# IDENTITY-GRPO: OPTIMIZING MULTI-HUMAN IDENTITY-PRESERVING VIDEO GENERATION VIA REINFORCEMENT LEARNING

**Xiangyu Meng**[1*], **Zixian Zhang**[2*], **Zhenghao Zhang**[1†], **Junchao Liao**[1], **Long Qin**[1], **Weizhi Wang**[1]
[1]Alibaba Group [2]Fudan University

## ABSTRACT

While advanced methods like VACE and Phantom have advanced video generation for specific subjects in diverse scenarios, they struggle with multi-human identity preservation in dynamic interactions, where consistent identities across multiple characters are critical. To address this, we propose Identity-GRPO, a human feedback-driven optimization pipeline for refining multi-human identity-preserving video generation. First, we construct a video reward model trained on a large-scale preference dataset containing human-annotated and synthetic distortion data, with pairwise annotations focused on maintaining human consistency throughout the video. We then employ a GRPO variant tailored for multi-human consistency, which greatly enhances both VACE and Phantom. Through extensive ablation studies, we evaluate the impact of annotation quality and design choices on policy optimization. Experiments show that Identity-GRPO achieves up to 18.9% improvement in human consistency metrics over baseline methods, offering actionable insights for aligning reinforcement learning with personalized video generation. Code and weights are publicly available at Identity-GRPO.

## 1 INTRODUCTION

The scalability of diffusion transformer architectures Peebles & Xie (2023) with full 3D attention has propelled the field toward high-quality visual content generation, enabling a broad spectrum of downstream applications, such as identity-preserving video generation or motion-controlled video generation Zhang et al. (2025a). Identity-preserving video generation, which aims to create high-fidelity videos with consistent human identity, has become a particularly prominent direction. Early methods like ConsisID Yuan et al. (2025b) and MovieGen Polyak et al. (2024) demonstrated excellence in single-identity video generation. More recently, advancements such as ConceptMaster Huang et al. (2025), Video Alchemist Chen et al. (2025), Tora2 Zhang et al. (2025b), Phantom Liu et al. (2025c), SkyReels-A2 Fei et al. (2025), and VACE Jiang et al. (2025) have extended this paradigm to multi-human generation, fundamentally transforming human-centric content creation pipelines.

However, for multi-human identity-preserving video generation (MH-IPV) task, models must simultaneously satisfy complex interactive instructions from text prompts while maintaining identity consistency across the entire video sequence. Even state-of-the-art models like VACE Jiang et al. (2025) and Phantom Liu et al. (2025c) often erroneously prioritize overall composition similarity over individual identity preservation. For example, these models may swap facial features between characters to fulfill a prompt (e.g., "Two people dancing with synchronized movements but distinct outfits"), resulting in coherent motion patterns but catastrophic identity misalignment.

Recent advances in Text-to-Video (T2V) generation have demonstrated that diffusion transformer models can better follow complex text prompts through post-training stages involving Reinforcement Learning from Human Feedback (RLHF) Ouyang et al. (2022). Methods like DiffusionDPO Wallace et al. (2024) and DenseDPO Wu et al. (2025), adapted from Direct Preference Optimization (DPO) Rafailov et al. (2023), offer a simpler alternative to RLHF by directly optimizing policies that align with human preferences under classification objectives. More recently, FlowGRPO Liu et al. (2025a) and DanceGRPO Xue et al. (2025) have shown more promising results in aligning

[*]Equal contribution.

video outputs with human preferences through Group Relative Policy Optimization (GRPO) Guo et al. (2025).

Despite the success of T2V alignment strategies, applying GRPO to identity-preserving video generation remains underexplored. A critical bottleneck lies in the absence of fine-grained reward models that can disentangle identity preservation from dynamic motion requirements. For T2V, high-level semantic guidance from text allows reward metrics like HPS-v2.1 Wu et al. (2023), CLIP score Radford et al. (2021), and VideoAlign Liu et al. (2025b) to effectively capture human preferences. However, in multi-human identity-preserving video generation scenarios, where entities must maintain unique visual characteristics while adhering to complex spatial-temporal interactions, existing reward signals like ArcFace Deng et al. (2019) exhibit high correlations with non-identity-related factors across frames, leading to what we term the "copy-and-paste" effect. Additionally, multimodal conditioning inputs, which consist of multiple reference images paired with corresponding text prompts, introduce significant variance in GRPO training due to substantial differences among group samples.

In this paper, we propose Identity-GRPO, the first preference-driven alignment strategy for MH-IPV scenarios. To address the critical challenge of reward modeling in this domain, we construct a preference-based multi-human similarity dataset (approximately 15k annotated examples) through a hybrid pipeline that strategically curates and evaluates generated video pairs from five advanced video generation models using a semi-automated framework combined with human labeling. This approach ensures scalability by surpassing manual annotation limits while maintaining strict alignment with human preferences via quality-controlled filtering. Leveraging this dataset, we train a specialized reward model capable of capturing fine-grained identity-consistent quality differences between paired video samples. To optimize practical implementation, we systematically evaluate key hyperparameters, including group size, clip ratio, prompt design, and initialization noise, and identify the most effective configuration for this task. Our experiments demonstrate that this reward model successfully aligns both VACE Jiang et al. (2025) and Phantom Liu et al. (2025c) models with human preference criteria, establishing a robust foundation for preference-driven MH-IPV.The contributions are summarized as below:

- We construct a large-scale high-quality annotated dataset (15k examples) for multi-human identity-preserving video generation, synthesized from five advanced video generation models. This dataset serves as a foundational resource for evaluating identity-preservation capabilities.
- We propose an identity consistency reward model and systematically investigate GRPO training configurations, analyzing the impact of design choices on reward signal effectiveness in multi-human scenarios.
- Our comprehensive experiments demonstrate that Identity-GRPO outperforms baseline methods (VACE and Phantom) by up to 18.9% and 6.5% on identity-consistency metrics, providing novel insights into the integration of reinforcement learning with customized visual synthesis for complex multi-human generation tasks.

## 2 METHOD

### 2.1 PREFERENCE DATASET CONSTRUCTION

To facilitate the training of our identity-consistent reward model, we construct a preference dataset for identity preservation by leveraging the OpenHumanVid Li et al. (2025a) dataset which is a human-centric video generation dataset. Recognizing that manual annotation of preference is prohibitively costly and labor intensive, thereby limiting the scalability of dataset, we curate an extensive, automatically labeled preference dataset to augment the reward model's training process.

#### 2.1.1 DATA-FILTERED PIPELINE

To construct a high-quality dataset for our task, we designed a multi-stage data filtering and augmentation pipeline. Initially, we address scene complexity and identity ambiguity by programmatically filtering the OpenHumanVid dataset. Qwen3 Yang et al. (2025a) is employed to parse captions and limit videos to a maximum of three subjects, followed by using Qwen2.5-VL Bai et al. (2025) to retain

only samples with clear, frontal human faces for maximum facial detail preservation. To prevent background interference in reference images, GroundingDINO Liu et al. (2024b) and SAM2 Ravi et al. (2024) are utilized to precisely segment and extract all human subjects. A key challenge in MH-IPV is the "copy-and-paste" problem Liu et al. (2025c); Yuan et al. (2025a). To mitigate this, we synthesize varied reference images of each subject from multiple perspectives using Flux.1 Kontext Labs et al. (2025). This forces the model to learn a robust identity representation independent of a single pose. The yield rate of the images edited by Flux.1 Kontext exceeds 80%, meeting our data requirements. The final pipeline yields a curated dataset of triplets formatted as `<edited reference image, prompt, original video>`.

#### 2.1.2 AUTOMATIC ANNOTATION

We use our data-filtered pipeline to generate inputs for five personalized video models (VACE-1.3B/14B Jiang et al. (2025), Phantom-1.3B/14B Liu et al. (2025c) and MAGREF Deng et al. (2025)) to synthesize videos. From these, two types of preference pairs are constructed: (1) `<original video, generated video>`. The generated video can be naturally viewed as a degraded representation of the original video. So, the original video is assigned the default preference. (2) `<generated video 1, generated video 2>`. Both videos are synthesized under identical conditions. The preference label is determined by a majority vote over multiple inferences from VLMs Bai et al. (2025); Wang et al. (2025a); Liu et al. (2025e). To ensure that preferences reflect identity consistency rather than extraneous factors, such as semantic relevance to the prompt, we filter the dataset by using a SOTA multi-modal vector model GME Zhang et al. (2024b). For each pair, we compute the text-video similarity for both videos. Pairs with a significant discrepancy in similarity scores are discarded. This process yields 10,000 automatically annotated preference pairs, which we denote as `Auto-labeled` data.

#### 2.1.3 HUMAN ANNOTATION

Pairwise annotation is a well-established method for capturing relative preferences. It has been demonstrated Liu et al. (2025b) to achieve higher inter-annotator agreement compared to pointwise methods that rely on the assignment of absolute scores. Therefore, in this work, we employ a pairwise human annotation framework to construct our preference dataset. We provide annotators with guidelines to ensure a clear understanding of the preference criteria. The guidelines instructed annotators to determine the preference labels by evaluating each video based on three key points: facial consistency with the reference image, visual quality over the whole video, and alignment with the text prompt. In the annotation interface, annotators were shown a set of reference images, a text prompt, and two generated videos based on these conditions. They were asked to select the video that better preserves the subjects' identities, with the options: *A better / Ties / B better*. A "*Ties*" vote indicates no discernible difference in identity preservation. Reference images were sourced from CelebA-HQ Karras et al. (2017) and our filtered OpenHumanVid data, with prompts also from OpenHumanVid. Each pair was evaluated by three annotators, and the final label was determined by a majority vote. Like automatic annotation, we filtered human-labeled dataset by using the GME Zhang et al. (2024b) model to ensure data quality. This process resulted in 5,000 high-quality, human-annotated preference pairs, denoted as `Human-labeled` data.

### 2.2 IDENTITY-CONSISTENT REWARD LEARNING

#### 2.2.1 PREFERENCE MODELING

We adopt Qwen2.5-VL-3BBai et al. (2025) as the reward model. We employ Bradley-Terry-with-Ties (BTT) Rao & Kupper (1967); Liu et al. (2024a) as the object function. BTT explicitly models the tied preferences and conforms to the category of labels in our dataset. In the identity-preserving video generation, two videos $y_A$ and $y_B$ are generated under identical conditions (i.e., the same reference images $x$ and prompt $t$). The BTT model defines the probabilities of each possible preference as follows:

$$P(y^A \succ y^B \mid x, t) = \frac{e^{r(x,t,y^A)}}{e^{r(x,t,y^A)} + \theta e^{r(x,t,y^B)}} \tag{1}$$

$$P(y^B \succ y^A \mid x, t) = \frac{e^{r(x,t,y^B)}}{\theta e^{r(x,t,y^A)} + e^{r(x,t,y^B)}} \tag{2}$$

$$P(y^B = y^A \mid x, t) = \frac{(\theta^2 - 1)e^{r(x,t,y^A)}e^{r(x,t,y^B)}}{(e^{r(x,t,y^A)} + \theta e^{r(x,t,y_B)})(\theta e^{r(x,t,y^A)} + e^{r(x,t,y^B)}))}, \tag{3}$$

where $r$ is the optimized reward function, and $\theta \geq 1$ is a parameter that controls the tendency towards ties, with a larger $\theta$ indicating a higher probability of ties. Following prior work Liu et al. (2024a), we set $\theta = 5$ and optimize the BTT model by minimizing negative log-likelihood loss:

$$\mathcal{L}_{\text{BTT}} = -\mathbb{E}_{(x,t,y^A,y^B)\sim D}\Big[\sum_{i\in\{y^A\succ y^B, y^B\succ y^A, y^A=y^B\}} \mathbb{I}(i) \log P(i \mid x, t)\Big], \tag{4}$$

where $\mathbb{I}(i)$ denotes the indicator function.

### 2.2.2 Learning with Auto-labeled Data

We propose a two-stage training methodology designed to make full use of the automatically annotated data. Our approach first refines the large, automatically labeled dataset through a consistency-based filtering protocol and then employs a joint training procedure with a dynamic, smooth sampling strategy.

The data refinement process unfolds as follows. A preliminary reward model, denoted as `RM_teacher`, is trained exclusively on the high-quality human-annotated dataset (`Human-labeled` data). This model serves as a proxy for human preferences. `RM_teacher` is used to infer preference scores for all pairs in the larger, automatically labeled dataset (`Auto-labeled` data). We filter the `Auto-labeled` dataset by retaining only the samples where the preference predicted by `RM_teacher` aligns with the original automatic label. This filtering protocol yields a refined dataset, `Filtered auto-labeled` dataset, which comprises approximately 48% of the original data in `Auto-labeled` dataset.

In the training stage, we perform joint training on the `Human-labeled` dataset and the `Filtered auto-labeled` dataset. To mitigate abrupt distributional shifts when combining these heterogeneous data sources, we introduce a smooth sampling strategy. This strategy employs a cosine scheduling mechanism to dynamically adjust the sampling proportion, $\alpha_t$, for `Filtered auto-labeled` dataset at each training step $t$:

$$\alpha_t = 0.5 \cdot (1 + \cos(\frac{\pi t}{T})), \tag{5}$$

where $t \in [0, T]$, $T$ is the total number of training steps. The value of $\alpha_t$ monotonically decreases from 1 to 0 as $t$ progresses from 0 to $T$. Consequently, the training curriculum commences with a high proportion of data from `Filtered auto-labeled` dataset and gradually transitions to prioritize the high-fidelity, human-annotated data from `Human-labeled` dataset. This method ensures training stability and maximizes the utility of both data sources. Our experimental results validate that this approach culminates in a superior-performing identity-consistent reward model.

## 2.3 Identity-GRPO Training

In this section, we first introduce the preliminary concepts of reinforcement learning, then present the sampling process of Identity-GRPO, and finally describe the method we propose to address the training challenges of GRPO in the MH-IPV task.

### 2.3.1 Reinforcement Learning

Following DDPO Black et al. (2023), the denoising process of the rectified flow can be formulated as a Markov Decision Process (MDP):

$$\begin{aligned} &s_t \triangleq (\boldsymbol{c}, t, \boldsymbol{z}_t),\ \pi(a_t|s_t) \triangleq p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{c}),\ P(s_{t+1}|s_t, a_t) \triangleq (\delta_{\boldsymbol{c}}, \delta_{t-1}, \delta_{\boldsymbol{z}_{t-1}}) \\ &a_t \triangleq \boldsymbol{z}_{t-1},\ R(s_t, a_t) \triangleq \begin{cases} r(\boldsymbol{z}_0, \boldsymbol{c}), & \text{if } t = 0 \\ 0, & \text{otherwise} \end{cases},\ \rho_0(s_0) \triangleq (p(\boldsymbol{c}), \delta_T, \mathcal{N}(\mathbf{0}, \mathbf{I})), \end{aligned} \tag{6}$$

where $s_t$ is the state at step $t$, $\pi(a_t|s_t)$ is the policy, $P(s_{t+1}|s_t, a_t)$ is the deterministic transition, $a_t$ is the action, $R(s_t, a_t)$ is the reward which is only given at the final step, and $\rho_0(s_0)$ is the initial state distribution.

In Identity-GRPO, the generative model samples a set of videos $\{\boldsymbol{z}_0^i\}_{i=1}^G$ from noise samples $\{\boldsymbol{z}_1^i\}_{i=1}^G$. Our reward model then assigns scores $r(\boldsymbol{z}_0^i, \boldsymbol{c})$ to this set of generated videos, and the advantage of each sample is computed by:

$$\hat{A}_t^i = \frac{r(\boldsymbol{z}_0^i, \boldsymbol{c}) - \text{mean}(\{r(\boldsymbol{z}_0^i, \boldsymbol{c})\}_{i=1}^G)}{\text{std}(\{r(\boldsymbol{z}_0^i, \boldsymbol{c})\}_{i=1}^G)}. \tag{7}$$

Then the policy model is optimized by maximizing the following objective function:

$$\mathcal{J}(\theta) = \mathbb{E}_{\{\boldsymbol{z}^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|\boldsymbol{c})} \frac{1}{G} \sum_{i=1}^{G} \frac{1}{T} \sum_{t=0}^{T-1} \left( \min \left( \rho_t^i(\theta) \hat{A}_t^i, \text{clip} \left( \rho_t^i(\theta), 1-\varepsilon, 1+\varepsilon \right) \hat{A}_t^i \right) \right), \tag{8}$$

where $\rho_t^i(\theta) = \frac{p_\theta(\boldsymbol{z}_{t-1}^i|\boldsymbol{z}_t^i, \boldsymbol{c})}{p_{\theta_{\text{old}}}(\boldsymbol{z}_{t-1}^i|\boldsymbol{z}_t^i, \boldsymbol{c})}$, and $\varepsilon$ is a hyper-parameter for clip-range. For the sake of notational simplicity, the KL divergence term is omitted in the formulation.

### 2.3.2 Video Sampling

In the current field of image and video generation, flow matching models have become the dominant approach due to their solid theoretical foundation and exceptional performance. Given a prompt $\boldsymbol{c}_p$ and its corresponding reference images $\boldsymbol{c}_r$ as conditions, flow matching models use transformer framework to predict the velocity field $\boldsymbol{v}_\theta(\boldsymbol{z}_t, \boldsymbol{c}, t)$ and:

$$\begin{aligned} \boldsymbol{c} &= \{\boldsymbol{c}_p, \boldsymbol{c}_r\} \\ \mathrm{d}\boldsymbol{z}_t &= \boldsymbol{v}_\theta(\boldsymbol{z}_t, \boldsymbol{c}, t)\mathrm{d}t, t \in [0, 1]. \end{aligned} \tag{9}$$

Following previous works Liu et al. (2025a); Xue et al. (2025), we derive the corresponding reverse SDE formulation:

$$\mathrm{d}\boldsymbol{z}_t = \left[ \boldsymbol{v}_\theta(\boldsymbol{z}_t, \boldsymbol{c}, t) + \frac{\sigma_t^2}{2t} (\boldsymbol{z}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{z}_t, \boldsymbol{c}, t)) \right] \mathrm{d}_t + \sigma_t \mathrm{d}\boldsymbol{w}, \tag{10}$$

where $\sigma_t$ introduces the stochasticity during sampling, and we use $\sigma_t = t$ in our paper.

### 2.3.3 Training Stability Strategies

How to sample a suitable set of videos for subsequent advantage calculation is crucial to the success of GRPO training. However, in MH-IPV, compared to the T2V task, the model input includes multiple modalities, which introduces significant variance and makes it difficult for sampled videos to support stable GRPO training. To address this, we propose two strategies in Identity-GRPO: prompt finetuning and initial noise differentiation. In addition, the variance between different modalities requires a larger number of videos to be used in a single parameter update — analogous to using a larger batch size in standard training. Otherwise, GRPO training is prone to instability or collapse.

**Prompt Finetuning.** In MH-IPV, different models exhibit varying sensitivity to the discrepancy between the prompt and the reference image. For example, VACE tends to follow the content of the prompt, whereas Phantom tends to preserve the content of the reference image. This discrepancy leads to reward models being unable to provide reasonable outputs, introducing significant instability during GRPO training. To address this issue, we employed Qwen2.5-VL-7B Bai et al. (2025) to generate prompts that contain accurate descriptions of the characters in the reference images.

**Initial Noise Differentiation.** In MH-IPV, due to the constraints imposed by the reference image, it is difficult to create significant differences in identity consistency among videos within the same group solely relying on the randomness introduced by the SDE. This limits the exploration space required for effective reinforcement learning training. Therefore, during sampling, we employ different initialization noises to amplify the diversity between generated videos.

Table 1: Preference accuracy on our multi-human identity-preserving preference benchmark. Red-colored font indicates improvement over the baseline when using few-shot examples.

| | ArcFace | Qwen2.5VL-3B | Qwen2.5VL-72B | InternVL3.5-38B | Ours |
|---|---|---|---|---|---|
| Accuracy | 0.772 | $0.430_{+4.4\%}$ | $0.657_{+3.6\%}$ | $0.685_{+2.0\%}$ | 0.890 |

Table 2: Ablation study on data source and data sampling strategies.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Human-labeled data | ✓ | | ✓ | | ✓ | ✓ | ✓ |
| Auto-labeled data | | ✓ | ✓ | | | | |
| Filtered auto-labeled data | | | | ✓ | ✓ | ✓ | ✓ |
| Random Sampling | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| Staged Sampling | | | | | | ✓ | |
| Smooth Sampling | | | | | | | ✓ |
| Accuracy | 0.853 | 0.664 | 0.812 | 0.785 | 0.877 | 0.824 | 0.890 |

**A Larger Video Number.** In Identity-GRPO, due to the variance between different modalities, using a limited number of videos in a single parameter update can lead to training instability or collapse. To increase the video number under limited computational resources, we reduce the resolution and number of frames in the sampled videos. Experiments demonstrate that this approach effectively stabilizes the training of Identity-GRPO.

## 3 Experiments

### 3.1 Reward Learning

**Training Setting.** We use LoRA Hu et al. (2022) to update the identity-consistent reward model. The vision encoder is also optimized to capture fine-grained details of human identity. Empirically, we set the learning rate to 2e-6. The global batch size is set to 32. We sample videos at 2 fps, with a resolution of approximately 832 $\times$ 480 pixels, which is the default resolution of VACE Jiang et al. (2025) and Phantom Liu et al. (2025c).

**Main Results.** We establish a preference benchmark for MH-IPV, which comprises 500 video samples meticulously annotated by human evaluators. We compare our reward model against ArcFace Deng et al. (2019), Qwen2.5-VL Bai et al. (2025) and InternVL3.5 Wang et al. (2025a). The quantitative results are presented in Table 1. VLMs, such as Qwen2.5VL and InternVL3.5, fail to effectively predict human preferences regarding subject consistency. This is particularly evident in small-scale model, such as the Qwen2.5VL-3B, which serves as our base model. Furthermore, the accuracy of ArcFace Deng et al. (2019) is insufficiently high (below 0.8), indicating its poor alignment with human perception and rendering it unsuitable as a reward signal for human feedback. In contrast, our model outperforms all other methods, showcasing its effectiveness in evaluating identity consistency in MH-IPV task.

**Ablation Study.** As shown in Table 2, we analyzed the effectiveness of human-annotated data and filtered auto-labeled dataset. The model trained exclusively on human annotated data achieves an accuracy of 0.853, which we establish as our baseline. In contrast, when trained solely on the raw, unfiltered auto-labeled data, the model's accuracy drops significantly to 0.664. After applying our filtering process, training on the filtered automated data yields an accuracy of 0.785, a substantial improvement of 0.12. Furthermore, when augmenting the human-annotated dataset with automated data, the model's accuracy reaches 0.877 with filtered auto-labeled data, surpassing the baseline. Conversely, using unfiltered automated data results in a lower accuracy of 0.812. These findings collectively underscore the necessity of our data filtering mechanism of auto-labeled data for improving data quality and subsequent model performance.

During the training phase, we explored several dynamic sampling strategies to effectively combine the filtered auto-labeled dataset and human-annotated dataset. These sampling strategies include:

Table 3: Evaluation results on our test set. **Bold**: Best Performance.

| Method | ID-Consistency ↑ | Aesthetics ↑ | GmeScore ↑ | Winning Rate ↑ |
|---|---|---|---|---|
| VACE-1.3B | 2.606 | 45.58% | 67.98% | 24% |
| VACE-1.3B+Identity-GRPO | **3.099** | **47.56**% | **68.35**% | **76**% |
| Phantom-1.3B | 3.809 | 44.13% | 67.56% | 37% |
| Phantom-1.3B+Identity-GRPO | **4.056** | **47.03**% | **68.47**% | **63**% |

(1) `Random Sampling` denotes that the two datasets are pooled and samples are drawn randomly from the combined collection for training. (2) `Staged Sampling` denotes that training stage is divided into two sequential stages of equal length in terms of training steps. The first stage involves training the model exclusively on a large-scale, filtered auto-labeled dataset. The objective of this stage is to enable the model to learn generalizable knowledge and robust feature representations of preferences. In the second stage, the training curriculum shifts to a smaller, high-quality dataset annotated by humans. This stage is designed to align the model's outputs with real human preferences, leveraging the precision of human-provided data. (3) `Smooth Sampling`, which denotes this approach utilizes a probabilistic cosine scheduling mechanism to dynamically adjust the sampling ratio of Filtered auto-labeled and Human-labeled data at each training step. The proportion of samples from Filtered auto-labeled data is high at the beginning of training and gradually shifts towards a higher proportion of samples from Human-labeled data. The results in Table 2 indicate that `Smooth Sampling` outperforms all the other three strategies, with a notable accuracy improvement of 0.066 over `Staged Sampling`. This is because `Smooth Sampling` mitigates the abrupt distributional shift that occurs in `Staged Sampling` strategy. In the latter, this shift causes the parameters from the first stage to become sub-optimal for the new data distribution in the second stage, forcing the model to learn the new feature space and leading to a periodic performance degradation.

### 3.2 GRPO Training

**Experiment Setup.** We utilized the reference images from the preference dataset we constructed, along with the corresponding prompts generated by our prompt finetuning strategy, for GRPO training. The test set consists of 100 samples, which are sufficient to cover a diverse range of scenarios. For quantitative evaluation, we compare the identity consistency reward score (abbreviated as ID-Consistency) produced by the reward model. To evaluate visual quality and text relevance, we use two evaluation metrics from OpenS2V Yuan et al. (2025a): AestheticScore christophschuhmann (2024) and GmeScore Zhang et al. (2024b). Additionally, we conduct a user study in which participants are asked to select videos with higher ID-Consistency between baseline and Identity-GRPO. The winning rate is then calculated based on user preferences.

**Training details.** Following previous works Liu et al. (2022); Xue et al. (2025), we employ a reduced sampling timestep during training. Specifically, we use 25 steps for sampling and 50 steps for evaluation. The experiments were conducted on 8 A100 GPUs. The sampled videos have a frame length of 33, with a resolution of $416 \times 240$. The number of sampling groups is set to 16, with a group size $G = 8$. The clip range $\varepsilon$ is set to $1 \times 10^{-3}$. We incorporate LoRA with $\alpha = 32$ and rank $= 32$.

**Main Results.** Figure 1 illustrates the performance of Identity-GRPO on the test set throughout the training process. Both models exhibited a clear upward trend during the GRPO training process. As detailed in Table 3, Identity-GRPO improved ID-Consistency of VACE-1.3B and Phantom-1.3B by up to 18.9% and 6.5%, respectively. The advantage of Identity-GRPO was also evident in the user study, achieving winning rates of 76% and 63%. Furthermore, Identity-GRPO enhances the aesthetic scores of both VACE-1.3B and Phantom-1.3B, while the GME scores remain largely stable throughout the fine-tuning process. These results demonstrate that Identity-GRPO effectively strengthens identity preservation without compromising visual quality or prompt-following performance.

Qualitative comparison results are provided in Figure 2. In several cases, the baseline model produced outputs that clearly mismatched the reference image, whereas Identity-GRPO consistently maintained high identity alignment.

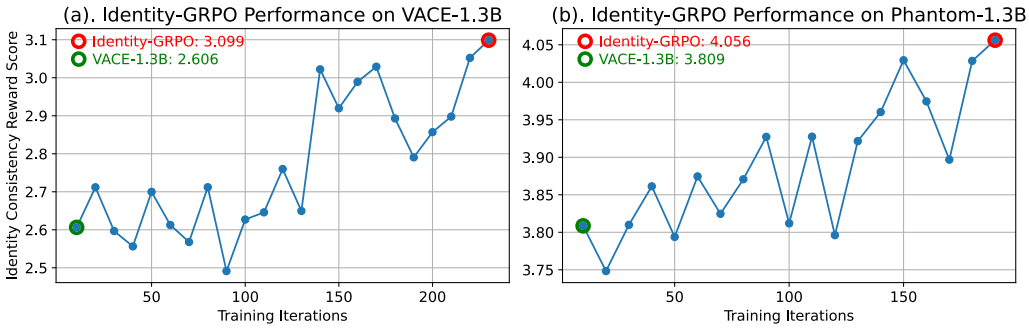


Figure 1: (a) and (b) respectively show the performance curves of Identity-GRPO on VACE-1.3B and Phantom-1.3B. Both exhibit a clear upward trend.

Table 4: Ablation study on group number and initial noise. **Bold**: Best Performance.

| Group Num | Different Initial Noise | ID-Consistency |
|---|---|---|
| 4 | | 2.588 |
| 4 | ✓ | 2.749 |
| 16 | | 2.718 |
| 16 | ✓ | **3.099** |

Table 5: Ablation study on different post-training methodologies. **Bold**: Best Performance.

| | VACE-1.3B | Supervised Fine-Tuning | Identity-GRPO |
|---|---|---|---|
| ArcFace similarity | 0.235 | 0.261 | **0.298** |
| ID-Consistency | 2.606 | 2.774 | **3.099** |

**Ablation Study.** As discussed in the Method section, due to the multimodal input conditions in the MH-IPV task, using a large number of videos in each parameter update, along with diverse initial noises, is crucial for achieving stable GRPO training. Therefore, we conducted ablation studies on the impact of the video number and the initial noise on the VACE-1.3B training process of Identity-GRPO. The number of videos is equal to the number of sampling groups multiplied by the group size. In our experiments, we fix the group size at 8 and vary the number of sampling groups. As shown in Table 4, which presents the final results under four training settings. When the video number is insufficient, GRPO training becomes unstable and consistently fluctuates. Using the same initialization noise when sampling a group of videos restricts the exploration space of GRPO, preventing the reward from increasing.

Leveraging our curated high-quality dataset, we conducted a comparative analysis of two post-training methodologies: Supervised Fine-Tuning (SFT) and Identity-GRPO. To evaluate identity preservation, we employed the ArcFace similarity score as metrics. As shown in Table 5, the baseline VACE-1.3B achieved a score of 0.235, and SFT reached 0.261, whereas our proposed Identity-GRPO yielded a superior score of 0.298. These results demonstrate that Identity-GRPO significantly outperforms SFT in maintaining identity consistency, validating its efficacy in enhancing the generation quality of video models when trained on high-quality datasets.

## 4 Conclusion

This paper presents Identity-GRPO, the first preference-driven alignment strategy for multi-human identity-preserving video generation (MH-IPV), addressing critical challenges in maintaining identity consistency during complex spatial-temporal interactions. By constructing a large-scale annotated dataset through a hybrid semi-automated labeling pipeline, we enable the training of a fine-grained

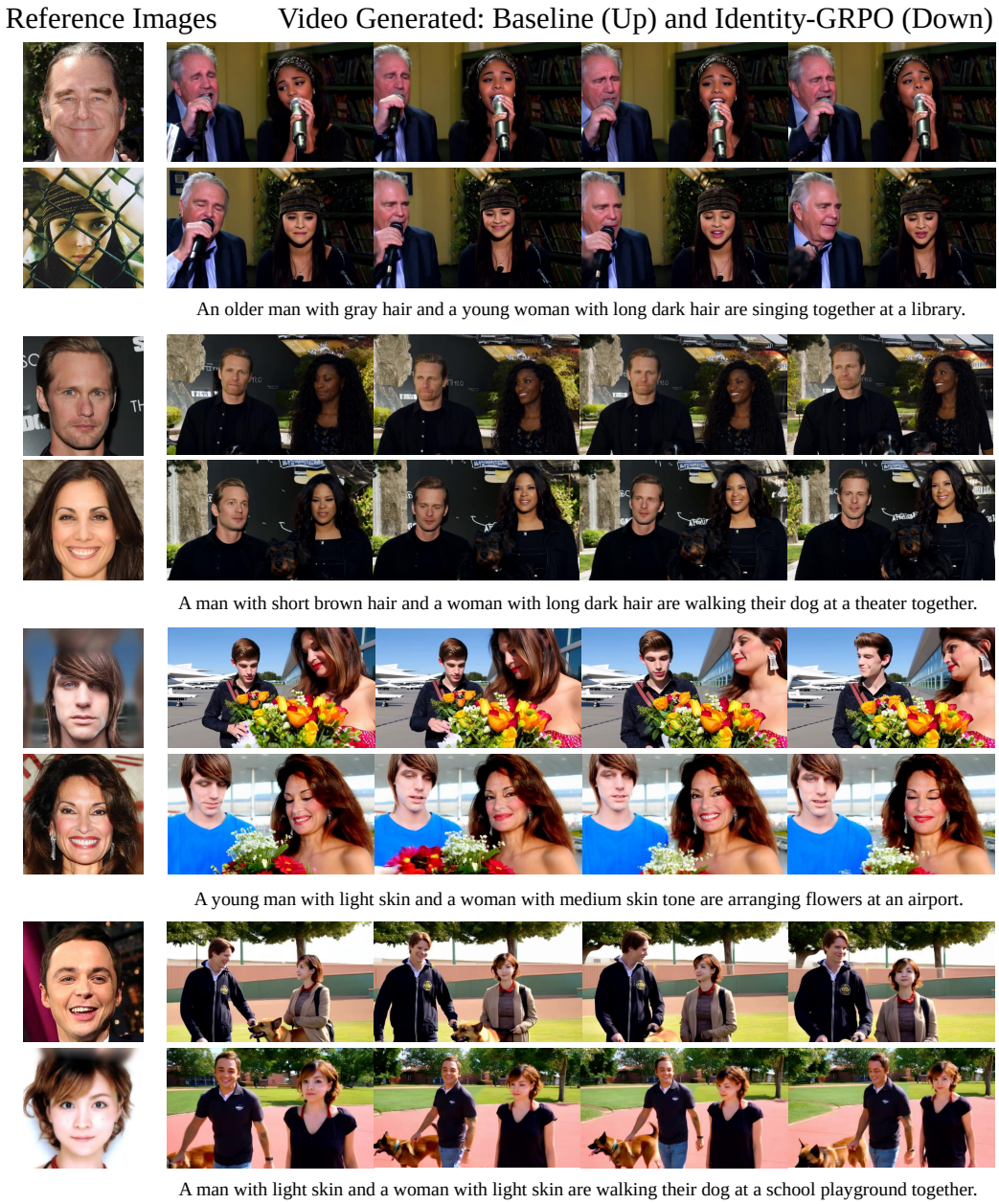


Figure 2: Visualization results for qualitative analysis. The first two groups show a comparison between VACE-1.3B and VACE-1.3B+Identity-GRPO, while the last two groups compare Phantom-1.3B with Phantom-1.3B+Identity-GRPO. In each group, the first row presents the results from the baseline model, and the second row shows the results generated by Identity-GRPO.

identity-consistency reward model tailored to disentangle identity preservation from dynamic motion requirements. Our systematic evaluation of GRPO training configurations identifies optimal hyper-parameters for MH-IPV scenarios, demonstrating that Identity-GRPO outperforms state-of-the-art baselines (VACE and Phantom) by 18.9% and 6.5% in identity-consistency metrics. These contributions establish a robust foundation for aligning reinforcement learning with human-centric video generation, offering actionable insights for advancing scalable, identity-preserving multi-human content creation.

## REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6099–6110, 2025.

christophschuhmann. improved-aesthetic-predictor. *improved-aesthetic-predictor Lab*, 2024. URL `https://github.com/christophschuhmann/improved-aesthetic-predictor/tree/main`.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. *ArXiv*, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

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

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023*. Neural Information Processing Systems Foundation, 2023.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shashank Gupta, Chaitanya Ahuja, Tsung-Yu Lin, Sreya Dutta Roy, Harrie Oosterhuis, Maarten de Rijke, and Satya Narayan Shukla. A simple and effective reinforcement learning method for text-to-image diffusion fine-tuning. *arXiv preprint arXiv:2503.00897*, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, October 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

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

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7752–7762, 2025a.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025b.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2(5):7, 2024.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. Criticbench: Benchmarking llms for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*, 2024.

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

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

Jinsong Liu, Dongdong Ge, and Ruihao Zhu. Reward learning from preference with ties. *arXiv preprint arXiv:2410.05328*, 2024a.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, October 2025c.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8009–8019, 2025d.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4122–4134, 2025e.

Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10844–10853, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL `https://arxiv.org/abs/2212.09748`.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *ArXiv*, 2023.

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

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

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

Pejaver V Rao and Lawrence L Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024a.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 2024b.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024.

Xiaomeng Yang, Zhiyu Tan, and Hao Li. Ipo: Iterative preference optimization for text-to-video generation. *arXiv preprint arXiv:2502.02088*, 2025b.

Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *Advances in Neural Information Processing Systems*, 37: 73366–73398, 2024.

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

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12978–12988, 2025b.

Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *arXiv preprint arXiv:2412.15159*, 2024a.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024b.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2063–2073, 2025a.

Zhenghao Zhang, Junchao Liao, Xiangyu Meng, Long Qin, and Weizhi Wang. Tora2: Motion and appearance customized diffusion transformer for multi-entity video generation. *arXiv preprint arXiv:2507.05963*, 2025b.

Hanyang Zhao, Haoxian Chen, Ji Zhang, David D Yao, and Wenpin Tang. Score as action: Fine-tuning diffusion generative models by continuous-time reinforcement learning. *arXiv preprint arXiv:2502.01819*, 2025.

Bingwen Zhu, Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Yidi Wu, Huyang Sun, and Zuxuan Wu. Aligning anime video generation with human feedback. *arXiv preprint arXiv:2504.10044*, 2025.

# A Related Work

## A.1 Multi-human Identity-preserving Video Generation

With the advancement of video foundational models Wan et al. (2025); Kong et al. (2024), fine-grained identity-preserving video generation becomes possible. Recently, SkyReels-A2 Fei et al. (2025) proposed an image-text joint embedding model to inject multi-element representations into existing video foundation models, to effectively learn cross-modal data form integration. Tora2 Zhang et al. (2025b) introduced a decoupled personalization extractor that generated comprehensive personalization embeddings for multiple open-set entities. HunyuanCustom Hu et al. (2025) designed a text-identity image fusion module based on LLM for enhanced multi-modal understanding. VACE Jiang et al. (2025) utilized a context adapter structure to inject different video generation task inputs into the model, allowing it to handle arbitrary video synthesis tasks, such as reference-to-video generation, multi-human identity-preserving video generation. Phantom Liu et al. (2025c) introduced a dynamic information injection scheme, allowing the insertion of one or more reference human, achieving a unified model architecture for single and multi-human identity-preserving video generation.

## A.2 Vision-language Reward Models

Reward models are crucial in aligning video generation models with human preferences. Recently, with the advancement of large vision-language models (VLM) Bai et al. (2025); Wang et al. (2025a), many methods use VLM to simulate human perception by performing visual quality score regression or preference learning. VideoScore He et al. (2024) trained video quality assessment models on human-annotated video labels. VisionReward Xu et al. (2024a) performs multidimensional evaluation through 64 fine-grained binary visual QA questions, producing human-aligned visual preference scores. VideoAlign Liu et al. (2025b) collected large-scale human-annotated video preference datasets and tuned VLM as a multidimensional reward model to evaluate three critical aspects for text-to-video task: visual quality, motion quality, and text-video alignment. AnimeReward Zhu et al. (2025) constructed a large-scale anime video dataset that incorporated human preferences for both visual appearance and visual consistency. Beyond these methods, some works such as LiFT Wang et al. (2024), IPO Yang et al. (2025b) and UnifiedReward Wang et al. (2025b), following VLM-as-a-judge methods Lin et al. (2024), leverage the intrinsic reasoning capabilities of VLMs, where VLMs are tuned to generate cirtic reasons followed by predicted preference labels. Despite these promising advances, these VLM-as-a-judge methods are not accurately predict human preferences in many cases. In order to overcome the problems existing in the current methods, we construct the first identity consistency reward model for MH-IPV task.

## A.3 Reinforcement Learning for Image and Video Generation

To apply Reinforcement Learning from Human Feedback (RLHF) to image and video generation, early methods either directly fine-tuned models using scalar reward signals Prabhudesai et al. (2023); Clark et al. (2023); Xu et al. (2024b); Prabhudesai et al. (2024) or employed Reward Weighted Regression (RWR) Peng et al. (2019); Lee et al. (2023); Furuta et al. (2024). Inspired by Proximal Policy Optimization (PPO) Schulman et al. (2017), policy gradient methods were later integrated into diffusion models and demonstrated effectiveness Black et al. (2023); Fan et al. (2023); Gupta et al. (2025); Miao et al. (2024); Zhao et al. (2025). However, PPO-based approaches suffer from

high computational costs and sensitivity to hyperparameters. To improve training efficiency, Direct Preference Optimization (DPO)-based methods Rafailov et al. (2023); Wallace et al. (2024); Dong et al. (2023); Yang et al. (2024); Liang et al. (2024); Yuan et al. (2024); Liu et al. (2025d); Zhang et al. (2024a) directly utilize human preference data as learning signals through a supervised loss objective. Recently, GRPO has shown superior performance in complex reasoning tasks. In the domain of image and video generation, Flow-GRPO Liu et al. (2025a) and Dance-GRPO Xue et al. (2025) have successfully introduced GRPO into flow matching models, enabling diverse sampling by reformulating the ODE as an equivalent SDE. To improve the efficiency of the optimization process, Mix-GRPO Li et al. (2025b) introduces a sliding window mechanism, applying SDE sampling and GRPO-guided optimization only within the window, while using ODE sampling outside of it.