Identity-Preserving Image-to-Video Generation via Reward-Guided Optimization

Supplementary Material

To see the dynamic effect of our method and visual comparisons, please refer to our supplementary video. This document includes the following contents:

- Visualizations on Wan 2.2 5B
- Visualizations of multi-person scenarios
- Hacking metric
- Details on truncation gradient step K
- Ablation of loss weights
- Local temporal constraints
- Different face recognition models
- Face score variation with training steps
- Social impacts.

A. Visualizations on Wan 2.2 5B

In the main paper, we provide quantitative comparisons of Wan 2.2 5B with and without our reward optimization. Here we present qualitative visual comparisons, as shown in Fig. 1. Our method achieves better identity consistency preservation compared to the Wan 2.2 5B baseline.

Figure 1. **Visualizations on Wan 2.2 5B.** Our method preserves identity consistency better than Wan 2.2 5B baseline.

B. Visualizations of multi-person scenarios

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† [4] 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. 2, our method achieves superior identity consistency in complex multi-person settings.

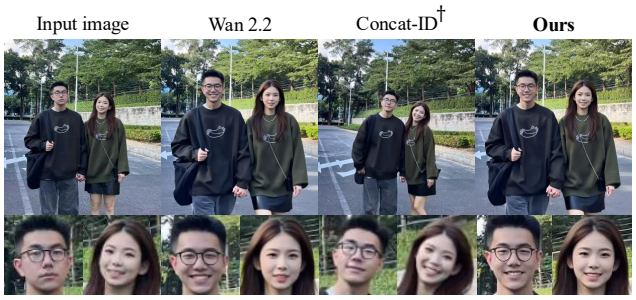

Figure 2. **Visual comparison on multi-person scenes.** Our method preserves identity consistency across frames better than other methods.

C. Hacking metric

We used a vision-language model (VLM) [1] 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:

> 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:
>
> - 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?

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

D. Details on truncation gradient step K

We found that as the truncated gradient step size increased, the improvement in identity consistency began to level off. Taking computational cost into account, we selected K=4.

K	1	2	3	4	5
FaceSim↑	0.6593	0.6712	0.6835	0.6942	0.6966

E. Ablation of loss weights

We supervise the model with a combination of face reward loss and KL-divergence regularization loss. In the main paper, we discuss how KL-divergence regularization prevents large-scale updates and distribution shift. Here we analyze the impact of different loss weights for these two components, as shown in Table 1. When the weight ratio is $\lambda_2/\lambda_1 = 0$ or $\lambda_2/\lambda_1 = 1$, KL regularization is either negligible or insufficient, resulting in a high rate of reward hacking. Conversely, when the weight ratio reaches $\lambda_2/\lambda_1 = 100$, KL regularization becomes overly restrictive, preventing effective reward-driven model optimization. Based on comprehensive evaluation, we adopt $\lambda_1 = 0.1, \lambda_2 = 1$ during training, corresponding to a weight ratio of $\lambda_2/\lambda_1 = 10$.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{Reward} + \lambda_2 \mathcal{L}_{KL}, \tag{1}$$

Table 1. **Ablation study on loss weights.**

λ_2/λ_1	0	1	10	100
FaceSim↑	0.7544	0.7215	0.6942	0.6371
Hacking↓	58%	52%	10%	9%

F. Local temporal constraints

We have conducted experiments to incorporate local temporal rewards. However, the results indicate that such constraints such constraints provide limited improvement in our task for several reasons. Firstly, the primary challenge is not pixel smoothness but identity fidelity. While VGG-based perceptual loss or noise consistency can improve low-level pixel smoothness and reduce jitter, they lack the discriminative power to maintain fine-grained facial identity. Secondly, local rewards inherently tend to propagate errors. If the identity in frame $t-1$ starts to degrade, a local temporal reward will supervise frame t to match that degraded identity. In contrast, our face pool acts as a global anchor, providing a stable and diverse reference set and covering a wide range of angles and expressions. Lastly, current I2V foundation models (e.g., Wan 2.2) already possess robust temporal attention mechanisms that handle local frame-to-frame consistency.

	Wan 2.2	LTC	Ours	Ours + LTC
FaceSim↑	0.5780	0.6128	0.6942	0.6961
Time Flickering↑	0.9676	0.9692	0.9690	0.9699

G. Different face recognition models

We employ the ArcFace model [2] as our facial embedding extractor and compute the similarity between embeddings of generated and ground-truth videos. To further validate the robustness of our reward framework, we additionally evaluate FaceSim using another face recognition model, CurriculumFace [3]. As shown in Table 2, our method consistently performs well across different face recognition models.

Table 2. **Different face recognition models.**

	Wan 2.2 A14B	***W/ reward model***
FaceSim-Arc↑	0.5780	**0.6942**
FaceSim-Cur↑	0.5805	**0.6989**

H. Face score variation with training steps

As shown in Fig. 3, the face score on the training set increases steadily throughout training, indicating that the identity reward effectively enhances facial identity preservation.

I. 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.

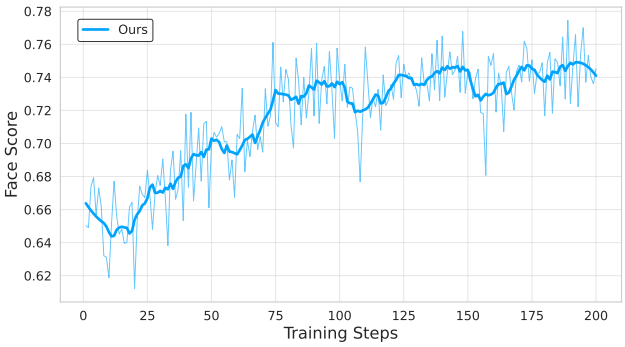

Figure 3. **Face score evolution during training in training set.** Scores increase steadily with training steps, indicating that the face reward effectively improves identity consistency.

References

[1] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2

[3] 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 *CVPR*, 2020. 2

[4] Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-id: Towards universal identity-preserving video synthesis. *arXiv preprint arXiv:2503.14151*, 2025. 1