Identity-Preserving Text-to-Video Generation by Frequency Decomposition

Supplementary Material

1. ConsisID Dataset

We propose a data pipeline to process and construct a high-quality ID-preservation video dataset as shown in Figure 2.

Data Curation Most existing identity-preserving datasets are image-centric [9, 31, 48, 59], with only a few focusing on videos [14, 85, 98]. However, these datasets primarily target facial regions, often cropping out relevant background content (*e.g.*, talking heads [69, 93]), which limits their broader applicability. To address this, we propose a pipeline to construct identity-preserving videos suitable for daily-life scenarios, as depicted in Figure 1. In line with the approach used by [89], we first constructed a set of search keywords (*e.g.*, "human," "woman," "man") and used them to retrieve videos from the internet. During this process, we excluded videos with *few views* or *likes* to ensure quality. Next, following [90], we apply PySceneDetect, Aesthetic, and Motion Score to filter out low-quality clips, ultimately getting a dataset of 130K high-quality clips.

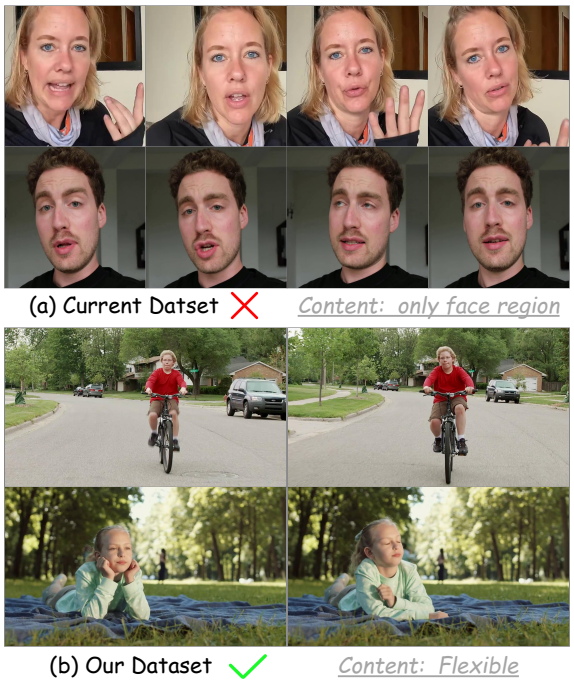


Figure 1. **Comparison between our in-house data and current Human Centric Video Dataset [19, 48, 85].** Our dataset is more flexible and diverse compared to previous ones.

Multi-view Face Filtering The purity of internet-sourced data is typically low, as full videos often include only brief segments featuring facial content. To address this, we first apply YOLO-Box [30] to extract frame-by-frame bounding box (bbox) information for the categories "face," "head," and "person". Video clips containing all three bounding boxes are retained. We then use YOLO-Pose [30] to detect facial keypoints (*e.g.* left eye, right eye, left ear, right ear, nose) and filter out video clips with low keypoint density to obtain clean ID-preservation video clips. To mitigate potential YOLO [30] errors, we set a tolerance threshold α. For instance, if β frames among frames 0 to 49 lack any of the three bounding boxes or valid keypoints, and $\beta < \alpha$, frames 0 to 49 are still retained as a complete video clip. To further enhance data quality, we discard video clips in which the "face" bbox occupies less than 6% of the frame.

Identical Verification A video may include multiple individuals, necessitating the assignment of a unique identifier to each person for subsequent training. Existing video tracking algorithms [1, 55, 92] lack robustness, often resulting in missed or incorrect detections. To address this, we utilize the previously obtained frame-by-frame bound-

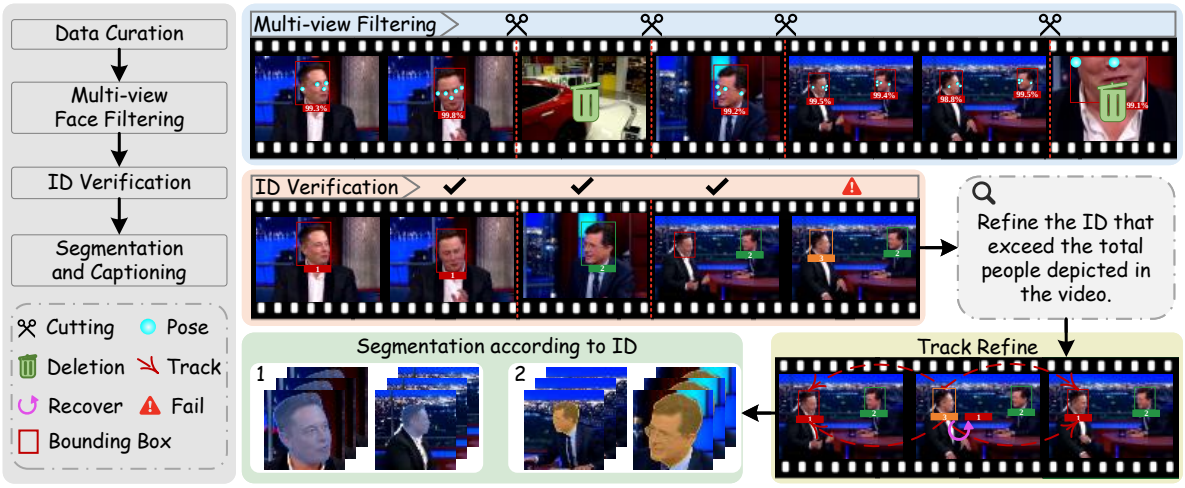


Figure 2. **Overview of the proposed Identity-Preserving Video Data Processing Pipeline.** First, we identify video clips with high facial density using bounding boxes (bbox) and key points. Next, we implement a tracking algorithm for ID verification, optimizing individual tracking IDs through the previous bounding box. Finally, we generate masks for each individual according to their unique IDs.

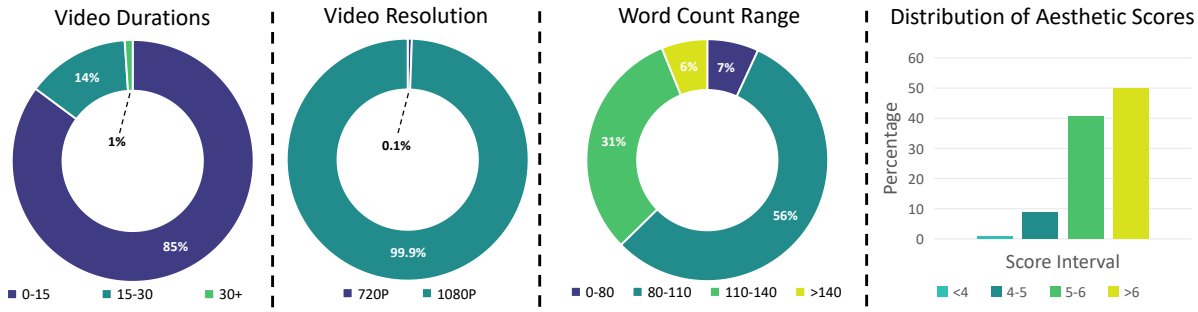


Figure 3. **Video Clips Statistics of our Dataset.** The dataset includes a diverse range of categories, durations and caption lengths, with most of the videos being in 1080P resolution.

ing box (bbox) to compute a unique identifier for each individual. Specifically, using the "face" bbox as an example, we first determine the maximum number of individuals, m, present in the video based on the bbox information, and then assign a unique identifier (not exceeding m) to each "face" bbox in the initial frame. We then apply forward propagation: for the n-th frame, each bbox is assigned a unique identifier, corresponds to those from frame $n-1$, based on the Intersection over Union (IoU) between all bboxes of frames $n-1$ and n. After completing forward propagation for all frames, we perform backward propagation: for each bbox in frame n, we assign a unique identifier that corresponds to frames $n-1$ and $n+1$, again based on the IoU between all bboxes of frames n and $n+1$. Ultimately, each bbox is assigned a unique identifier, facilitating precise tracking of each person's location across video.

Segmentation and Captioning To facilitate the application of dynamic mask loss, we first input the highest-confidence bounding box (bbox) for each category obtained in the previous step into SAM2 [54] to generate the corresponding masks for each person's "face," "head," and "person." Subsequently, SAM2's [54] tracking signals are used to further refine the unique identifiers assigned earlier. We then employ Time-Aware Annotation [90], leveraging Qwen2-VL-72B [67], to produce high-quality captions for the video clips. Data statistics are presented in Figure 3.

2. Additional Experimental Results

2.1. Comparison with Closed-source Method

In this section, we compare our ConsisID with Vidu 1.5 [6], the only available video generation model capable of performing the Identity-Preserving Text-to-Video (IPT2V) task. Notably, Vidu 1.5 [6] is a closed-source model, and it remains unclear whether it is tuning-free or tuning-based. As Vidu's API is not publicly available, we can only manually collect output to evaluate its performance. Due to constraints on large-scale generation, we randomly select 60 prompts from the evaluation dataset. Each prompt is paired with a unique reference image, resulting in the generation of

Figure 4. **Seamlessly integrate the frame interpolation model [28] into ConsisID.** The newly added frames are clear, which shows that the original video generated by ConsisID is coherent.

	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
Vidu 1.5 [6]	0.36	0.39	32.87	215.42
ConsisID	**0.52**	**0.54**	**33.08**	**163.68**

Table 1. **Quantitative Comparison with Close-source Method.** ConsisID performs best on most metrics, especially FaceSim, which is the most important metric of IPT2V. "↓" denotes lower is better. "↑" denotes higher is better.

	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓	Speed (s) ↓
DreamVideo [72]	0.03	0.03	26.03	237.91	3600+
MotionBooth [73]	0.05	0.06	24.42	287.90	600+
Magic-Me [46]	0.09	0.10	23.14	237.35	500+
ConsisID	**0.46**	**0.47**	**27.45**	**181.97**	100+

Table 2. **Quantitative Comparison with Tuning-based Methods (on single Nvidia H100).** ConsisID can generate high-quality id-preserving videos in a very short time, achieving the best balance among methods. "↓" denotes lower is better. "↑" higher is better.

	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
CogVideoX-5B-I2V [83]	0.37	0.38	**28.53**	201.69
EasyAnimate v4 [76]	0.15	0.15	27.95	235.68
OpenSora-Plan v1.3 [37]	0.31	0.32	27.25	224.99
DynamiCrafter512 [74]	0.25	0.26	29.76	212.13
ConsisID	**0.46**	**0.47**	27.45	**181.97**

Table 3. **Quantitative Comparison with I2V Methods.** End-to-end methods yield higher-quality id-preserving videos than two-stage methods. "↓" denotes lower is better. "↑" higher is better.

60 videos by each method. The quantitative results are presented in Table 1, indicating that ConsisID consistently outperforms Vidu 1.5 [6] across all four automatic metrics. The qualitative analysis, illustrated in Figure 13, further supports these findings. Both ConsisID and Vidu 1.5 follow prompts effectively to generate the specified actions, backgrounds, and attributes. However, Vidu's outputs contain noticeable artifacts (*e.g.*, the flower in case 1 and the tree in case 2) and exhibit lower visual quality compared to ConsisID. Moreover, in terms of preserving identity features, Vidu 1.5 retains only high-frequency facial characteristics (*e.g.*, hair color and hairstyle in case 1, facial shape in case 2), while failing to maintain intrinsic ID features essential for IPT2V. Consequently, individuals generated by Vidu 1.5 do not align consistently with the reference images.

2.2. Comparison with Tuning-based Methods

In this section, we compare ConsisID with several tuning-based methods, including DreamVideo [72], MotionBooth [73] and Magic-Me [46]. These methods require parameter tuning for each new identity input prior to inference. Due to computational and time constraints, we randomly select 1 reference image per ID from our evaluation dataset and each image is evaluated on 45 randomly selected prompts, resulting in a total of 1,350 video sequences generated per method. The results are presented in Figure 13 and Table 2. As shown, ConsisID consistently outperforms existing tuning-based methods despite having a shorter inference time (on single Nvidia H100). This superior performance is likely due to the fact that the latter are designed for open-domain tasks (*e.g.*, people, objects, animals, plants), which encompass a wide range and consequently fail to effectively capture the nuanced distinctions of individual identity features, leading to suboptimal results.

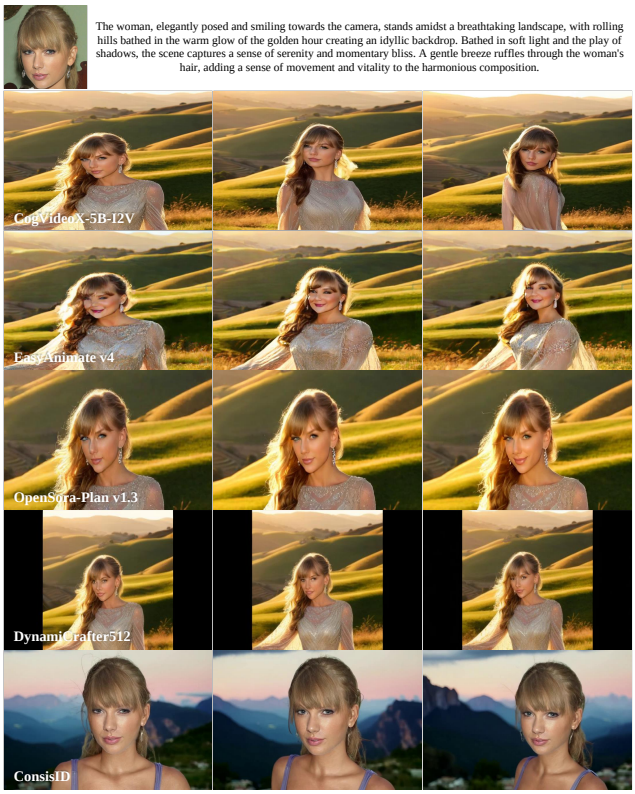


Figure 5. **Qualitative comparison with I2V Methods.** It is clear that standard I2V models encounter challenges in generating high-quality identity-preserving videos.

	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
w/o CLIP	0.40	0.38	26.87	142.23
w/o FaceExtractor	0.36	0.37	27.76	193.99
w/o Noise ζ	0.37	0.38	27.52	193.46
Loss †	0.35	0.37	26.53	167.11
Loss ††	0.39	0.38	26.89	216.69
Loss ‡	0.29	0.30	24.77	150.80
ConsisID †	0.40	0.40	27.38	256.29
ConsisID	**0.46**	**0.47**	**27.45**	**181.97**

Table 4. **Fine-grained Ablation Study.** Each component of ConsisID plays a crucial role in generating high-quality videos. "↓" denotes lower is better. "↑" higher is better.

2.3. Comparison with I2V Methods

In this section, we compare ConsisID with several image-to-video methods, leveraging the identity-preserving image model [36]. As shown in Figure 5, Table 3 and Table 5, the I2V foundation models [74, 76, 83] clearly demonstrates considerable temporal decay in identity preservation. While OpenSora-Plan [37] achieves higher fidelity due to its lower motion amplitude, it does not align with the real video. In contrast, only the proposed ConsisID consistently preserves identity throughout the entire video.

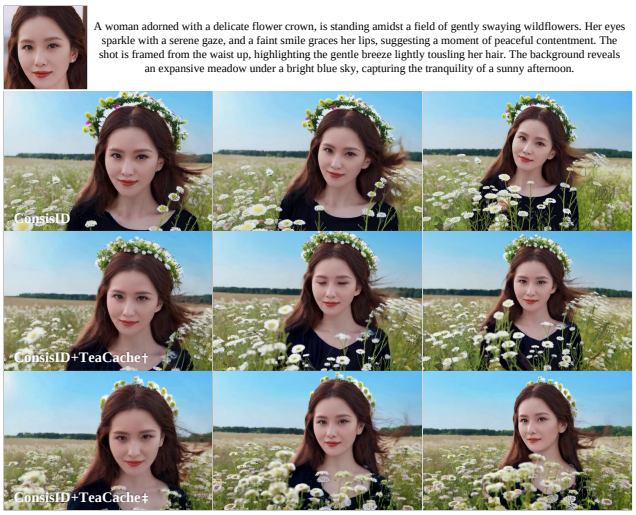


Figure 6. **Increasing Inference Speed with the help of TeaCache [38].** ConsisID can seamlessly integrate into existing inference acceleration frameworks without much quality degradation, demonstrating its strong scalability.

2.4. Fine-grained Ablation Study

In this section, we conduct more detailed ablation studies. Due to limited computational resources, we extract approximately 30K video samples from the ConsisID-Dataset for the following experiments. The batch size is reduced to 16, and the training duration is set to one epoch, with all other settings remaining unchanged. The experiments including:

- **Ablation on the Local Facial Extractor**: This experiment aims to assess the distinct roles of CLIP and the Face Extractor in capturing high-frequency facial features, specifically *w/o CLIP* and *w/o Face Extractor*.
- **Ablation on Noise in Cross-Face Loss**: This experiment examines whether introducing subtle noise into the input image improves the model's generalization capability, specifically *w/o Noise* ζ.
- **Ablation on the Weight Ratio of Loss**: This experiment investigates the individual roles of MSE Loss, Dynamic Mask Loss, and Dynamic Cross Loss to the training process. *Loss* † corresponds to a mix ratio of 2:1:1, *Loss* †† to a ratio of 1:2:1, and *Loss* ‡ to a ratio of 1:1:2.

The results are shown in Figure 7 and Table 4, where *ConsisID* † represents the complete model trained on the subset data. It can be concluded that CLIP is essential for acquiring the semantic information necessary for video editing, as removing it leads to a significant decrease in the CLIPScore. FaceExtractor plays a critical role in maintaining facial consistency, with its removal resulting in a drop in FaceSim scores. Both Noise ζ and Dynamic Cross Loss contribute positively to the model's generalization performance; however, an overemphasis on the latter may prevent convergence. MSE Loss accelerates convergence, while

Figure 7. **Fine-grained Ablation Study.** After the removal of CLIP, the model loses essential semantic information necessary for editing. FaceExtractor and MSE Loss play a critical role in maintaining consistency of facial features. Over-reliance on Dynamic Mask Loss may result in the loss of background information. Noise ζ and Dynamic Cross Loss are vital for the model's generalization; without them, the model struggles to produce results beyond the training data.

Model	Memory	Paramaters	Speed
ID-Animator [19]	8GB	1.5B	~11s
CogVideoX-5B-I2V †† [83]	42GB	5.2B	~210s+7s
EasyAnimate v4 †† [76]	15GB	1.8B	~78s+7s
OpenSora-Plan v1.3 †† [37]	37GB	2.6B	~282s+7s
DynamiCrafter †† [74]	17GB	2.4B	~26s+7s
CogVideoX-5B-I2V [83]	42GB	5.2B	~210s
ConsisID	44GB	5.7B	~214s
ConsisID+TeaCache † [38]	44GB	5.7B	~137s
ConsisID+TeaCache ‡ [38]	44GB	5.7B	~103s

Table 5. **Computation Overhead of Different Methods (on single Nvidia A100).** Compared to the baseline, ConsisID introduces only a minimal overhead to achieve the IPT2V task. †† means generating video with the help of PhotomakerV2 [36].

Figure 8. **Style Transfer Applications.** Despite being trained on real facial data, ConsisID demonstrates a remarkable generalization by generating anime-style videos in a zero-shot manner.

Dynamic Mask Loss enhances focus on facial features, thereby improving identity consistency. However, excessive reliance on Dynamic Mask Loss may lose the ability to generate background content. The complete model, integrating all components, yields optimal performance.

2.5. Ablation on the Number of Inference Steps

To assess the impact of varying the number of inference steps on model performance, we conduct an ablation study within the inference phase of ConsisID. Given constraints on computing resources, 60 prompts are randomly selected from the evaluation dataset. Each prompt is paired with a unique reference image, leading to the generation of 60 videos for each setting. Using a fixed random seed, we vary the inversion step parameter t across values of 25, 50, 75, 100, 125, 150, 175, and 200. The results are illustrated in Figure 9 and Table 6. Although theoretical expectations [24, 62, 63] suggest that increasing the number of inference steps would continuously enhance the generation quality, our findings indicate a non-linear relationship where quality peaks at $t = 50$ and subsequently declines. Specifi-

	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓	Speed (s) ↓
$t = 25$	0.50	0.53	30.43	184.44	**50+**
$t = 50$	**0.52**	0.54	**33.08**	**163.68**	100+
$t = 75$	0.43	0.52	31.92	200.86	160+
$t = 100$	0.46	**0.55**	32.25	212.74	220+
$t = 125$	0.42	0.51	32.38	185.85	270+
$t = 150$	0.34	0.40	32.41	186.56	330+
$t = 175$	0.35	0.42	29.98	186.99	390+
$t = 200$	0.33	0.39	31.18	166.79	440+

Table 6. **Effect of the Inference Steps by Quantitative Analysis (on Nvidia H100).** "↓" denotes lower is better. "↑" higher is better.

cally, at $t = 25$, the model produces incomplete garlands; at $t = 75$, it fails to generate upper body clothing; beyond $t = 125$, it loses critical low-frequency facial information, resulting in distorted facial features; and beyond $t = 150$, the visual clarity progressively deteriorates. We infer that the initial stages of denoising process are dominated by low-frequency information, such as generating the outline of a face, while the later stages focus on high-frequency details, such as intrinsic facial features. $t = 50$ is just the optimal setting to balance these two stages.

2.6. Increasing Inference Speed

As shown in Table 5, ConsisID requires about 44 GB of GPU memory to decode 49 frames with output resolution 720x480 ($W \times H$), while baseline needs 42 GB of GPU memory. The inference time of ConsisID is almost identical to the baseline, with only an additional 0.5B parameters, yet it achieves the IPT2V functionality that the baseline cannot, demonstrating the efficiency of the proposed method. In addition, ConsisID can seamlessly integrate with training-free inference acceleration strategies, achieving minimal degradation in visual quality, as illustrated in Figure 6. Specifically, TeaCache† corresponds to setting $rel\_l1\_thresh = 0.1$, while TeaCache‡ corresponds to setting $rel\_l1\_thresh = 0.15$. The rel_l1_thresh regulates the trade-off between generation quality and speed.

2.7. Generating Higher FPS Videos

Due to limited resources, ConsisID can only generate 49 frames at 8 frames per second (fps). Although the resulting video is coherent, the frame rate falls below the human perceptual threshold for smoothness, which is approximately 16 fps. Therefore, a frame interpolation model [28] can be applied to post-process the output video, increasing the frame rate to 16 fps, as illustrated in Figure 4. The results indicate that after interpolation, the video maintains a high level of clarity, suggesting that the original frames generated at fps 8 are sufficiently coherent.

2.8. Style Transfer Applications

Figure 8 demonstrates the generalization capability of ConsisID. Beyond generating realistic, customized videos, the framework effectively processes stylized prompts while preserving the identity of animated characters in a zero-shot manner. These capabilities are expected to significantly advance video content creation.

2.9. More Cases on ID-preserving Videos

Due to space constraints, to assess the robustness and generalizability of our method, we present more ID-preserving video generation results in Figures 14, 15 and 16, covering different people and different text prompts. ConsisID not only generates faces that match the identity of the reference image but also adheres to the text prompt, allowing for control over the character's expressions, attire, actions, age, background, and even camera angles (*e.g.*, detailed close-ups, wide panoramic views). These results substantiate the effectiveness of the Global / Local Facial Extractors, and the Consistency Training Recipe can enhance performance.

3. Additional Experimental Details

3.1. Visualization of Different Injection Methods

To enhance the explanation of *Identity Signal Injection in DiT* presented in the main text, we visualized various schemes, as shown in Figure 10. For simplicity, the visualization of the text branch is omitted. ConsisID employs scheme (c), which injects high-frequency information between the Attention and FFN modules, while integrating low-frequency signals (with facial key points) into the shallow layers of the network, achieving optimal result.

3.2. Validation of the Automatic Metrics

In order to assess the effectiveness of the different metrics, we preform a cross-validation using the results of the user study. Specifically, we obtain FaceSim-Arc [13], FaceSim-Cur, CLIPScore [21] and FID [22] scores for each video in the questionnaire. We then identify the highest scoring option for each metric and compared these results with the questionnaire responses, as shown in Figure 11. Although CLIPScore [21] reflect model performance reasonably well, their alignment with human perception remains limited. In particular, FID [22] showed an inverse relationship with human perception, with the lower quality ID-Animator [19] receiving a higher score. Therefore, the quantitative results presented in the main text should be interpreted cautiously.

3.3. Details of Resource

We employ Nvidia H100 (x40) and A100 (x8) for all the experiments. All implementations are conducted on the basis of the official code using the PyTorch framework.

3.4. Details of Evaluation Models

Vidu [6]. Vidu1.5 is currently the only closed-source model supporting tuning-free IPT2V. It can generate videos

Figure 9. **Effect of the Inference Steps t.** Overall quality does not improve consistently as t increases, but first improves and then declines. This may be because the early steps are dominated by low frequency, whereas the later steps are dominated by high frequency.

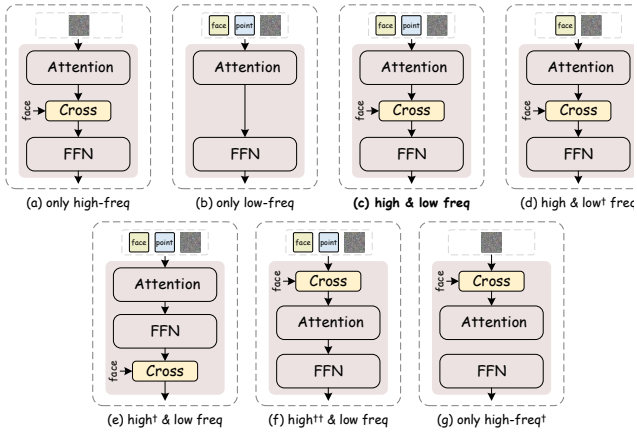


Figure 10. **Visualization of Different Methods of Injecting Control Signals into DiT.** Only (c), which injects high-frequency information between the Attention and FFN modules, while incorporating low-frequency signals, including facial key points, into the shallow layers of the network, resulting in optimal performance.

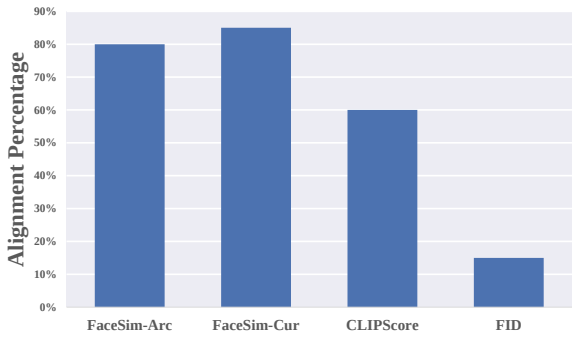


Figure 11. **Cross Validation between Automatic Metrics and Human Perception.** Existing metrics show limited alignment with human, particularly CLIPScore and FID.

of 4 or 8 seconds in length, with resolutions of 480p, 720p, or 1280p, and aspect ratios of 16:9, 9:16, or 1:1. We used its official default settings to generate 4-second, 480p, 16:9 videos for best comparison.

ID-Animator [19]. ID-Animator is the sole open-source model currently supporting tuning-free IPT2V. It utilizes a UNet-based architecture and is designed to generate 2-second (16-frame) videos at a resolution of 512×512. We utilized its official default configuration of DreamBooth (realisticVisionV60B1) to generate videos for comparison.

DreamVideo [72]. DreamVideo is an open-source model supporting tuning-based IPT2V. It utilizes a UNet-based architecture and is designed to generate 4-second (32-frame) videos at a resolution of 256 × 256. For fairness, we use only one reference image and the official default settings (e.g., steps, learning rate) to train the model, and then generate videos for comparison.

MotionBooth [73]. MotionBooth is an open-source model supporting tuning-based IPT2V. It utilizes a UNet-based architecture and is designed to generate 576×320×24 and 512×320×16 videos, respectively. For fairness, we use only one reference image and the official default settings (e.g., 512×320×16, steps) to train the model, and then

generate videos for comparison. Due to the requirement to specify the direction of camera movement, we fix the camera to move to the left.

Magic-Me [46]. Magic-Me is an open-source model supporting tuning-based IPT2V. It utilizes a UNet-based architecture and is designed to generate 8-second (16-frame) videos at a resolution of 1024×1024. For fairness, we use only one reference image and the official default settings (e.g., steps, learning rate) to train the model, and then generate videos for comparison.

CogVideoX [83], EasyAnimate [76], OpenSora-Plan [37], DynamiCrafter [74]. These models are open-source foundational generation models supporting image-to-video generation. While CogVideoX, EasyAnimate, and OpenSora-Plan are based on DiT architecture, DynamiCrafter employs a UNet-based architecture. Due to the lack of support for IPT2V in all these models, the process begins by generating the initial frame using PhotomakerV2 [36]. Subsequently, the respective models are used to generate the subsequent frames: CogVideoX-5B-I2V, EasyAnimate v4, OpenSora-Plan v1.3, and DynamiCrafter512. Due to the differences in supported resolution and length, we use the official default settings to ensure optimal performance.

3.5. Details of Implementation

(1) The section of *Quantitative Analysis* requires each model to generate 13,500 videos (30×5×90). To minimize computational overhead, we select only 2 reference images per ID, each with 90 text prompts in the section *Effect of the Identity Signal Injection in DiT* (30×2×90); 60 IDs each with 1 text prompt in the section *Comparison with Tuning-based Methods* and *Ablation on the Number of Inference Steps* (60×1×1); and select only 1 reference image per ID, each with 45 text prompts for the remaining experiments (30×1×45), including the *Comparison with I2V Methods*, *Fine-grained Ablation Study*, etc. **(2)** For all the baselines used in this paper, including Vidu [6], ID-Animator [19], DreamVideo [72], MotionBooth [73] and Magic-Me [46], we use the default settings from their official websites (*e.g.*, resolution, fps, inference steps, training steps, etc.) to ensure optimal results. For MotionBooth [73], since it requires the direction of camera movement to be specified, we set the camera to move to the left, which is also the official setting. **(3)** The evaluation dataset consists of 30 individuals, including celebrities, ordinary people, and those of diverse skin tones, as demonstrated by the qualitative results presented in this paper. This diversity enhances the comprehensiveness and reliability of the experimental data. Furthermore, the text prompts cover a wide range of expressions, actions, and backgrounds, providing a thorough assessment of the generalizability of IPT2V.

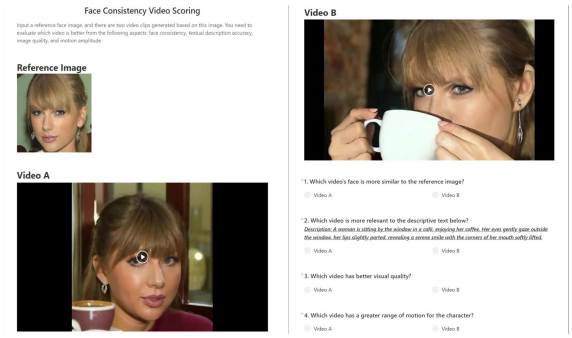


Figure 12. **Visualization of the Questionnaire for User Study.**

3.6. Details of Human Evaluation

To illustrate the user study intuitively, we provide a visualisation of the questionnaire in Figure 12. In addition, to increase the reliability and diversity of the questionnaire, we implement the following rules:

- The presentation order of different videos is randomized to reduce cognitive bias among respondents.
- A sliding verification upon submission is required to confirm that all responses are submitted manually and not by automated bots.
- Each IP address is restricted to a single submission, and users are required to log in before voting to ensure each individual could submit only once.
- Questionnaires where 90% of responses selected the same option (all A or all B) are discarded.
- The primary voting population consisted of undergraduate, master's, and PhD students from universities, along with a portion of the general public outside the field.
- The validity of data is assessed based on the time spent completing the questionnaire; responses with completion times of less than two minutes are excluded, as it typically takes 2–5 minutes to complete.
- Validity is also evaluated based on response distribution. Since the A/B options are randomly assigned for each question, we discarded questionnaires where 90% of responses favored a single option (either all A or all B).

4. Additional Statement

4.1. Support for Key Findings

These findings are not merely our observations but are synthesized from and validated by existing diffusion and ViT literature, with additional support provided by our experiments in Main Sec. 4.5. **Finding 1)** [[53] (Sec. 3.2), [61] (Sec. 1)] highlight that diffusion models' noise prediction is fundamentally low-level and benefits from a U-Net bias. In Main Sec. 4.5, Main Table 3 (f–g) shows training instability when low-frequency features are omitted, reinforcing their crucial role. **Finding 2)** [[88] (Sec. 1)] shows the im-

portance of high-frequency features in the diffusion model. [[5] (Abs), [3] (Sec. 1)] note that vision transformers' challenges with capturing high-frequency detail. Convergent evidence in Main Figure 7 (f) indicates transformers' high-frequency handling warrants deeper investigation. In Main Sec. 4.5, Main Figure 7 confirms improvements when decoupling high- and low-frequency signals, despite a legend error we have corrected.

4.2. How to ensure Fine-grained Feature by LFE?

For how to ensure that the feature output by Local Facial Extractor remains fine-grained: Q-Former serve as a fusion mechanism rather than an extractor. Among the input into it, Face Extractor, as a face recognition backbone, plays a central role by inherently extracting fine-grained features invariant to non-identity attributes (e.g., expression, posture). CLIP provides secondary semantic features for editing, while Dropout [2, 26] are employed to it to maintain the Q-Former's fine-grained feature dominance. Main Table 3 shows both modules are distinct and complementary.

4.3. Why not base on UNet?

Sora [8], based on the DiT architecture, exhibits significant potential in simulating the physical world. Recently, foundational models for visual generation have shifted from UNet [16, 74] to DiT [50, 76, 83, 96, 97], owing to the latter's scalability and superior performance. Accordingly, our ConsisID is based on DiT instead of UNet architecture.

4.4. Can we Generate Image?

Despite being trained exclusively on video data, ConsisID can leverage the generalization capabilities of CogVideoX [83] to generate high-quality, identity-preserving images. This is achieved by either setting the $frame$ parameter to 1 or extracting the first frame of a video as an image.

4.5. Ethics Statement

ConsisID is capable of generating high-quality, realistic human videos. However, it also presents potential negative impacts, as the technique may be utilized to produce deceptive content for fraudulent activities. It is important to recognize that any technology is susceptible to misuse [78–81].

4.6. Reproducibility Statement

First, we have explained the implementation of ConsisID in detail in section 3. Second, we have explained the details of training and inference in section 4. Finally, the data and codes used in this work will be open-source online.

4.7. Copyright Statement

The training data is sourced from in-house datasets, and only a subset of the data (CC BY 4.0 license) will be made publicly available. The video content exclusively features humans, and any NSFW content is detected and removed based on the video captions. The videos come from different regions of the world to ensure they are representative.

4.8. Limitations and Future Work

Existing metrics fail to accurately assess the capabilities of different ID preservation models. While ConsisID can generate realistic and natural videos based on a text prompt, commonly used metrics such as CLIPScore [21] and FID [22] exhibit minimal differences compared to previous methods. A promising approach is to develop a metric that better aligns with human perception.

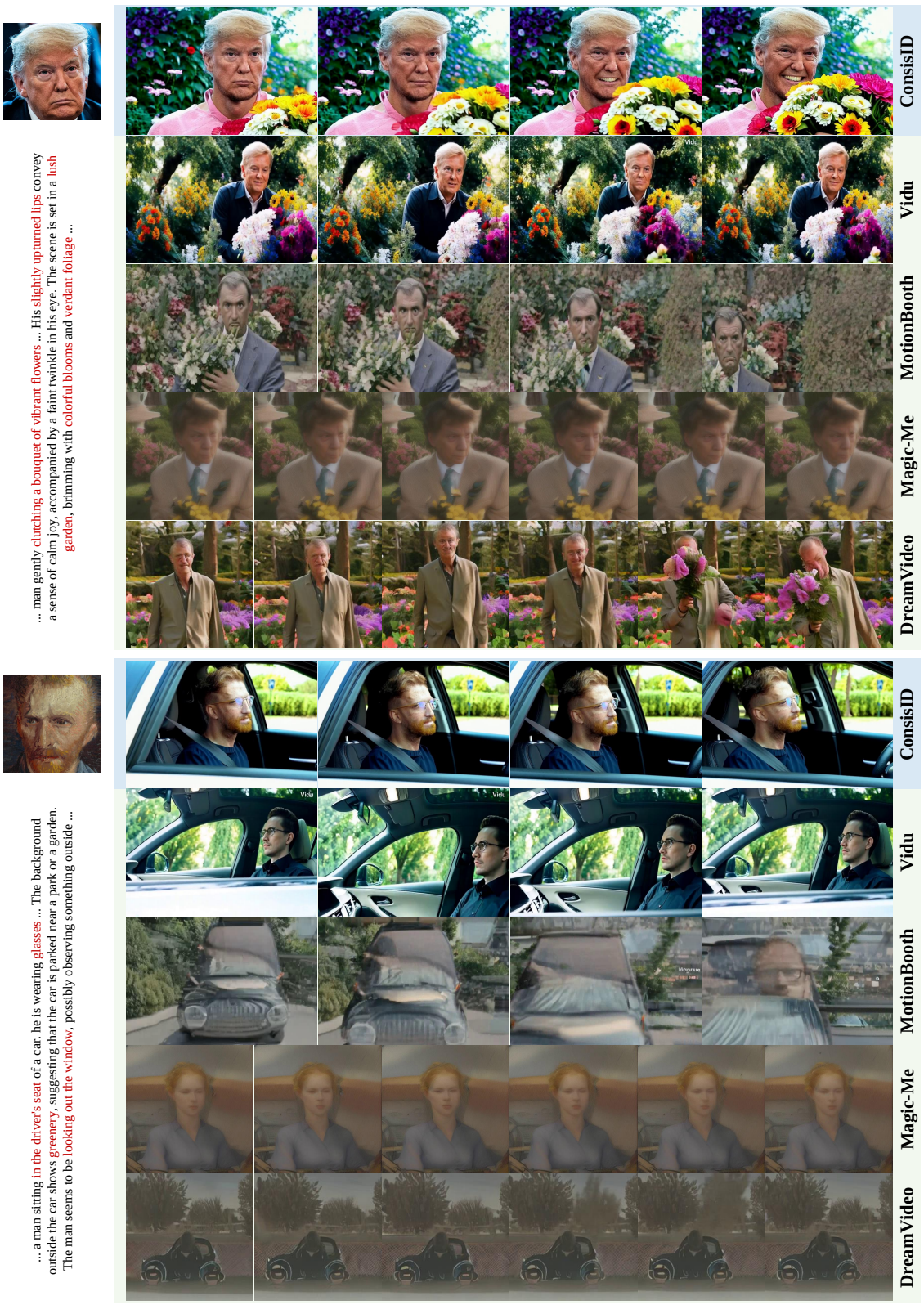


Figure 13. **Quantitative comparison with Closed-source and Tuning-based Identity-Preserving Videos Generation methods.** ConsisID achieves advantages in identity preservation, visual quality, motion amplitude, and text relevance. Moreover, only ConsisID, Vidu [6] and MotionBooth [73] can generate aesthetically pleasing *horizontal* videos, while the others [46, 72] can only produce *square* videos.

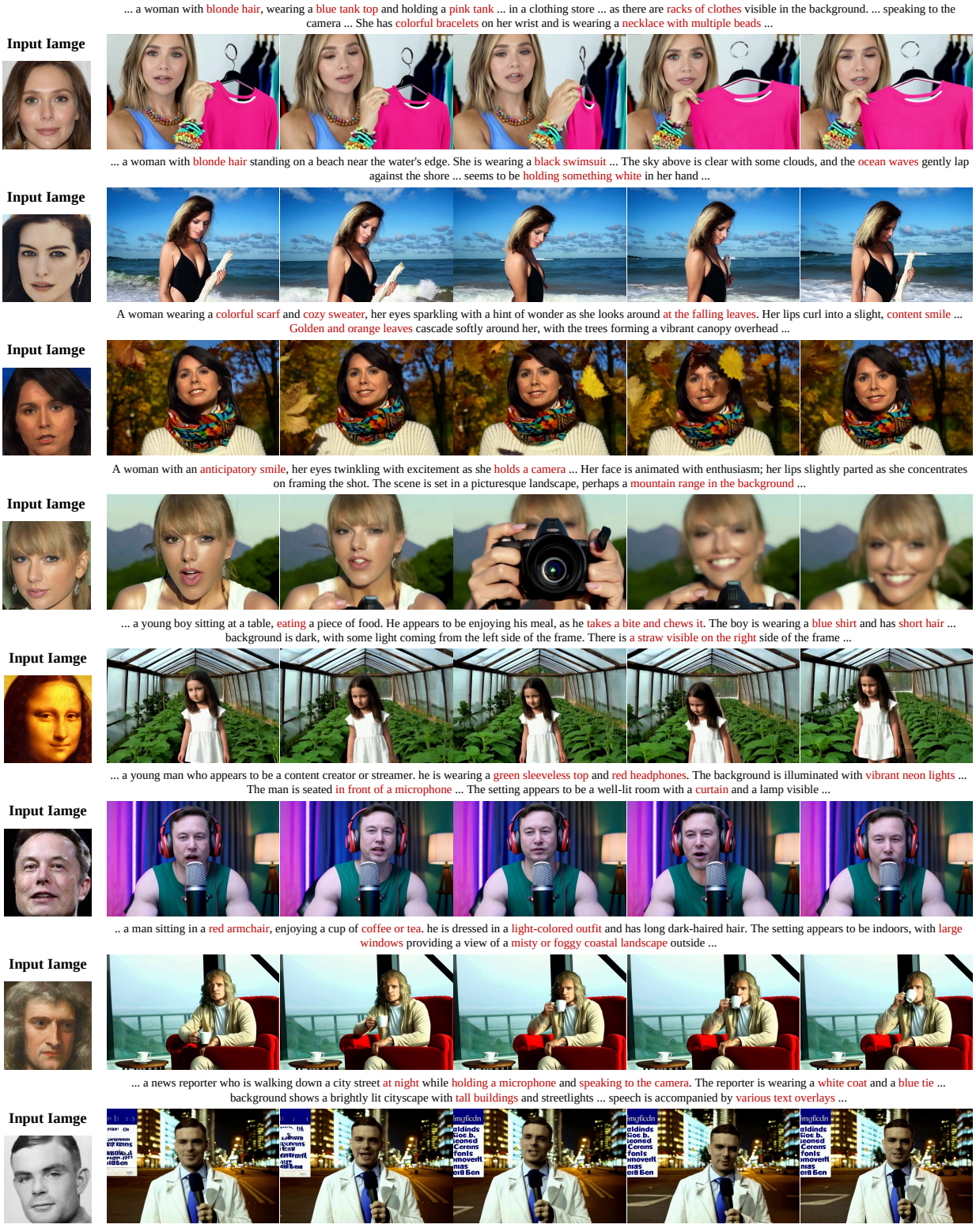
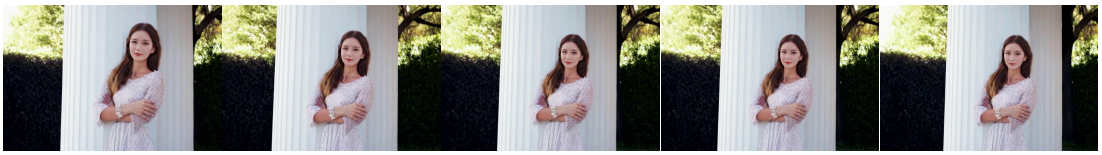


Figure 14. **Showcases of Identity-Preserving Videos Generated by ConsisID.** Our method consistently generates realistic videos that match the input identity while enabling precise control through text prompts, demonstrating significant practical utility.

... a woman ... leaning against a tall, white column ... Her arms are crossed, and she appears to be posing for the camera. The background is lush with greenery, including trees and bushes, suggesting that the location is a park or garden ...

Input Iamge



... a young man ... in a snowy park ... wearing a colorful winter jacket with a floral pattern ... shows a snowy landscape with trees, benches, and a metal fence ... he smiles and gives a thumbs-up gesture towards the camera ...

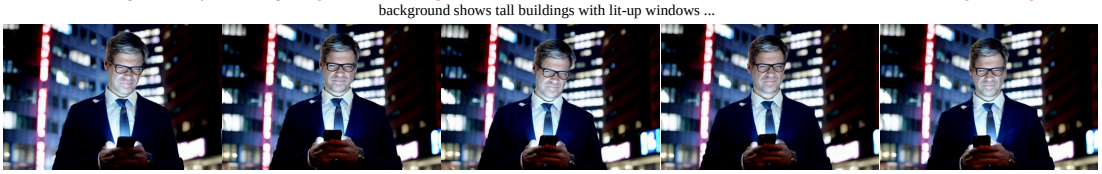
Input Iamge

... a woman standing in front of a large screen ... She is wearing a purple top and appears to be presenting or speaking about technology-related topics. The background includes a cityscape with tall buildings, suggesting an urban setting ...

Input Iamge

... a man walking down a city street at night, engrossed in his smartphone. He is dressed in a formal suit and tie ... street is illuminated with various neon lights and signs ... background shows tall buildings with lit-up windows ...

Input Iamge

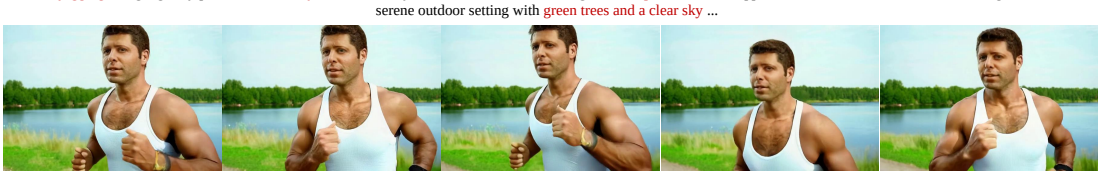


... a woman dressed as a mermaid, swimming underwater. She is wearing a silver tail and a matching top, which is adorned with colorful patterns. The woman has long, wavy hair that flows freely underwater. The water around her is clear and blue, with sunlight filtering through the surface ...

Input Iamge

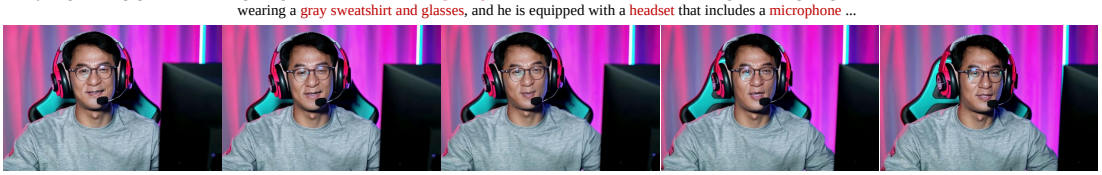
... a man jogging along a grassy path next to a body of water, likely a lake or river. he is wearing a white sports cloth and appears to be focused on his run. The background shows a serene outdoor setting with green trees and a clear sky ...

Input Iamge



... a young man engaged in an intense gaming session. he is seated in a gaming chair ... The chair has a distinctive design, featuring a logo or emblem on the backrest. The man is wearing a gray sweatshirt and glasses, and he is equipped with a headset that includes a microphone ...

Input Iamge



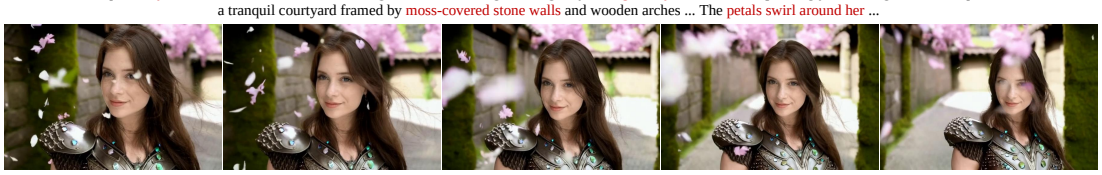
... a young man walking through a park during sunset ... he is wearing a sleeveless top with a geometric pattern and denim shorts ... long, dark hair that falls over his shoulders ··· holds a skateboard ... orange and green in color. The park is lush with greenery, and the sunlight filters through the trees ...

Input Iamge

Figure 15. **More Showcases of Identity-Preserving Videos Generated by ConsisID.**

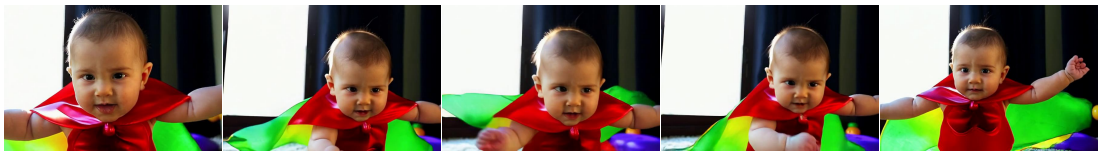
... a woman in exquisite hybrid armor adorned with iridescent gemstones, standing amidst gently falling cherry blossoms. Her piercing yet serene gaze hints at quiet determination ... a tranquil courtyard framed by moss-covered stone walls and wooden arches ... The petals swirl around her ...

Input Iamge



... a baby wearing a bright superhero cape, standing confidently with arms raised in a powerful pose ... with eyes wide and lips pursed in concentration, as if ready to take on a challenge. The setting appears playful, with colorful toys scattered around and a soft rug underfoot ...

Input Iamge



... a boy walking along a city street, filmed in black and white on a classic 35mm camera. His expression is thoughtful, his brow slightly furrowed as if he's lost in contemplation ... the cityscape is filed with vintage buildings, cobblestone sidewalks, and softly blurred figures passing by, their outlines faint ...

Input Iamge

... a man standing at an easel, focused intently as his brush dances across the canvas. He wears a paint-splattered apron ... The setting, filled with scattered art supplies, open paint tubes, and unfinished sketches pinned to the wall ... A large window on one side allows sunlight to stream in ...

Input Iamge

... a little girl with pigtails, a backpack, and a bright smile, skipping joyfully down a bustling city street. She holds a bunch of colorful balloons that sway as she moves, her eyes wide with excitement. The backdrop of vibrant buildings and street vendors ...

Input Iamge

... a girl sitting on a stool, playing a guitar with her fingers moving skillfully across the strings, creating a soulful melody. Her eyes are closed, and there's a slight smile on her face, conveying a sense of deep connection to the music ...

Input Iamge

... a man with a rugged beard, wearing a leather jacket, riding a vintage motorcycle along a desert highway ... eyes narrowed slightly against the wind, as the setting sun casts a warm glow over the landscape ... arid land with occasional cacti and rocky outcrops ...

Input Iamge

... a man celebrating his birthday, holding a piece of cake while he prepares to blow out the candles. He is smiling warmly, and as he closes his eyes to make a wish ... Around him, soft, ambient lighting from nearby candles casts a warm glow, and there are a few balloons and decorations in the background ...

Input Iamge

Figure 16. **More Showcases of Identity-Preserving Videos Generated by ConsisID.**