MagicID: Hybrid Preference Optimization for ID-Consistent and Dynamic-Preserved Video Customization

Supplementary Material

A. Implement Details

To construct the preference video repository, we generate 20 prompts using LLM model. Then we sample 100 and 20 videos with fine-tuned T2V model and the initial T2V model respectively, which are incorporated with the static videos inflated from the reference images. After the pair selection, we sample the top 100 pairs as our preference data. In the training stage, we employ the AdamW optimizer configured with a learning rate of 2e-5 and a weight decay parameter of 1e-4. We first fine-tune the model for 1000 steps in the initial training stage. Then we use our training method to optimize for 5000 steps. During inference, we use 50 steps of DDIM sampler and classifier-free guidance with a scale of 7.5. We generate 61-frame videos with 720 × 1280 spatial resolution. All experiments are conducted on a single NVIDIA H00 GPU.

B. Baseline Details

We compare our method with both optimization methods, such as Magic-Me and Dreambooth with LoRA, and encoder-based methods such as IDAnimator and ConsisID. Specifically, Magic-Me is a recent T2V customization method that trains extended keywords and injects it into HunyuanVideo. Besides, we compare with Dreambooth-LoRA, which uses traditional reconstructive loss during training. For a fair comparison, we train them for the same total steps with our method.

C. DPO Objective Function

This section provides a systematic derivation of the Direct Preference Optimization (DPO) formula, detailing the derivation of the DPO objective function and constructing a preference-driven optimization framework based on the reward function and KL divergence. The goal of DPO is to maximize rewards while aligning the policy with a baseline model. The objective function is defined as:

$$\max_{\pi} \mathbb{E}_{x\in X, y\in\pi}\left[r(x,y)\right] - \beta \cdot \mathbb{D}_{\text{KL}}\left[\pi(y|x)||\pi_{\text{ref}}(y|x)\right], \tag{10}$$

where D_{KL} denotes the Kullback-Leibler divergence between the learned policy π and a reference policy π_{ref}, enforcing consistency with the baseline model. To simplify, this objective is reformulated as a minimization problem:

$$\min_{\pi} \mathbb{E}_{x\in X, y\in\pi}\left[\log\frac{\pi(y|x)}{\pi^*(y|x)} - \log Z(x)\right], \tag{11}$$

where $Z(x)$ is defined as:

$$Z(x) = \sum_{y} \pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right). \tag{12}$$

This reformulation leads to the final optimization objective:

$$\min_{\pi} \mathbb{E}_{x\sim D}\left[\mathbb{D}_{\text{KL}}(\pi(y|x)||\pi^*(y|x))\right]. \tag{13}$$

Under the minimization of KL divergence, the policy $\pi(y|x)$ adheres to the following form:

$$\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\cdot\exp\left(\frac{1}{\beta}r(x,y)\right). \tag{14}$$

Reversing this equation yields the reward function:

$$r^*(x,y) = \beta\log\frac{\pi(y\mid x)}{\pi_{\text{ref}}(y\mid x)}. \tag{15}$$

Incorporating the Bradley-Terry model, the cross-entropy loss function $\mathcal{L}$ is defined, which quantifies the difference between the preferred and non-preferred responses. This loss function is essential for deriving the gradient necessary to optimize the DPO objective:

$$\mathcal{L} = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\ln\sigma\left(\beta\log\frac{\pi(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \tag{16}$$

where σ denotes the sigmoid function, which maps the difference in log-probabilities to a range of $[0, 1]$. Differentiating $\mathcal{L}$ provides the gradient needed to optimize the DPO objective with respect to the preference data.

D. Applying DPO Strategy into our MagicID

In adapting Dynamic Preference Optimization (DPO) to our MagicID task, we consider a pairwise preference video data $P = \{(c, v_0^w, v_0^l)\}$. In this dataset, each example contains their text prompts c and a pair of videos (v_0^w, v_0^l) generated by a reference model p_{ref}, where $v_0^w \succ v_0^l$ indicates that humans prefer v_0^w over v_0^l. The goal of DPO is to train a new model p_θ so that its generated videos align with human preferences rather than merely imitating the reference model. However, directly computing the distribution $p_\theta(v_0|c)$ is highly complex, as it requires marginalizing over all possible generation paths $(v_1, \ldots, v_T)$ to produce v_0, which is practically infeasible.

To address this challenge, researchers leverage Evidence Lower Bound (ELBO) by introducing latent variables $v_{1:T}$. The reward function $R(c, v_{0:T})$ is defined to measure the quality of the entire generation path, allowing the expected reward $r(c, v_0)$ for given c and v_0 to be formulated as:

$$r(c, v_0) = \mathbb{E}_{p_\theta(v_{1:T}|v_0,c)}[R(c, v_{0:T})] \tag{17}$$

In DPO, a KL regularization term is also included to constrain the generated distribution relative to the reference distribution. Here, an upper bound on the KL divergence is used, converting it to a joint KL divergence:

$$\mathbb{D}_{KL}[p_\theta(v_{0:T}|c) \parallel p_{\text{ref}}(v_{0:T}|c)] \tag{18}$$

This upper bound ensures that the distribution of the generated model $p_\theta(v_{0:T}|c)$ remains consistent with the reference model $p_{\text{ref}}(v_{0:T}|c)$, preserving the model's generation capabilities while optimizing human preference alignment. Plugging in this KL divergence upper bound and the reward function $r(c, v_0)$ into the objective function, we obtain:

$$\max_{p_\theta} \mathbb{E}_{c\sim\mathcal{D}_c, v_{0:T}\sim p_\theta(v_{0:T}|c)}[r(c, v_0)] - \beta\mathbb{D}_{KL}[p_\theta(v_{0:T}|c) \parallel p_{\text{ref}}(v_{0:T}|c)] \tag{19}$$

The definition of this objective function is optimized over the path $v_{0:T}$. Its primary goal is to maximize the reward for the reverse process $p_\theta(v_{0:T})$ while maintaining distributional consistency with the original reference reverse process. To optimize this objective, the conditional distribution $p_\theta(v_{0:T})$ is directly used. The final DPO-MagicID loss function $L_{\text{HPO}}(\theta)$ is expressed as follows:

$$\mathcal{L}_{\text{HPO}}(\theta) = -\mathbb{E}_{(v_0^w, v_0^l)\sim P} \log\sigma\left(\beta\mathbb{E}_{v_{1:T}^w\sim p_\theta(v_{1:T}|v_0^w),\, v_{1:T}^l\sim p_\theta(v_{1:T}|v_0^l)}\left[\log\frac{p_\theta(v_{0:T}^w)}{p_{\text{ref}}(v_{0:T}^w)} - \log\frac{p_\theta(v_{0:T}^l)}{p_{\text{ref}}(v_{0:T}^l)}\right]\right) \tag{20}$$

By applying Jensen's inequality, the expectation can be moved outside of the $\log\sigma$ function, resulting in an upper bound. This simplifies the formula and facilitates optimization. After applying Jensen's inequality, the upper bound of the loss function is given by:

$$\mathcal{L}_{\text{HPO}}(\theta) \leq -\mathbb{E}_{(v_0^w, v_0^l)\sim P}\, \mathbb{E}_{v_{1:T}^w\sim p_\theta(v_{1:T}|v_0^w),\, v_{1:T}^l\sim p_\theta(v_{1:T}|v_0^l)}\, \log\sigma\left(\beta\left[\log\frac{p_\theta(v_{0:T}^w)}{p_{\text{ref}}(v_{0:T}^w)} - \log\frac{p_\theta(v_{0:T}^l)}{p_{\text{ref}}(v_{0:T}^l)}\right]\right) \tag{21}$$

To handle the complexity of calculating high-dimensional video sequence probabilities with a total of $T = 1000$ time steps, we employ an approximation approach. We introduce an approximate posterior $q(v_{1:T}|v_0)$ for the subsequent time steps and utilize the Evidence Lower Bound (ELBO) to approximate $\log p_\theta(v_{0:T})$. Then, by expressing $p_\theta(v_{0:T})$ and $q(v_{1:T}|v_0)$ as products of conditional probabilities at each time step, we achieve a stepwise sampling approach. The final approximate expression is:

$$\begin{aligned}\log p_\theta(v_{0:T}) \approx \mathbb{E}_{q(v_t|v_{t-1}),t\sim\{1..T\}}\Bigg[&\log\frac{p_\theta(v_0)}{q(v_0)} \\ &+ \log\frac{p_\theta(v_t|v_{t-1})}{q(v_t|v_{t-1})}\Bigg].\end{aligned} \tag{22}$$

Since $q(v_t|v_{t-1})$ is a conditional probability distribution that generally sums to 1, the KL divergence can be expressed as:

$$\mathbb{D}_{KL}\left(q(v_t|v_{t-1}) \,\|\, p_\theta(v_t|v_{t-1})\right) = \log \frac{q(v_t|v_{t-1})}{p_\theta(v_t|v_{t-1})}. \tag{23}$$

Based on Eqs. (22) and (23), we rewrite $\log p_\theta(v_{0:T})$ as:

$$\begin{aligned} \log p_\theta(v_{0:T}) \approx \mathbb{E}_{q(v_{1:T}|v_0),t\sim\{1..T\}} \left[\log \frac{p_\theta(v_0)}{q(v_0)}\right] \\ -\mathbb{D}_{KL}\left(q(v_t|v_{t-1}) \,\|\, p_\theta(v_t|v_{t-1})\right). \end{aligned} \tag{24}$$

Moreover, the derivation of $\log p_{\text{ref}}(v_{0:T})$ is consistent with that of $\log p_\theta(v_{0:T})$. Based o Eq. (24), we can rewrite $\Delta(v_{0:T})$ as:

$$\log \frac{p_\theta(v_{0:T})}{p_{\text{ref}}(v_{0:T})} = -\mathbb{D}^{\theta}_{KL} + \mathbb{D}^{\text{ref}}_{KL} + C. \tag{25}$$

By rewriting the KL divergence in terms of noise prediction, we can express it as follows:

$$\mathbb{D}^{\theta}_{KL} \propto ||\epsilon - \epsilon_\theta(v_t, t)||^2 \quad \mathbb{D}^{\text{ref}}_{KL} \propto ||\epsilon - \epsilon_{\text{ref}}(v_t, t)||^2. \tag{26}$$

Finally, based on Eqs. (8), (21) and (25), the complete form of the DPO loss function for our MagicID is:

$$\begin{aligned} \mathcal{L}_{\text{HPO}}(\theta) = \mathbb{E}_{(v_0^w, v_0^l)\sim\mathcal{D},\, t\sim\{1..T\}} \Bigg[\beta \log \sigma \Bigg(\\ \left(||\epsilon_w - \epsilon_\theta(v_t^w, t)||^2 - ||\epsilon_w - \epsilon_{\text{ref}}(v_t^w, t)||^2\right) \\ - \left(||\epsilon_l - \epsilon_\theta(v_t^l, t)||^2 - ||\epsilon_l - \epsilon_{\text{ref}}(v_t^l, t)||^2\right) \Bigg)\Bigg]. \end{aligned} \tag{27}$$

E. More Results

As shown in Fig. 9, Fig. 10, Fig. 11, Fig. 12, and Fig. 13, we present more customization results of MagicID. They showcase it achieves consistent identity and preserves natural motion dynamics, which provides further evidence of its promising performance.

F. Reproducibility Statement

We make the following efforts to ensure the reproducibility of MagicID: (1) Our training and inference codes together with the trained model weights will be publicly available. (2) We provide training details in the appendix, which is easy to follow. (3) We provide the details of the human evaluation setups.

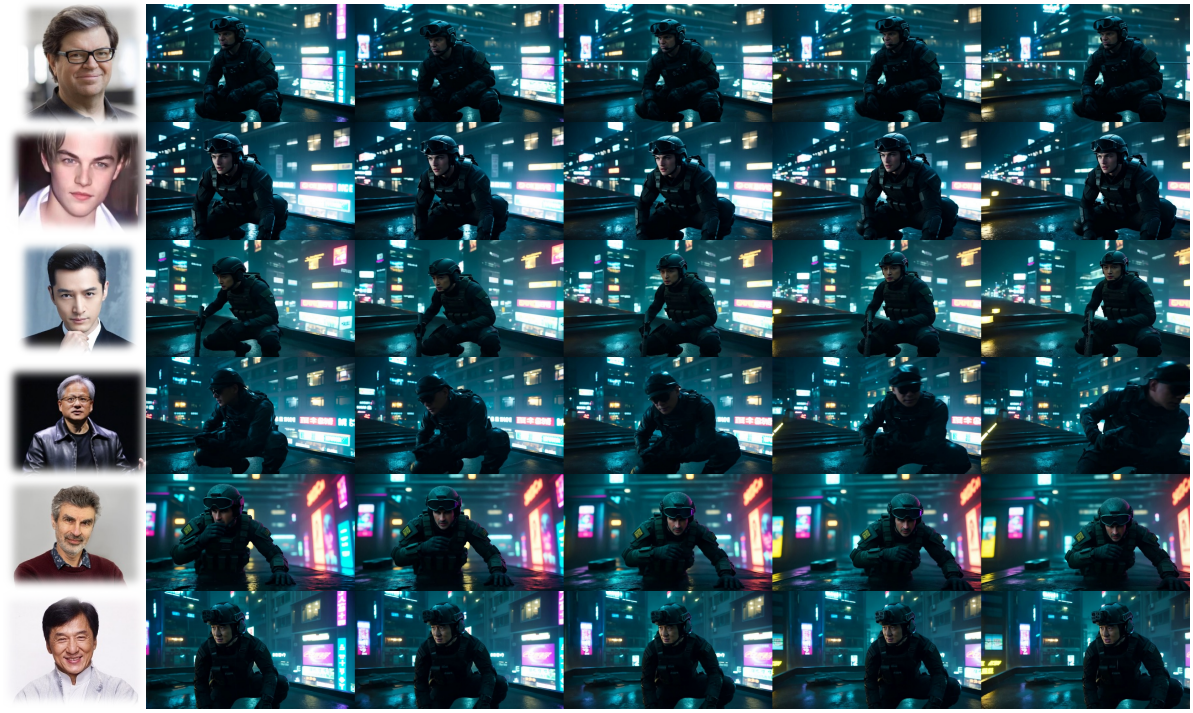
G. Impact Statement

Our main objective in this work is to empower novice users to generate visual content creatively and flexibly. However, we acknowledge the potential for misuse in creating fake or harmful content with our method. Thus, we believe it's essential to develop and implement tools to detect biases and malicious use cases to promote safe and equitable usage.

Reference

"A firefighter man charges through billowing smoke, his helmet visor smudged with ash. His face glows orange from the flames as he grips a hose tightly, water arcing toward a collapsing building. Sweat streaks down his determined expression."

Figure 9. More results of MagicID.



Reference

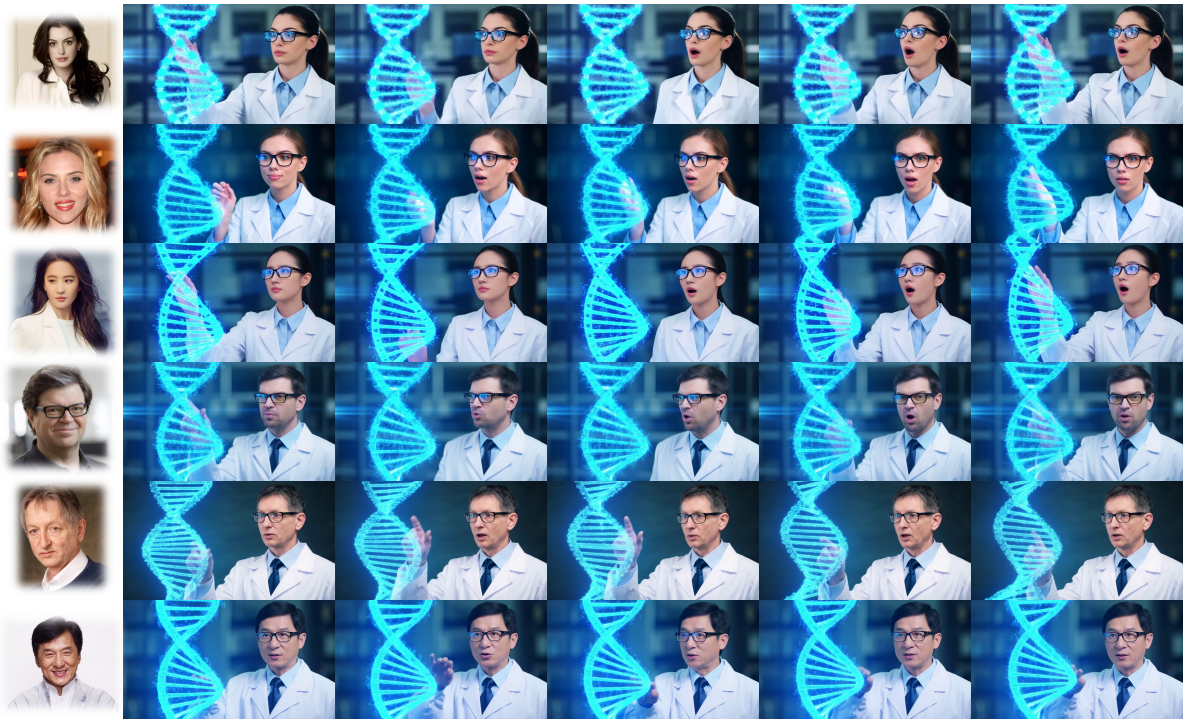
"A man in a tactical stealth suit with a helmet crouches on a rain-soaked rooftop. His jaw is clenched in concentration as he surveys the neon-lit alleyways below, reflections of holographic advertisements flickering across his visor."

Figure 10. More results of MagicID.

Reference

A tribal warrior man pauses at the edge of a jungle waterfall, his face painted with crimson symbols. He grips a spear tightly, muscles tensed as his gaze locks onto movement in the mist-shrouded valley below.

Figure 11. More results of MagicID.



Reference

"A scientist man in a lab coat stands before a glowing holographic DNA strand, his glasses reflecting blue light. His eyebrows shoot up in astonishment, mouth slightly open as he reaches out to touch the shimmering data stream."

Figure 12. More results of MagicID.

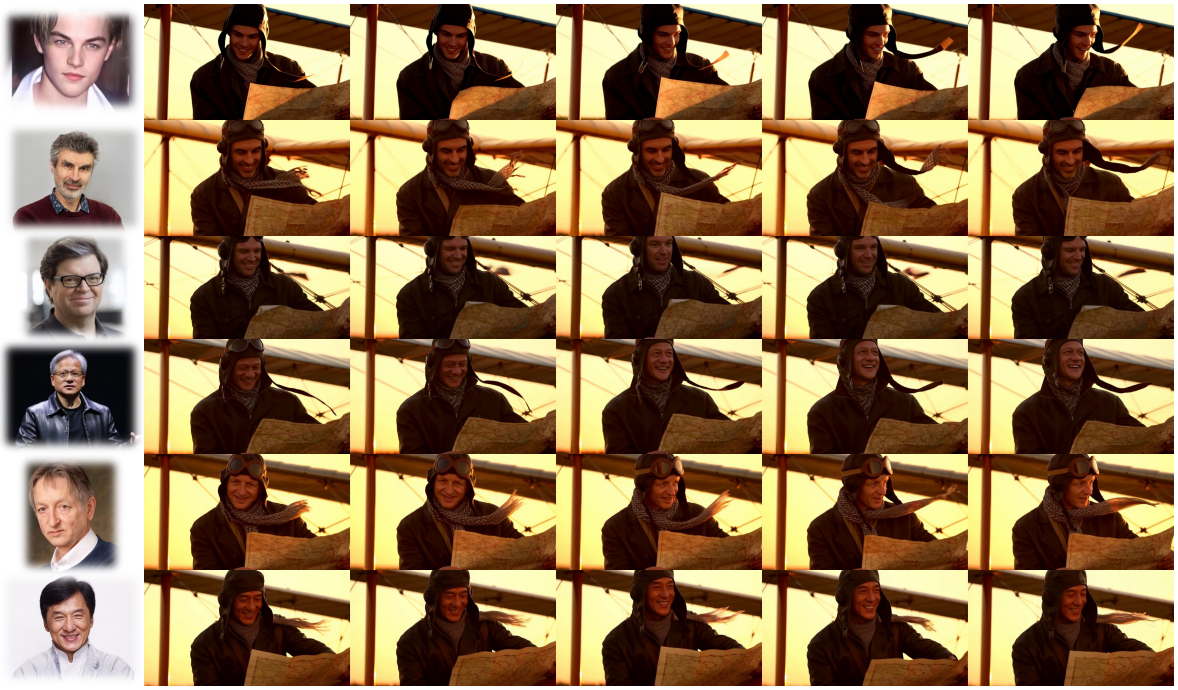


Figure 13. More results of MagicID.