introduces a facial scoring mechanism which treats faces in ground-truth videos as multi-view facial feature pools, thereby enhancing generalization and curbing copy-paste phenomenon. Finally, Section 4.3 incorporates a KL-divergence regularization to stabilize training and prevent overfitting to the reward signal.

4.1 IDENTITY-PRESERVING REWARD.

Facial reward feedback learning. We propose a direct and effective approach for adapting video diffusion models to preserve facial identity throughout generated video sequences, by leveraging a differentiable facial reward function. The core idea is to guide the generation process through an explicit optimization objective that maximizes face identity consistency between the generated video frames and ground-truth video frames. To achieve this, we compute a facial reward score R_{face} based on the cosine similarity between the Arcface (Deng et al., 2019) embeddings of the GT video and the generated video. ArcFace (Deng et al., 2019) has been widely recognized for its strong discriminative power in capturing fine-grained facial features, making it well-suited as a perceptual metric for identity preservation. Our goal is to fine-tune the parameters θ of a diffusion model such that videos generated by the sampling process maximize the differentiable reward function R_{face} :

$$J(\theta) = \mathbb{E}_{x_T \sim N(\mathbf{0}, \mathbf{I})} [R_{face}(\text{sample}(\theta, x_T))] \quad (3)$$

where $\text{sample}(\theta, x_T)$ denotes the sampling process from time $t = T \rightarrow 0$.

Prior identity-preserving approaches typically rely on supervised fine-tuning with teacher forcing, which induces exposure bias: during training the model conditions on intermediate states derived from real samples, whereas at inference it must condition on its own generated states. This will be magnified as a long-term consistency issue in video identity preservation: small gradual errors cannot be seen and corrected by the training process, ultimately manifested as identity features drifting over time, being attracted by the "average face", and identity jumps occurring under occlusion/extreme postures. By using facial reward feedback learning, starting from random noise and generating and learning in an inference manner, the training distribution can be aligned with

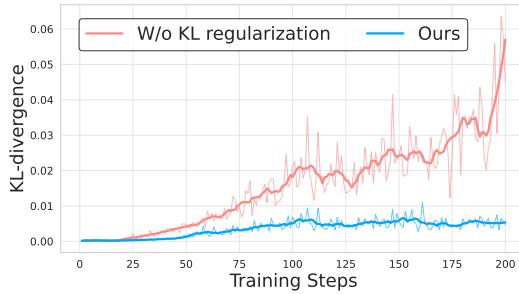
216
217
218
219
220
221
222
223
224
225

Figure 3: **The effect of KL regularization on KL divergence across training steps.** The model trained with KL regularization (blue) maintains a low and stable divergence, whereas the model without regularization (red) exhibits a rapid and volatile increase.

the inference distribution, directly optimizing the long-term identity consistency objective, thereby significantly alleviating the damage of exposure bias to identity preservation.

We efficiently learn the facial reward by backpropagating its gradient to update the diffusion model parameters θ . To reduce memory and accelerate optimization, we adopt the DRaFT (Clark et al., 2023) truncation strategy, backpropagating only through the last K sampling steps. Since identity preservation is driven by appearance details, and AlignProp (Prabhudesai et al., 2023) shows that later timesteps are most influential for these fine-grained details, this truncation does not degrade performance. This concentrates capacity where identity cues reside, reduces memory/compute, and prevents drift of global dynamics, and truncated gradient is given by:

$$\nabla_{\theta} R_{face}^K = \sum_{t=0}^K \frac{\partial R_{face}}{\partial x_t} \cdot \frac{\partial x_t}{\partial \theta} \quad (4)$$

4.2 FACIAL SCORING MECHANISM

We introduce a facial scoring mechanism (FSM) that treats the faces in the ground-truth video as a reference pool. Previous methods (Li et al., 2024; Xie et al., 2025; Yuan et al., 2025) computes identity loss by measuring face similarity either between generated frames and a reference image or between generated frames and their time-aligned ground-truth frames. However, the former encourages copy-paste phenomenon (the human in the generated video strictly maintain the expression in the given first frame) and suppresses facial expression diversity, while the latter provides weak supervision under teacher forcing, since the generated frame is already close to its corresponding ground-truth frame. In contrast, thanks to generating videos from random noise, we can treat the faces of all frames in the ground-truth video as a pool and, for each generated frame, calculate the average face similarity to all ground-truth frames. This objective encourages the people in the generated video to resemble those in the ground-truth video while allowing natural variation across frames and providing a broader, more informative reward signal. This similarity serves as the reward signal for our facial reward model, formalized as:

$$s_i = \frac{1}{F} \sum_{j=1}^F \cos(\phi(\hat{x}_i), \phi(x_j)), \quad (5)$$

$$R_{face} = \frac{1}{F'} \sum_{i=1}^{F'} s_i, \quad (6)$$

where $X = \{x_1, \dots, x_F\}$ are ground-truth frames, $\hat{X} = \{\hat{x}_1, \dots, \hat{x}_{F'}\}$ are generated frames, ϕ is the face encoder, and \cos denote cosine similarity. s_i is the average similarity between each generated frame i and all ground-truth frames, and R_{face} is the score of the face reward model.

4.3 KL-DIVERGENCE REGULARIZATION

We propose a multi-step KL-divergence regularization over the reverse-time sampling trajectory to stabilize training and mitigate reward hacking. Rather than optimizing solely by reward model, we

270

271
272
Table 1: **Quantitative comparisons.** Our method achieves more consistent face similarity than the
baseline, without compromising its performance on other dimensions.

273

Method	FaceSim↑	SC↑	BC↑	AQ↑	IQ↑	TF↑	DD↑	MS↑
In-house I2V model	0.4769	0.9768	0.9777	0.6641	0.7291	0.9895	8.93	0.9943
<i>W/ reward model</i>	0.6960	0.9811	0.9810	0.6641	0.7259	0.9913	8.31	0.9951
Wan 2.2 A14B (Wan et al., 2025)	0.5780	0.9508	0.9705	0.6588	0.7273	0.9676	19.45	0.9804
<i>W/ reward model</i>	0.6942	0.9536	0.9720	0.6607	0.7253	0.9690	19.17	0.9813

278



279

280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
Figure 4: **Qualitative comparison before and after integrating our framework.** Our method
achieves more stable generation and superior identity preservation compared to the baseline.300
301
constrain the sampling process from x_T by penalizing deviations between the optimized model θ
and the original model θ_{ref} at each gradient step:

$$\begin{aligned}
 D_{KL}(p_\theta(x_{0:T}) || p_{\theta_{ref}}(x_{0:T})) &= \sum_{t=1}^K \omega_t D_{KL}(p_\theta(x_{t-1}|x_t) || p_{\theta_{ref}}(x_{t-1}|x_t)) \\
 &= \sum_{t=1}^K \omega'_t \|v_\theta(x_t, t) - v_{\theta_{ref}}(x_t, t)\|^2.
 \end{aligned} \tag{7}$$

308
309
310
311
312
where $p_\theta(x_{0:T})$ is the path distribution induced by the sampling procedure starting from x_T to x_0 , ω_t
and ω'_t are step weights, and $v_\theta, v_{\theta_{ref}}$ denote the velocity parameterizations. As shown in Fig. 3, the
model trained with KL-divergence regularization maintains a low and stable KL divergence, which
indicates the regularization updates are well-constrained, preventing large-scale shifts.

313

5 EXPERIMENTS

314

5.1 EXPERIMENTAL SETUP

315

316

317
318
319
320
321
322
323
Implementation Details. We utilize both our in-house I2V model (15B dense model with both
MMDiT block and single DiT block) and Wan2.2 27B-A14B (Wan et al., 2025) as our image-to-
video foundational model. For Wan2.2 27B-A14B (Wan et al., 2025), we keep all modules frozen
except the low-noise expert model. In order to improve training efficiency, we employ Wan2.2-
Lightning, the distilled version of Wan2.2 (Wan et al., 2025) which requires only 8 steps without the
need of Classifier-free guidance (CFG). For the training process, we employ the Adam optimizer
configured with a learning rate of 2e-5, and train 100 steps with a batch size of 64. The truncation
gradient step K is 4, the facial reward weight is 0.1, and the KL-divergence loss weight is 1.

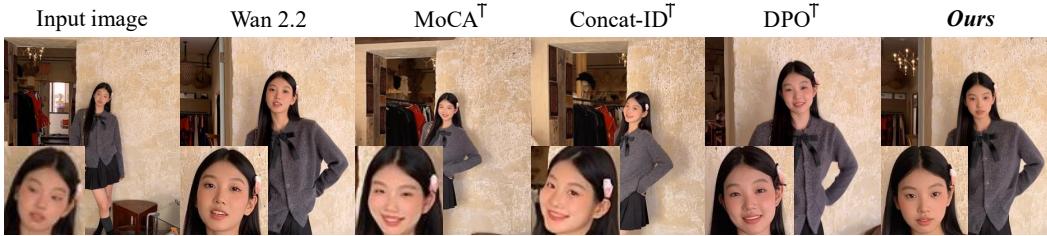


Figure 5: **Qualitative comparison with other methods.** Our method preserves identity information more faithfully than other methods.

Table 2: **Comparison with other methods.** Our method achieves the highest face similarity among all compared methods.

Method	FaceSim \uparrow
Wan 2.2 (Wan et al., 2025)	0.5780
+ MoCA † (Xie et al., 2025)	0.5820
+ Concat-ID † (Zhong et al., 2025)	0.6056
+ DPO † Rafailov et al. (2023)	0.6284
+ Ours	0.6942

Table 3: **Ablation study on different training frameworks.** Our method outperforms SFT and CLIP reward in preserving face similarity.

Method	FaceSim \uparrow
Wan2.2 (Wan et al., 2025)	0.5780
+ SFT	0.6392
+ CLIP	0.6099
+ Ours	0.6942

Datasets. We collect high-quality videos with 960p resolution from the Internet and detect faces using SCRFID (Guo et al., 2021). To emphasize small-face scenarios, we retain clips in which the largest face bounding box per frame does not exceed 100×100 pixels. We discard videos with insufficient face coverage ($< 40\%$ of frames containing a detectable face) and use Qwen2.5-VL (Bai et al., 2025), a vision–language model, to remove videos where faces are occluded by objects (e.g., phones, masks). For the remaining data, we extract face embedding of each face frame in the video.

Evaluation metrics. We evaluate the identity consistency on a small-face evaluation set of 600 scenes, by computing face similarity (FaceSim) between the generated frames and the input image. In addition, to comprehensively evaluate the overall quality of generated videos, we also report VBench-I2V (Huang et al., 2024) metrics on its official evaluation set, including Subject Consistency (SC), Background Consistency (BC), Aesthetic Quality (AQ), Imaging Quality (IQ), Time Flickering (TF), Dynamic Degree (DD), and Motion Smoothness (MS).

5.2 COMPARISONS

Comparisons on facial reward. To validate the effectiveness of our facial reward model, we use the in-house I2V model and Wan2.2 I2V 27B-A14B (Wan et al., 2025) as baselines and compare their performance before and after integrating our framework. As shown in Table 1, the FaceSim metric improves markedly: face similarity for the in-house I2V model increases by 45.9%, and Wan2.2 (Wan et al., 2025) improves by 20.1%. For general video quality metrics in VBench-I2V (Huang et al., 2024), introducing the reward model does not degrade the models’ original capabilities. Fig. 4 further provides qualitative comparisons, where our method exhibits more stable generation and stronger identity preservation.

Comparisons with other methods. Lacking I2V identity-preservation baselines, we adapt two state-of-the-art T2V methods, MoCA † (Xie et al., 2025) and Concat-ID † (Zhong et al., 2025), to I2V model (Wan et al., 2025) and compare them with our method. MoCA † (Xie et al., 2025) supervises identity consistency through a latent space identity loss, whereas Concat-ID † (Zhong et al., 2025) appends reference face tokens to the video token sequence and employs 3D self-attention for fusion. As shown in Table 2, our method achieves markedly better facial consistency than these two methods, indicating that facial reward-guided learning achieves better alignment with the desired outcomes. This advantage is further validated qualitatively in Fig. 5, where our method exhibits superior identity retention. We further evaluate multi-person scenarios in Appendix A.3, where our method naturally generalizes and demonstrates improved consistency.

378

379

380

Table 4: **Ablation study on reward hacking.** Our method enhances facial consistency without noticeable hacking.

381

382

383

384

	Wan2.2	W/o KL	W/o FSM	Ours
FaceSim \uparrow	0.5780	0.7544	0.7388	0.6942
Hacking \downarrow	7%	58%	52%	10%

385

386

387



388

389

390

391

392

393

394

395

Figure 6: **Visual comparison on different training frameworks.** Our method achieves more stable generation and superior identity preservation compared to others.

396

397

398

399



400

401

402

403

404

405

406

407

408

409

410

411

412

413

Figure 7: **Ablation study on reward hacking.** Without KL-divergence regularization or the FSM module, the generated video overly adheres to the input image, resulting in facial rigidity and reward hacking phenomenon. However, our method enables accurate, expressive prompt-following behavior, such as opening eyes.

414

415

416

417

418

419

420

421

Comparison with DPO. By bypassing an explicit reward model, DPO (Rafailov et al., 2023; Wallace et al., 2024) learns directly from human preferences and simplifies the pipeline. Yet the objective is purely relative: gradients come from preferred–rejected likelihood differences, leaving the method vulnerable to confounds and without calibrated, absolute utility. However, reward-guided learning uses pretrained identity encoders to directly provide dense, calibrated signals. Fig. 5 and Table 2 show that our reward-based method outperforms DPO in terms of face similarity.

5.3 ABLATION STUDY

422

423

424

425

426

427

SFT. To evaluate the effectiveness of our reward-based training framework, we compare it with supervised fine-tuning (SFT). Specifically, we incorporate facial similarity as a loss function into the training objective as an auxiliary loss during supervised fine-tuning. This encourages the model to preserve identity during generation. Despite this explicit supervision on face similarity, we observe that SFT still struggles to preserve face identity. In contrast, our reward-based approach leads to more robust and consistent performance, as shown in Fig. 6 and Table 3.

428

429

430

431

Reward model. We employ the ArcFace model (Deng et al., 2019) as our facial embedding extractor and compute the similarity between embeddings of generated and ground-truth videos. To validate the effectiveness of our reward model, we also experiment with the CLIP image encoder (Radford et al., 2021) to extract facial embeddings and compute similarity scores as the reward signal. However, as shown in Fig. 7 and Table 3, the CLIP-based encoder underperforms compared to Ar-

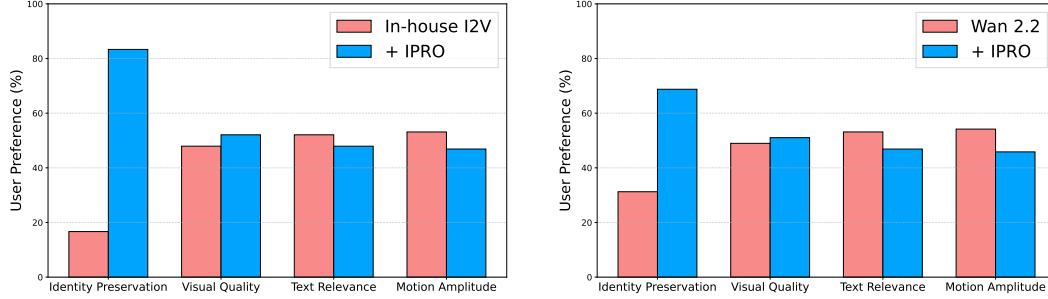
432 cFace, indicating that ArcFace is better suited for capturing fine-grained facial identity features in
 433 our framework.

434 **Different choices of gradient steps.** To evaluate the effectiveness of selecting final gradient steps
 435 for backpropagation, we conduct experiments comparing high-noise and low-noise initial steps. As
 436 shown in Tab. 5, using final (low-noise) gradient steps achieves better identity consistency.
 437

438 **Reward hacking.** Simply using the face reward model to optimize the model can easily lead to
 439 reward hacking, where the generated faces in the video over-adhere to the input image, resulting in
 440 stiff faces, lack of expression, and poor response to prompt. To address this, we propose a novel
 441 facial scoring mechanism (FSM) and a multi-step KL-divergence regularization to enhance general-
 442 ization. As shown in Fig. 7, our method ensures facial consistency while providing better response
 443 to prompt. We further leverage the state-of-the-art video understanding VLM Gemini 2.5 Pro (Co-
 444 manici et al., 2025) to identify facial reward hacking in generated videos, reporting the hacking rate
 445 on the evaluation set as a metric. As shown in Table 4, while ablating the KL regularization and
 446 FSM leads to higher facial similarity, it significantly increases the reward hacking phenomenon.
 447 More details on the hacking metric are provided in the appendix.

448 5.4 HUMAN EVALUATIONS

450 We conduct a user study to assess our method from a human perspective. Specifically, we randomly
 451 sample 50 images from the small-face evaluation set and we use different approaches to generate
 452 videos with identical settings. During the study, participants are shown the input image and two
 453 videos (ours vs. baseline) in randomized order. A total of 96 volunteers are invited to select the video
 454 that performs better in terms of Identity Preservation, Visual Quality, Text Alignment, and Motion
 455 Amplitude, or choose “no preference” if uncertain. The results, summarized in Fig. 8, indicate that
 456 our method achieves better identity preservation than baselines, while performing comparably in
 457 other dimensions.



468 **Figure 8: User study.** User study results show that our IPRO significantly outperforms both the
 469 in-house I2V model and Wan 2.2 I2V in identity preservation.
 470

471 6 CONCLUSION

472 In this paper, we present IPRO, a reinforcement learning-based framework for image-to-video dif-
 473 fusion models that substantially improves identity preservation without modifying the underlying
 474 architecture or introducing auxiliary identity modules. Our method addresses key shortcomings of
 475 prior approaches by (i) leverage a facial reward model to enhance identity consistency (ii) introduc-
 476 ing a robust facial scoring mechanism that aggregates multi-view facial features from ground-truth
 477 videos, and (iii) stabilizing learning with a multi-step KL regularization that preserves the founda-
 478 tion model’s original capabilities. We hope that our work can bring identity preservation in I2V and
 479 its reward framework into the sight of a broader community and motivate further research.
 480

481 **Limitations and Future Work.** This work focuses on facial identity preservation in image-to-video
 482 generation. However, preserving identity also entails consistency of non-facial attributes (jewelry,
 483 accessories, clothing), which remain underexplored. We plan to investigate the unified identity
 484 reward model that covers these aspects in future work.
 485

486 REFERENCES
487

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

493 Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models
494 on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

495 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit
496 Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
497 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
498 bilities. *arXiv preprint arXiv:2507.06261*, 2025.

499

500 Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin
501 loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision*
502 and pattern recognition

503 Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redis-
504 tribution for efficient face detection, 2021.

505

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

509

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

512 Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianx-
513 ing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua
514 Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative
515 models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-
516 nition*, 2024.

517

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

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

523

524 Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi
525 Wu, and Deng Cai. Personalvideo: High id-fidelity video customization without dynamic and
526 semantic degradation. *arXiv preprint arXiv:2411.17048*, 2024.

527 Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. Magicid:
528 Hybrid preference optimization for id-consistent and dynamic-preserved video customization.
529 *arXiv preprint arXiv:2503.12689*, 2025.

530

531 Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan,
532 Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv*
533 *preprint arXiv:2505.05470*, 2025a.

534

535 Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin
536 Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv*
537 *preprint arXiv:2501.13918*, 2025b.

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

540 Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-
 541 image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.
 542

543 Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak.
 544 Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

545 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
 546 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
 547 models from natural language supervision. In *International conference on machine learning*, pp.
 548 8748–8763. PMLR, 2021.

549 Rafael Raffailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
 550 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances
 551 in neural information processing systems*, 36:53728–53741, 2023.

552 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
 553 Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathemati-
 554 cal reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

555 Wenhui Song, Hanhui Li, Jiehui Huang, Panwen Hu, Yuhao Cheng, Long Chen, Yiqiang Yan,
 556 and Xiaodan Liang. Lavieid: Local autoregressive diffusion transformers for identity-preserving
 557 video creation. *arXiv preprint arXiv:2508.07603*, 2025.

558 Bram Wallace, Meihua Dang, Rafael Raffailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam,
 559 Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using
 560 direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision
 561 and Pattern Recognition*, pp. 8228–8238, 2024.

562 Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu,
 563 Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative
 564 models. *arXiv preprint arXiv:2503.20314*, 2025.

565 Yuji Wang, Moran Li, Xiaobin Hu, Ran Yi, Jiangning Zhang, Han Feng, Weijian Cao, Yabiao Wang,
 566 Chengjie Wang, and Lizhuang Ma. Identity-preserving text-to-video generation guided by simple
 567 yet effective spatial-temporal decoupled representations. *arXiv preprint arXiv:2507.04705*, 2025.

568 Qi Xie, Yongjia Ma, Donglin Di, Xuehao Gao, and Xun Yang. Moca: Identity-preserving text-to-
 569 video generation via mixture of cross attention. *arXiv preprint arXiv:2508.03034*, 2025.

570 Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao
 571 Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation.
 572 *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

573 Bowen Xue, Qixin Yan, Wenjing Wang, Hao Liu, and Chen Li. Stand-in: A lightweight and plug-
 574 and-play identity control for video generation. *arXiv preprint arXiv:2508.07901*, 2025a.

575 Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei
 576 Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv
 577 preprint arXiv:2505.07818*, 2025b.

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

581 Shanghai Yuan, Jinfu Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and
 582 Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceed-
 583 ings of the Computer Vision and Pattern Recognition Conference*, pp. 12978–12988, 2025.

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

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

594
595

A APPENDIX

596
597

A.1 THE USE OF LARGE LANGUAGE MODELS(LLMs)

598
599
600
601
602
603

We used large language models solely for language polishing (grammar and spelling correction, improving clarity and flow, and shortening long sentences). Specifically, LLMs were not used to conceive ideas, design the method or experiments, implement code, analyze results, write proofs, or conduct literature searches or citation selection. All LLM suggestions were reviewed and substantively edited by the authors; all technical statements, formulas, algorithms, and conclusions are authored and verified by us.

604
605
606
607

We also used a vision-language model (VLM) as an auxiliary metric to detect “reward hacking” in generated videos—i.e., over-adherence to the first frame that yields unnaturally rigid faces, suppressed expressions, and poor responsiveness to prompts. The vision-language (VLM) was given the following prompt:

608
609
610
611
612

“Analyze the following video and evaluate whether it exhibits signs of ‘reward hacking’ in the context of facial consistency optimization. Specifically, determine if the face remains too rigidly consistent with the first frame throughout the video, resulting in unnatural or overly static facial appearance, lack of natural expression changes, minimal facial dynamics, or absence of expected motion (e.g., subtle shifts due to speech, emotion, or camera movement). Consider the following aspects:

613
614
615
616
617
618
619
620
621
622

- Facial Motion: Is there realistic and natural variation in facial expressions or muscle movements across frames?
- Consistency vs. Stiffness: While the identity remains consistent, does the face appear unnaturally frozen or overly stabilized?
- Contextual Appropriateness: Given the content (e.g., talking, emotional expression, head movement), is the level of facial motion appropriate?
- Visual Artifacts: Are there any signs of blending, warping, or smoothing artifacts that suggest aggressive enforcement of similarity to the first frame?

623
624

Is the video over-optimized (i.e., reward hacking)? Only answer yes or no”

625
626

A.2 FACE SCORE VARIATION WITH TRAINING STEPS

627
628

The face score on the training set increases steadily throughout training, indicating that the identity reward effectively enhances facial identity preservation.

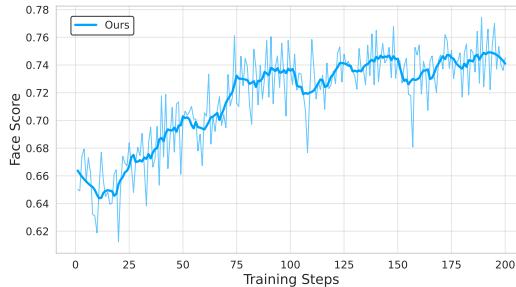
629
630
631
632
633
634
635
636
637
638639
640
641

Figure 9: **Face score evolution during training in training set.** Scores increase steadily with training steps, indicating that the face reward effectively improves identity consistency.

642
643

A.3 VISUALIZATIONS OF MULTI-PERSON SCENARIOS

644
645
646
647

Compared to injecting facial features directly into the base model, our RL-based approach not only requires no architectural modifications or additional modules, but also naturally generalizes to multi-person scenarios. Concat-ID[†] (Zhong et al., 2025) injects reference face tokens to the video token sequence and employs 3D self-attention for feature fusion. However, as the number of characters

in a scene increases, this method struggles to maintain consistent identity representations across frames. In contrast, as shown in Fig. 10, our method achieves superior identity consistency in complex multi-person settings.

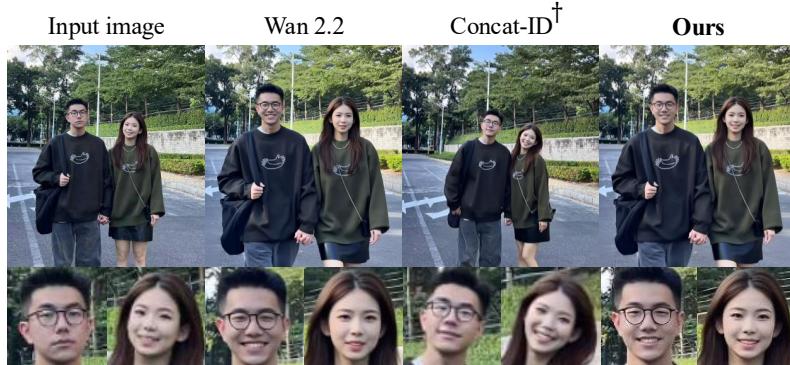


Figure 10: **Visual comparison on multi-person scenes.** Our method preserves identity consistency across frames better than other methods.

A.4 SOCIAL IMPACTS

Our RL-based image-to-video diffusion method, which improves human identity preservation, offers clear benefits for creative production, accessibility, and scientific research. However, it may also amplify risks of non-consensual deepfakes and misinformation. To mitigate these risks, we will work continuously with legal and ethics experts and impacted communities to iteratively advance safety measures while maintaining the technology's benefits.