Appendix

Table of Contents

A. Details of Dataset Construction

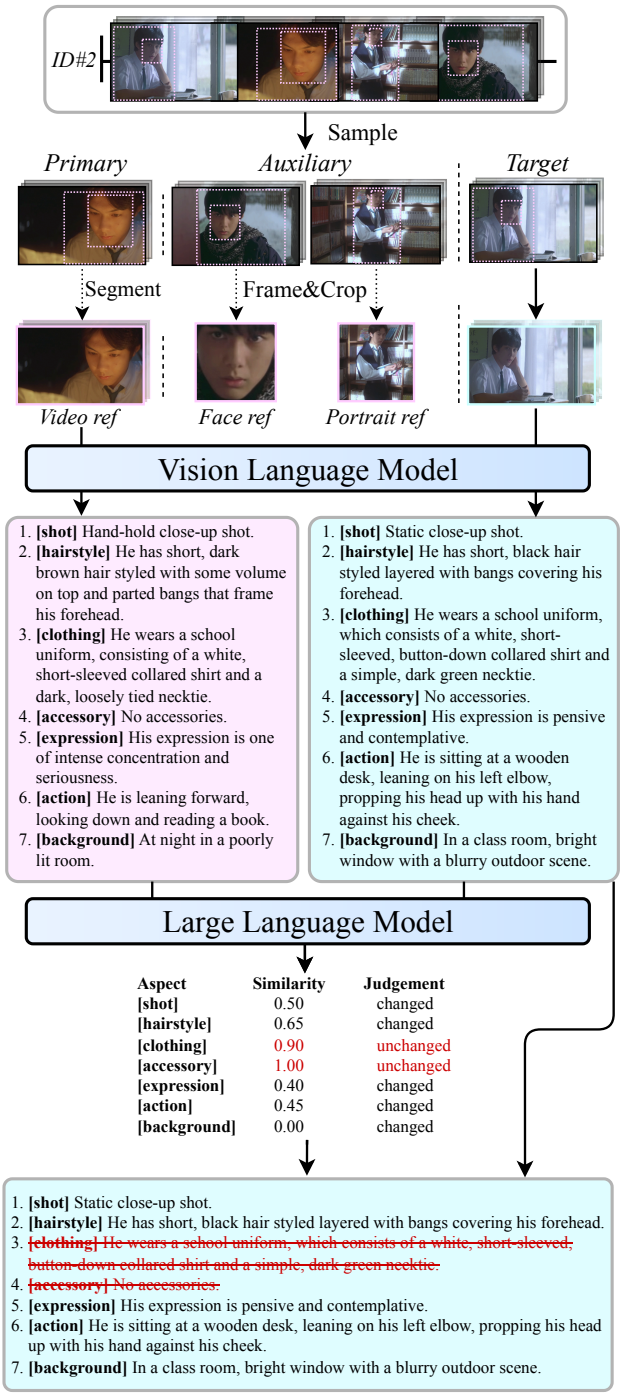

Figure 6. A construction example of the visual references and the differential prompt for the supervised training.

In our paper, we introduce a data pipeline for constructing diverse visual references and differential prompts to facilitate supervised training. Here, we provide a detailed description with a construction example in Fig. 6. As illustrated, the whole pipeline starts from sampling one video as the primary reference, several videos as auxiliary references, and one video as the target. All the reference videos randomly undergo operations to produce different forms of visual references. Temporal segmentation produces video references. Sampling one frame and cropping the facial region produces face image references. Sampling one frame and cropping the region around the person produces portrait image references. Afterwards, the primary reference and the target video are respectively input to a VLM for detailed descriptions in the seven portrait-video aspects. The two descriptions are then input to an LLM to quantify the similarity between them in terms of the seven aspects. Aspects with a similarity score exceeding a threshold γ, which is set to 0.8 in practice, are deemed unchanged. The unchanged aspects are removed from the description of the target video to produce the final differential prompt.

B. More Experimental Details

We utilize the text input mode of Gemini-2.5-Pro as the LLM and the text-video input mode as the VLM. All the training is carried out on NVIDIA A100 80GB GPUs. The supervised training starts from standard Wan-5B using LoRA, which takes 4,000 GPU hours. The RL tuning starts from the merged weights after supervised training, also using LoRA, which takes additional 500 GPU hours. Throughout the training, all the important settings are listed below:

Table 2. Experimental settings of AnyID.

Parameter	Value
Video height	704
Video width	1280
Video frame	121
FPS	24
LoRA rank	128
LoRA alpha	128
Batchsize	64
Train timesteps	1000
Train shift	5.0
Optimizer	AdamW
Learning rate	1e-4
Weight decay	0.001
Sample timesteps	50
Sample shift	5.0
Sample guidance scale	5.0

C. Details of Benchmark

Celebrities. To construct a comprehensive and objective evaluation set, we must initially choose target individuals with consistent appearances, abundant and easily accessible visual records, and preferably well-known to the public, for better intuitive evaluation of the generated fidelity. Hence, we utilized LLM to generate a list of 50 celebrities with a balanced distribution of ethnicity and gender, spanning diverse fields such as AI academia, entrepreneurship, athletics, acting, and music. The list of celebrities is as follows:

- 001 Sam Altman
- 002 Jensen Huang
- 003 Elon Musk
- 004 Jun Lei
- 005 Yun Ma
- 006 Shou Zi Chew
- 007 Feifei Li
- 008 Yann LeCun
- 009 Stephen Curry
- 010 Ming Yao
- 011 Kobe Bryant
- 012 Luka Doncic
- 013 LeBron James
- 014 Lionel Messi
- 015 Cristiano Ronaldo
- 016 Long Ma
- 017 Hanyu Yuzuru
- 018 Eileen Gu
- 019 Ziyi Zhang
- 020 Yuyan Peng
- 021 Andy Lau
- 022 Yi Zhang
- 023 Liying Zhao
- 024 Jackie Chan
- 025 Crystal Liu
- 026 Benedict Cumberbatch
- 027 Morgan Freeman
- 028 Leonardo DiCaprio
- 029 Robert Downey Jr.
- 030 Will Smith
- 031 Emma Watson
- 032 Andrew Garfield
- 033 Rowan Atkinson
- 034 Anne Hathaway
- 035 Scarlett Johansson
- 036 Keira Knightley
- 037 Halle Berry
- 038 Satomi Ishihara
- 039 Hiroshi Abe
- 040 Jay Chou
- 041 Rihanna
- 042 Taylor Swift
- 043 Eason Chan
- 044 Billie Eilish
- 045 JJ Lin
- 046 Mika Nakashima
- 047 Lana Del Rey
- 048 Ed Sheeran
- 049 Bruno Mars
- 050 Beyoncé

Visual References. For each celebrity, we elaborately collect four portrait images and one video from the website, ensuring that they are free of visual filters, with clear and unobstructed faces, and captured under bright and sufficient lighting conditions. The four portraits are chosen to exhibit significant angular differences and expression variance, offering a range of prior perspectives. The video length is approximately 10 seconds, without shot cuts, typically sourced from interview footage due to its clarity of facial features and stable camera set.

Prompts. To comprehensively evaluate the generative capabilities across diverse scenarios, we devise a systematic process of prompt construction for the IPT2V and IEPT2V tasks. We start with the construction for the IPT2V task, where all seven portrait-video elements are expected to change. We instruct LLM to generate 50 prompts in the predefined format comprising seven elements, with 25 for males and 25 for females. In the instructions, we emphasized the necessity for a wide variety of scenes and styles, such as indoor settings, landscapes, science fiction, fantasy, elegance, and grandeur. We also stress the need for an appropriate degree of change in characters' expressions and movements. Subsequently, we randomly remove descriptions of hairstyle, clothing, or accessory from the prompts to generate 50 differential prompts for the IEPT2V task. These prompts can be randomly combined with sets of visual references to generate up to 750 evaluation cases for each task. In the evaluation of our paper, we sample 50 sets of IPT2V cases and 50 sets of IEPT2V cases to compute metrics and draw comparisons, resulting in Tab. 1 and Fig. 3. Note that when for the baselines, we transform the prompts into a format suitable for their text encoders by simplifying or paraphrasing them. Here, to briefly showcase the format of the generated prompts, three examples are provided below:

- Handheld shot, medium shot. The person has short, messy ash-blonde hair with bangs casually falling over his forehead. He is wearing a navy-and-white striped sailor shirt. His expression is filled with curiosity and excitement, eyes wide open. Slowly, he tilts his head back, following something in the sky with his gaze, until his face is almost pointing upward toward the heavens. Then, he gradually lowers his head and flashes an open, cheerful smile at the

camera. The character's action takes place by the railing on a seaside promenade, with the sea breeze gently blowing through his hair. One hand shields his eyes from the sun while the other points to the sky. The video setting is a sunny coastal boulevard, with a backdrop of the deep blue ocean and white seagulls. The lighting is strong midday sunlight, creating a fresh, free-spirited, and energetic summer atmosphere.

- Static shot, medium shot. The person has long, straight black hair tied neatly into a ponytail, adorned with an intricate golden crown. He wears a lavish red silk robe embroidered with golden dragon patterns. His expression is solemn and majestic, with sharp, piercing eyes that exude authority without effort. Slowly, he turns his head from the left side toward the front, his gaze softening from its initial intensity into a compassionate gentleness. The character's posture shows him seated regally on a high-backed chair, hands folded across his chest. As he moves his head, his shoulders shift slightly, every motion deliberate and steady. The video setting is the terrace of an ancient palace, with red pillars in the background and distant rolling green mountains. The lighting comes from bright natural daylight, bathing his dignified face and highlighting the gold embroidery on his robes, evoking a grand, majestic, and historically rich ambiance.
- Push-in shot, transitioning from medium to close-up. The person has a sleek black high ponytail. She wears a form-fitting navy-blue pilot uniform paired with a flight helmet, the visor pushed up onto her head. Her expression initially reflects post-mission relaxation, with a hint of fatigue in her eyes. Upon hearing instructions through her communicator, her gaze sharpens instantly. Slowly turning her head, she lifts the corner of her mouth into a confident smile, ready to embrace the next challenge. The character's action depicts her seated inside the cockpit of a mech or combat aircraft. The video setting is a high-tech hangar, with other mechs undergoing maintenance in the background. The lighting consists of complex instrument panel glows within the cockpit, combined with the bright operational lights of the hangar, creating layered illumination full of tense anticipation for battle.

D. Analysis of Multiple References

The core contribution and motivation of AnyID is the introduction of multiple visual references. To further analyze the benefits brought by multiple references, we carry out a comparison in Fig. 7, from which we can draw the following conclusions: (1) Multiple references can provide prior information on static facial features of the target person, including facial markers such as freckles and moles. Taking the particular example of "Obama" in Fig. 7, for instance, he has a mole on one side of his nose. Only a single-angle reference lacks prior knowledge on this, making it completely impossible for the model to generate high-fidelity videos. This is the ill-posedness we have been emphasizing. By comparison, multiple references provide a holistic prior for the person, enabling the model to reconstruct the person with the highest possible fidelity. (2) Multiple references can provide prior information on dynamic features such as muscle movements and behavioral habits. As shown in the "Downey Jr." example in Fig. 7, when references in the same expression are provided, the dynamics of the generated character will deviate. If we provide an additional reference for the expected expression, the results will improve significantly.

In summary, multiple references can provide richer prior information, thereby significantly enhancing the fidelity of the model's output. Based on this, we also recommend providing references with a variety of expressions and angles when using our AnyID, rather than references with minimal variation.

E. Detailed Ablation Study

Primary Reference&Differential Prompt Design. Our AnyID features the design of a primary-reference generation and a differential prompt. The primary reference acts like a visual anchor, which provides element priors such as hairstyle and clothes. The differential prompt indicates the user-expected direction of modification, while the unmentioned aspects are expected to remain unchanged. Such design endows AnyID with the ability to follow the changes specified in the prompt and perfectly maintain the consistency of unmentioned elements. To prove this, we construct an ablating model of AnyID w/o primary reference&differential prompt design and compare it with the full model in Fig. 8. As shown, without such a design, the model's preservation of elements will be entirely controlled by the prompt, which is a weak constraint. As a result, the generations fail to preserve the hairstyle or the clothing from the reference. By comparison, the full model naturally anchors to the primary reference and succeeds in preserving the elements from it.

Reinforcement Learning. AnyID introduces a human-centric reinforcement learning (RL) strategy to achieve the best identity fidelity and prompt controllability. To evaluate its effect, we illustrate the generations of AnyID w/o reinforcement learning (i.e., the model after supervised training) and the full model in Fig. 9. As shown, the model without RL exhibits frequent flaws such as inconsistencies in details when preserving the elements or unsatisfactory

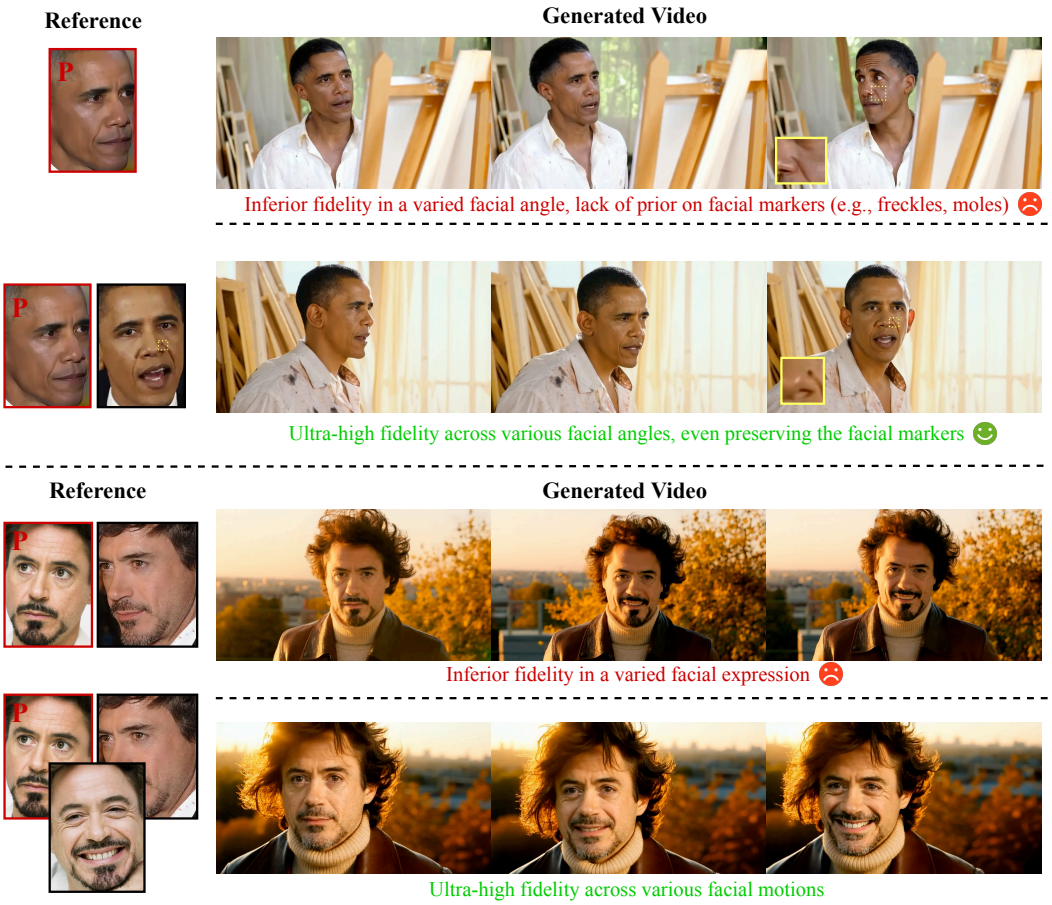

Figure 7. Illustration of ablating the primary reference&differential prompt design.

facial fidelity when preserving the identity. By comparison, benefiting from the RL tuning based on human feedback, the full model stably demonstrates perfect identity fidelity, element consistency, and prompt controllability.

F. Implementation of Preference Annotation

The preference annotation process in our paper consists of two tracks: identity fidelity and prompt controllability. In practice, we build a frontend interface with two modes for these two tracks. Fig. 10 shows the identity fidelity mode, where all the clips from the source ID-group are provided to ensure that annotators have comprehensive prior information about the character's identity. The prompt is not provided to avoid distractions. Annotators need to choose the preferred video with higher fidelity from the two generation options. By comparison, Fig. 11 shows the prompt controllability mode, where only the primary reference and the differential prompt are used. Annotators need to choose the preferred video that better changes according to the prompt while preserving consistency in aspects not specified by the prompt.

Figure 8. Illustration of ablating the primary reference&differential prompt design.

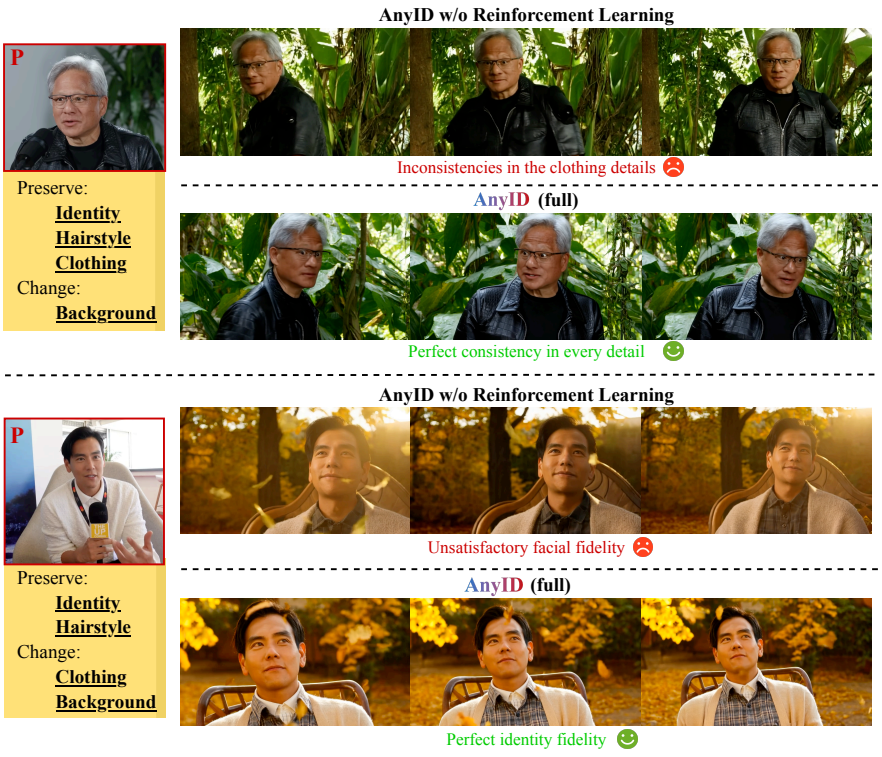

Figure 9. Illustration of ablating the reinforcement learning.

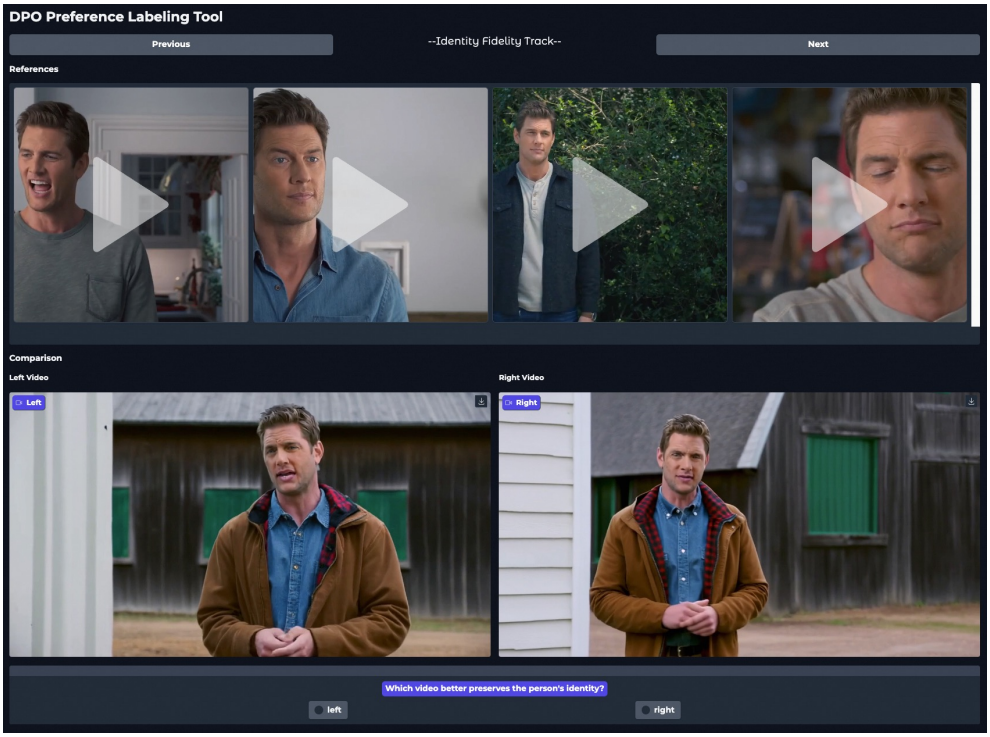

Figure 10. A preference annotation example of the identity fidelity track.

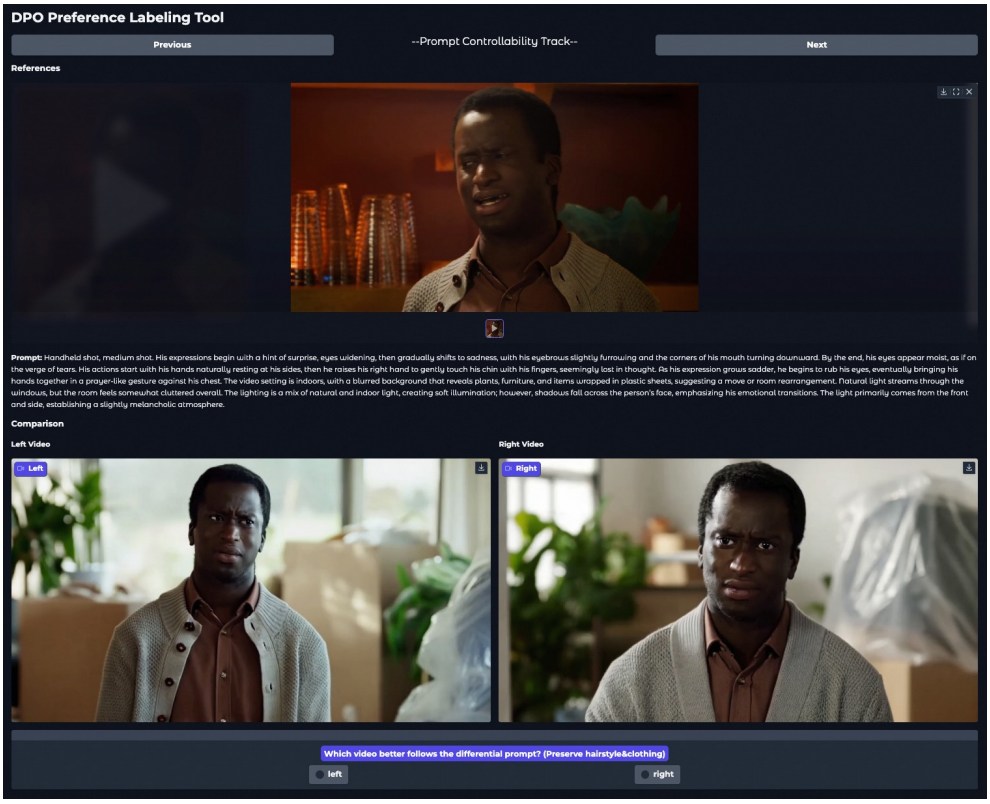

Figure 11. A preference annotation example of the prompt controllability track.

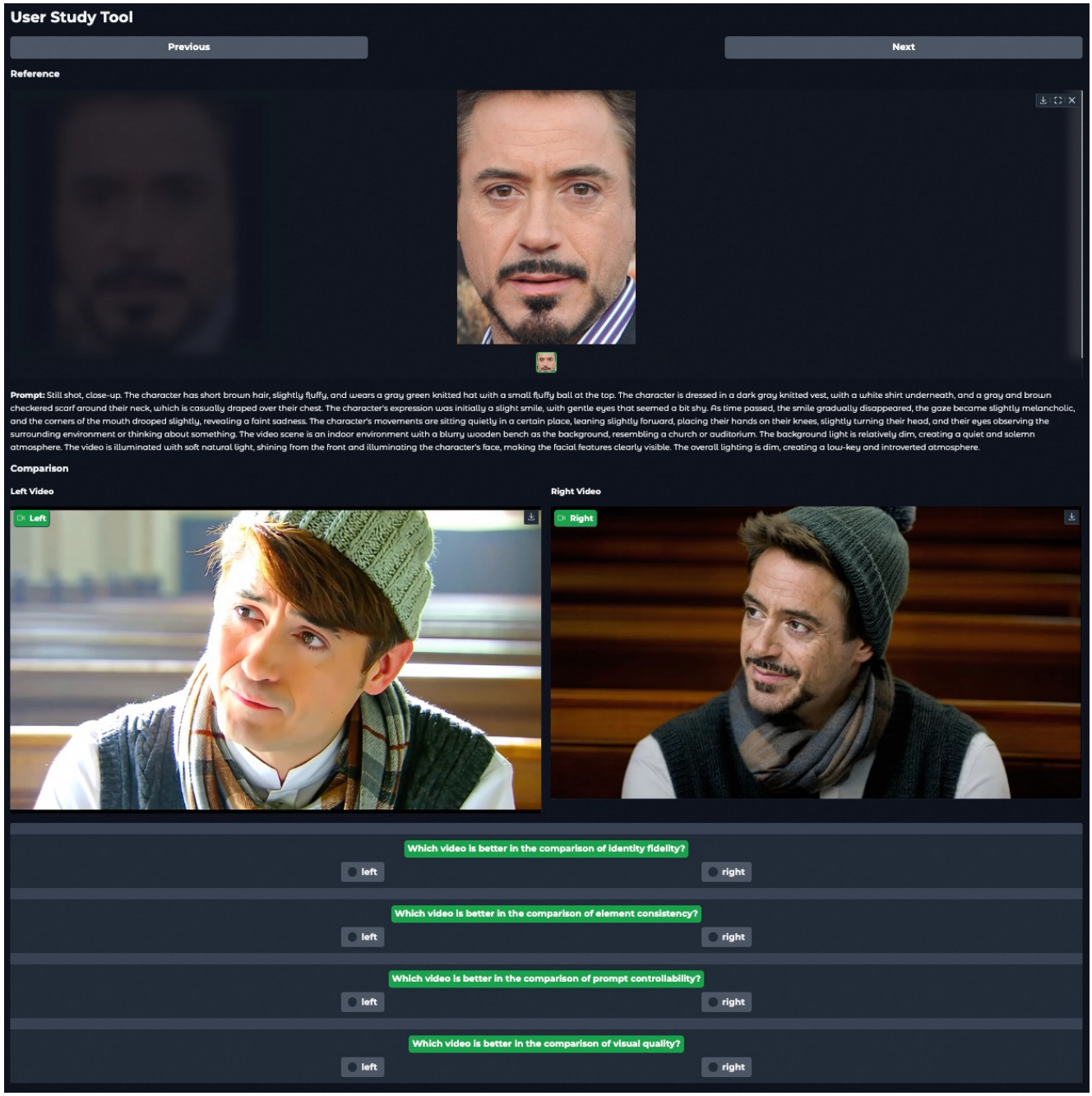

Figure 12. An example of the user study.

G. Implementation of User Study

To conduct the user study, we recruit volunteers and provide them with basic training based on the instructions below. We also build up a frontend interface shown in Fig. 12. The user study is based on the evaluation set of celebrities. In each matchup, only the primary reference of the celebrity and the prompt are provided since the faces of celebrities are already widely recognized. Participants need to select the winner in four dimensions, under the guidance of the instructions.

Instructions for Participants

Thank you for participating in our study!

We appreciate your willingness to contribute to our research. Below, you will find detailed instructions on how to complete the tasks involved in this study. Please read this information carefully before beginning.

1. Purpose of the Study

The goal of this study is to evaluate different methods for generating identity-preserving portrait videos given a visual reference of a celebrity and a prompt. Your feedback will help us evaluate which method produces videos with the best quality. Your participation is highly valuable to us!

2. Task Overview

You will be asked to compare two methods by evaluating their outputs in a series of one-on-one matchups. Each matchup will present results generated by our proposed method and a random baseline method (both anonymous). You will be given the visual reference along with the prompt. Note that the prompt only specifies the required changes, while aspects not mentioned in the prompt should remain consistent with the reference. Your task is to vote for your preferred option in each matchup based on the following criteria:

- **Identity Fidelity**: Which method better preserves the dynamic identity of the celebrity (e.g., facial features, expressions, and overall appearance) across all frames and shots in the video?
- **Element Consistency**: Which method better preserves the consistency of the unmentioned elements, such as hairstyle or clothing? (Note that this item is optional; some matchups will have it while others will not.)
- **Prompt Controllability**: Which method more accurately changes the aspects specified in the given prompt (e.g., appearance, expressions, actions, and background) in the generated video?
- **Visual Quality**: Which method produces a video with higher visual quality, considering factors like sharpness, smoothness of motion, and absence of artifacts (e.g., blurriness or unnatural textures)?

3. How to Complete the Task

- You will complete **40 matchups** in total.
- For each matchup, you will see two outputs side by side.
- To cast your vote, click on the button corresponding to your preferred option.
- There are no right or wrong answers. Please choose the option that feels best to you.

4. Time Commitment

The study should take approximately **20-30 minutes** to complete. You can work at your own pace, but we recommend completing the task in one sitting to ensure consistency.

5. Voluntary Participation

Your participation in this study is entirely **voluntary**. You are free to withdraw at any time without penalty or explanation. If you decide to withdraw, your responses will not be included in the analysis.

6. Privacy and Data Use

- Your responses will be anonymized and used solely for research purposes.
- No personally identifiable information will be collected or stored.

7. Questions or Concerns

If you have any questions about the study or encounter technical issues, please contact us. We are happy to assist you.

8. Consent

By proceeding with the study, you confirm that:

- You have read and understood the instructions.

- You agree to participate voluntarily.
- You understand that you can withdraw at any time without penalty.

Thank you again for your time and effort. Let's get started!

H. Societal Impacts and Safeguards

The introduction of AnyID represents a significant leap forward in identity-preserving video generation, with far-reaching societal implications. By enabling ultra-fidelity synthesis of personalized characters from diverse and heterogeneous identity references—ranging from static portraits to dynamic video clips—AnyID empowers creators, educators, filmmakers, and developers to produce highly consistent, controllable, and expressive visual content with unprecedented ease. This capability lowers barriers to professional-grade video creation, fostering inclusivity in digital storytelling, virtual communication, gaming, and personalized education. Moreover, the fine-grained attribute-level control offered by our primary-referenced generation paradigm opens new avenues for creative expression, adaptive avatars, and immersive user experiences in metaverse and AR/VR applications.

Nevertheless, such powerful generative capabilities entail substantial societal risks that must be proactively addressed. The potential for malicious use—such as generating convincing deepfakes, non-consensual impersonations, or deceptive media—poses serious threats to individual privacy, public trust, and information integrity. Additionally, the automation of high-quality video production may disrupt traditional creative workflows, potentially displacing roles in video editing, animation, and performance capture, thereby necessitating workforce reskilling and ethical transitions in creative industries. To mitigate these concerns, AnyID incorporates robust safeguards inherited from the Wan foundation model suite, including built-in content moderation filters and output watermarking for provenance tracking. Besides, in alignment with responsible AI principles, we commit to full transparency: all code, models, and datasets will be publicly released under licenses that prohibit misuse and require attribution. By balancing innovation with ethical stewardship, we aim to ensure that AnyID serves as a force for creative empowerment while safeguarding against societal harm.