Research Article

MotionCharacter: Identity-Preserving and Motion Controllable Human Video Generation

Haopeng Fang[1,2], Di Qiu[2], He Tang[1]

1. Huazhong University of Science and Technology, China; 2. Meituan, China

Recent advancements in personalized Text-to-Video (T2V) generation highlight the importance of integrating character-specific identities and actions. However, previous T2V models struggle with identity consistency and controllable motion dynamics, mainly due to limited fine-grained facial and action-based textual prompts, and datasets that overlook key human attributes and actions. To address these challenges, we propose MotionCharacter, an efficient and high-fidelity human video generation framework designed for identity preservation and fine-grained motion control. We introduce an ID-preserving module to maintain identity fidelity while allowing flexible attribute modifications, and further integrate ID-consistency and region-aware loss mechanisms, significantly enhancing identity consistency and detail fidelity. Additionally, our approach incorporates a motion control module that prioritizes action-related text while maintaining subject consistency, along with a dataset, Human-Motion, which utilizes large language models to generate detailed motion descriptions. For simplify user control during inference, we parameterize motion intensity through a single coefficient, allowing for easy adjustments. Extensive experiments highlight the effectiveness of MotionCharacter, demonstrating significant improvements in ID-preserving, high-quality video generation.

Corresponding authors: Di Qiu, qiudihk@gmail.com; He Tang, hetang@hust.edu.cn

 doi.org/10.32388/S2EWVR

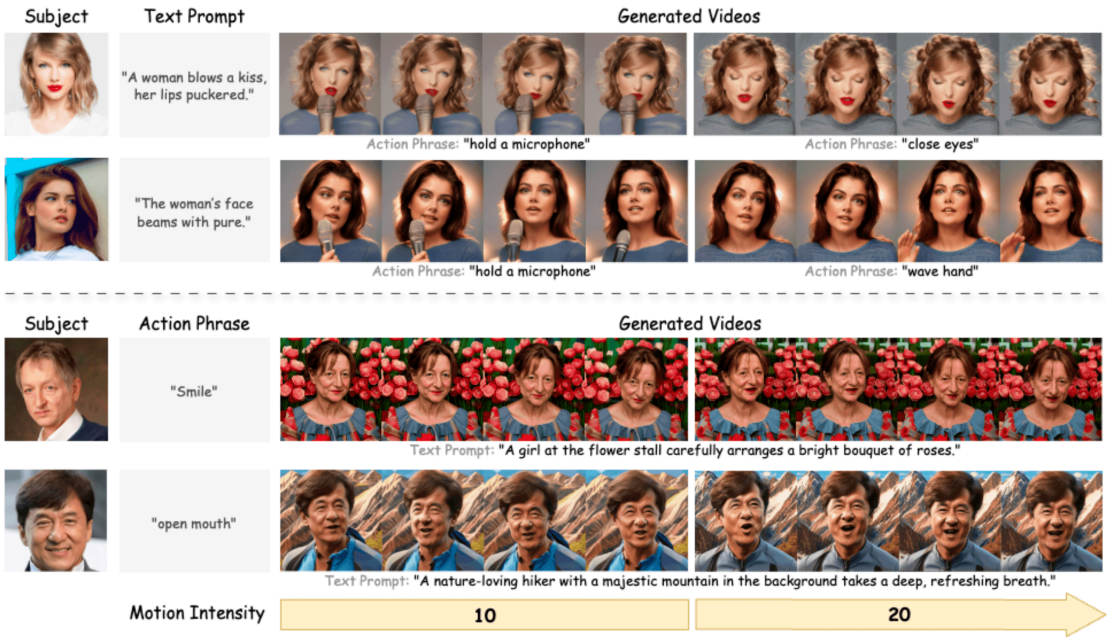

Figure 1. Given a single reference facial image, MotionCharacter can generate identity-consistent video outputs across text prompts, action phrases, and motion intensities. The upper section demonstrates its capability to accurately follow specific action phrases, while the lower section highlights its fine-grained motion control achieved by varying user-defined motion intensities.

1. Introduction

High-quality, personalized, and controllable human video generation has gained significant traction, with applications spanning social media, virtual avatars, and personalized content creation. Recent advancements in text-driven video generation models[1][2][3][4][5][6][7][8][9][10][11] have driven substantial progress in this field, yet several challenges remain, particularly in maintaining identity consistency and achieving fine-grained control over motion instructions. In response, recent approaches[12][13][14][15][16] in subject-driven Text-to-Video (T2V) generation have explored ways to address the challenge of producing high-quality videos that accurately depict specific individuals while following motion instructions consistently.

However, most of these approaches rely on separate training for each identity (ID), limiting their scalability and flexibility in practical applications. Some works such as ID-Animator[16] have advanced the field by enabling identity-specific video generation using any reference facial image without requiring additional fine-tuning, but it still lacks the ability to finely control the intensity of motion,

constraining its responsiveness to nuanced motion prompts. This limitation underscores the need for methods that can offer both identity preservation and precise motion control in T2V generation. Besides, the typical prompts provided by users to T2V models include descriptions of the entire video content, encompassing scene details and human motion. However, text-based prompts alone are often insufficient to capture fine-grained motion dynamics accurately. For example, phrases like “open mouth” or “eyes closed” omit key details such as movement speed or intensity, which are essential for nuanced motion dynamics. While a more intuitive approach might be to use motion-only prompts, current T2V models exhibit limited sensitivity to such concise motion instructions, often failing to reflect the intended subtleties in generated videos. In addition, the existing facial text-video datasets such as CelebV-Text[17] primarily focus on emotional changes, neglecting essential human attributes and actions, rendering them inadequate for identity-preserving video generation tasks.

In this paper, we propose MotionCharacter, a human video generation framework specifically designed for identity preservation and fine-grained motion control. To achieve high-fidelity and identity consistency for human video generation, we introduce an ID-preserving module and employ a combination of face embedding and CLIP[18] embedding, allowing the model to retain high identity fidelity while also being flexible enough to accommodate dynamic modifications of attributes such as expressions or actions based on user prompts. Additionally, we integrate a composite loss function that incorporates both ID-consistency loss and region-aware loss components that direct the model’s attention to critical facial regions, addressing common issues like distortion or blurriness in features such as lips and teeth, and enhance identity consistency and detail fidelity in personalized T2V generation.

To enhance the model’s responsiveness to motion instructions, we introduce a specialized motion control module that prioritizes action-related text cues while preserving identity consistency throughout the video sequence. Complementing this module, we present a new dataset, Human-Motion, which employs large language models (LLMs) to produce comprehensive and nuanced motion descriptions tailored for identity-preserving video synthesis. Furthermore, to facilitate user control over motion dynamics, we introduce a parameterized motion intensity coefficient, allowing users to easily adjust the scale of movement during inference. Together, these advancements improve the model’s ability to accurately follow motion instructions and generate realistic, personalized content. Through extensive experiments, we present qualitative, quantitative, and user study results that validate the effectiveness of

our method in terms of identity consistency and adherence to motion instructions. In summary, our contributions are as follows:

- We propose a framework, named MotionCharacter, designed to enhance identity consistency and fine-grained motion controllability in human video generation.
- We introduce a motion control module that prioritizes action-related text while maintaining subject consistency, which enables more precise control over motion dynamics and improves the generation of high-quality human videos based on textual descriptions.
- We propose region-aware loss to improve attention to critical facial regions, ensuring high-quality and identity-preserving video generation while maintaining accurate motion dynamics.

2. Related Work

Text-to-Video Diffusion Model

Recent advancements in diffusion models[19][20][21] have positioned them as a prominent method in generative modeling, especially in text-to-video (T2V) generation. The Video Diffusion Model[1] was among the first to utilize a space-time-factored U-Net architecture for unconditional video generation, effectively modeling video distributions in pixel space. Building on this, AnimateDiff[2] advanced the field by incorporating a motion module into the Stable Diffusion framework, enhancing its ability to generate videos from textual prompts. Subsequent developments included Imagen Video[3] and Make-a-Video[4], which introduced sequential models for T2V generation, focusing on pixel-space representations. In response to the challenges associated with high-dimensional video data, latent video diffusion model[5] proposed leveraging latent diffusion models to operate within the latent space of an auto-encoder. This latent approach has gained traction, leading to a proliferation of methods such as ModelScope[6], LAVIE[7], MagicVideo[8], VideoCrafter[9][10], each contributing to the ongoing evolution of T2V models and their applications.

Identity-Preserving Image Generation

The task of identity-preserving image generation aims to synthesize visual content that maintains the unique characteristics of a specific individual while allowing for variations in pose, expression, and other attributes. Most techniques focus on facial images, using methods like texture-based approaches and latent space manipulation to preserve identity. While fine-tuning methods such as Low-Rank

Adaptation[22], Textual Inversion[23], and DreamBooth[24] can customize models for ID-specific images, they often require dedicated training for each identity. Recent approaches utilize embeddings as conditional inputs to guide the generation process, enabling explicit identity control. For example, IP-Adapter[25] integrates an adapter module with identity embeddings into a pre-trained model for efficient identity preservation. PhotoMaker[26] stacks multiple images to mitigate identity-irrelevant features. InstantID[27] allows real-time, identity-specific generation with a lightweight architecture, reducing the need for case-specific training. PuLID[28] employs a multi-stage process to refine identity embeddings, enhancing fidelity and visual quality.

Subject-Driven Text-to-Video Generation

Subject-driven text-to-video generation aims to incorporate specific characters or subjects into synthesized videos while allowing for text-based control over actions, styles, and sequences. Previous methods such as VideoBooth[12], DreamVideo[13], MagicMe[14], and CustomCrafter[15] have explored learning-based frameworks to combine visual identity with motion dynamics. However, these approaches often require separate training for each individual, which can limit their scalability and flexibility. In contrast, the recent development of ID-Animator[16] has demonstrated the ability to achieve zero-shot, training-free capabilities, allowing for identity incorporation without extensive retraining. Nonetheless, this method still lacks fine-grained control over motion intensity. Specifically, the inability to adjust the strength of movements can hinder the model's effectiveness in generating realistic and dynamic animations that align with user specifications. In response to these limitations, our approach, MotionCharacter, enables simultaneous control of both appearance and motion without necessitating retraining during inference, thereby enhancing the usability and efficiency of subject-driven video generation.

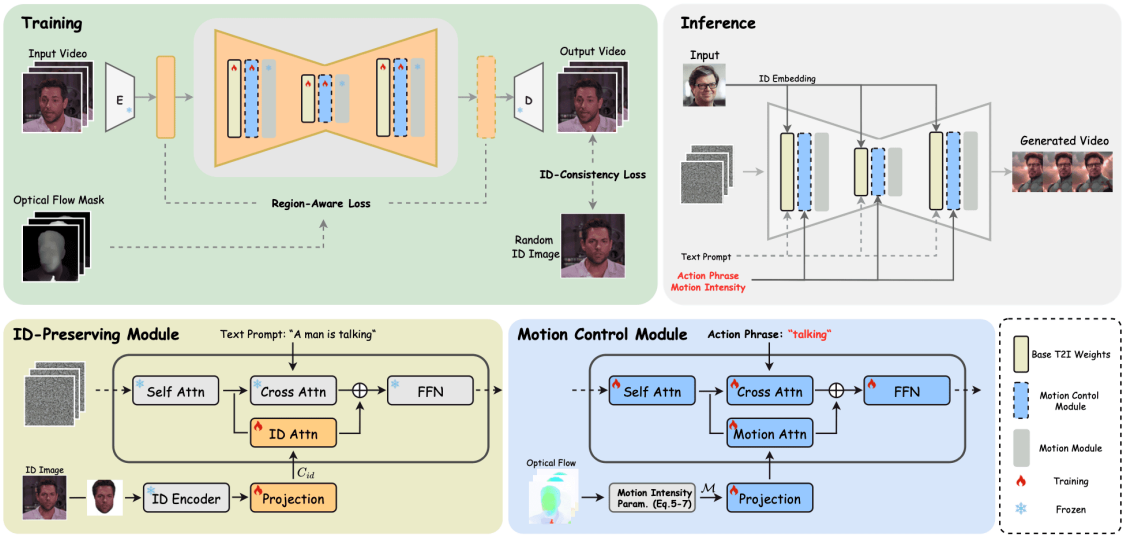

Figure 2. Framework overview. Our proposed framework comprises three core components: the ID-Preserving Module, the Motion Control Module, and a composite loss function. The loss function incorporates a Region-Aware Loss to ensure high motion fidelity and an ID-Consistency Loss to maintain alignment with the reference ID image. During training, motion intensity $\mathcal{M}$ is derived from optical flow. At inference, human animations are generated based on user-defined motion intensity $\mathcal{M}$ and specified action phrases, enabling fine-grained and controllable video synthesis.

3. Methodology

3.1. Problem Formulation

Personalized human video generation aims to create vivid clips consistent in character identity and motion based on a reference image and text prompt. To achieve this goal, we propose a novel model named MotionCharacter which accurately reflects identity information, captures action-based motion, and maintains smooth visual transitions. Formally, given a reference ID image $\mathcal{I}$, a text prompt $\mathcal{P}$, an action phrase $\mathcal{A}$, and a motion intensity $\mathcal{M}$, the model $\mathcal{F}$ is designed to produce video $\mathcal{V}$ by:

$$\mathcal{V} = \mathcal{F}(\mathcal{I}, \mathcal{P}, \mathcal{A}, \mathcal{M}). \tag{1}$$

Technically, we elaborately design the structure from two aspects, ID-Preserving Optimization and Motion Control Enhancement. Besides, we propose a new Human Motion dataset which is specially curated and annotated for training high-fidelity human video generation.

doi.org/10.32388/S2EWVR

3.2. ID-Preserving Optimization

ID Content Insertion

Since the adopted pretrained text-to-video (T2V) diffusion model[2] lacks identity-preserving capabilities, we first intend to introduce an ID-Preserving Adapter into the backbone to emphasize identity-specific regions and reduce irrelevant background interference. As illustrated in Fig. 2, the ID-Preserving Adapter extracts the identity embedding C_{id} from the reference image $\mathcal{I}$ and injects the identity embedding C_{id} into the diffusion model through cross-attention.

Specifically, the face region is first isolated from the reference image $\mathcal{I}$ to filter the interference of the background region. Then the face region image is processed in parallel to a pre-trained CLIP image encoder[18] and a face recognition model ArcFace[29] to obtain the broad contextual identity embeddings E_{clip} and the fine-grained identity embeddings E_{arc}, respectively. To effectively combine global context with fine-grained identity details, we employ cross-attention to fuse the CLIP and ArcFace embeddings:

$$C_{id} = \text{Proj}(\text{Attn}(E_{arc}W_q', EW_k', EW_v')), \tag{2}$$

where W_q', W_k', and W_v' are learnable parameters, with E_{arc} as the query and the combined embedding $E = E_{clip} + E_{arc}$ as the key and value. Following cross-attention, a projection layer Proj is applied to align the dimension with the text embedding, thereby generating the final identity embedding C_{id} for the reference image $\mathcal{I}$.

Inspired by recent work on image prompt adapters[25][27], the identity embedding C_{id} in MotionCharacter is regarded as an image prompt embedding and is used alongside text prompt embeddings to provide guidance for the diffusion model. This procedure can be expressed as:

$$z' = \text{Attn}(Q, K^t, V^t) + \lambda \cdot \text{Attn}(Q, K^i, V^i), \tag{3}$$

where the parameter λ controls the balance between text guidance and identity preservation. Here, $Q = zW_q$ is derived from the latent representation z, while $K^i = C_{id}W_k^i$ and $V^i = C_{id}W_v^i$ are identity-specific key and value matrices obtained from the identity embedding C_{id} of the reference image $\mathcal{I}$. Similarly, K^t, and V^t are the key and value matrices for the text cross-attention.

ID-Consistency Loss

Current stable diffusion models are usually trained with the MSE loss over each pixel, which is insufficient to ensure identity preservation between the reference image and the generated videos. To

address this issue, we introduce an ID-Consistency loss during training phase to maintain the identity information.

Specifically, at a specific diffusion step $\hat{t}$, the diffusion model can estimate the noise-free latent $\hat{z}_0$ from a noisy latent $z_{\hat{t}}$ by DDIM reversion process. Then, the estimated $\hat{z}_0$ is passed to a VAE decoder to reconstruct the frame, denoted as X^f. Therefore, the ID-Consistency loss $\mathcal{L}_{id}$ across the sequence of N frames can be calculated by:

$$\mathcal{L}_{id} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\phi(I)\cdot\phi(X_i^f)}{|\phi(I)||\phi(X_i^f)|}, \tag{4}$$

where $\cdot$ denotes the dot product, ϕ denotes the face recognition backbone[29], $\phi(X_i^f)$ and $\phi(I)$ represent the face embedding of each generated frame i and the reference identity image I, respectively.

3.3. Motion Control Enhancement

Although text prompt embedding shows control capacity to some degree, it is still insufficient to capture fine-grained motion dynamics. To address this challenge, we propose a spatial-aware motion control module with motion intensity to enhance the controllability. Besides, a region-aware loss is employed to enhance spatial coherence and realism in dynamic regions such as the face.

Motion Control Module

We regard the control capacity of the model as lying in two aspects: one is the faithfulness of the motion description, and the other is the magnitude of motion intensity. To achieve this goal, we introduce extra action phrase and motion intensity as the conditions in the proposed model.

We first extract the action phrase $\mathcal{A}$ from the original text prompt $\mathcal{P}$, for example, “talking” from “a man is talking”. We resort to MiniGPT[30] to implement this extraction process automatically. Then the action phrase is fed to CLIP text encoder[18] to obtain the action embedding E_M which captures the semantic intent of the motion.

Considering the magnitude of motion intensity is hard to define directly, we employ an optical flow estimator to extract the optical flow magnitude of the video as the motion intensity. Specifically, given a video clip $\mathcal{V}^{in} = \{v_i^{in}\}_{i=1}^{N}$, where N is the number of frames, we first extract the optical flow of each pixel between two adjacent frames by:

$$f_{i,(x,y)} = \Theta(v_i^{in}, v_{i+1}^{in}), \tag{5}$$

where (x, y) denotes the position of each pixel, and Θ is an optical flow estimation model. We use RAFT[31] as Θ for efficient and accurate optical flow estimation. Then the mean optical flow value τ_i can be calculated by simply averaging $f_{i,(x,y)}$. Afterward, we take τ_i as the threshold to produce binary mask $M_{i,(x,y)}$. Specifically, when the magnitude of the optical flow exceeds τ_i, set the corresponding position in $M_{i,(x,y)}$ to 1; otherwise set it to 0. Consequently, the mean foreground optical flow value $f_{i,fg}$ can be easily obtained by:

$$f_{i,fg} = \frac{1}{S} \sum_{x=1}^{H} \sum_{y=1}^{W} f_{i,fg}(x, y) = \frac{1}{S} \sum_{x=1}^{H} \sum_{y=1}^{W} M_{i,(x,y)} * f_{i,(x,y)}, \tag{6}$$

where $f_{i,fg}(x, y)$ is the foreground optical flow at each pixel (x, y). S denotes the number of the foreground pixels. The motion intensity $\mathcal{M}$ of the video is defined as follows:

$$\mathcal{M} = \frac{1}{N-1} \sum_{i=0}^{N-1} f_{i,fg}. \tag{7}$$

Subsequently, motion intensity $\mathcal{M}$ is projected through a multi-layer perceptron (MLP) to generate a motion embedding E_M aligned with the dimensionality of the action embedding E_A.

As illustrated in Fig. 2, two parallel cross attention modules (Cross Attn and Motion Attn) are adopted in the motion control module to insert the action embedding E_A and motion embedding E_M. The process is formally represented as follows:

$$Z'' = \text{Attn}(Q', K^a, V^a) + \alpha \cdot \text{Attn}(Q', K^m, V^m), \tag{8}$$

where $Q' = Z'W'_q$ is relevant with the output of ID-Preserving Module. K^a, V^a and K^m, V^m are the key-value pairs derived from the action embedding E_A and the motion embedding E_M, respectively. The parameter α balances the influence of motion intensity within the combined attention output Z''.

Method	Dover Score ↑	Motion Smoothness ↑	Dynamic Degree ↑	CLIP-I ↑	CLIP-T ↑	Face Similarity ↑
IPA-PlusFace[25]	0.797	0.985	0.325	0.587	0.218	0.480
IPA-FaceID-Portrait[25]	0.849	0.984	0.191	0.545	<u>0.223</u>	0.531
IPA-FaceID-PlusV2[25]	0.813	<u>0.987</u>	0.085	0.575	0.217	**0.617**
ID-Animator[16]	<u>0.857</u>	0.979	<u>0.433</u>	<u>0.607</u>	0.204	0.546
Ours	**0.869**	**0.998**	**0.449**	**0.633**	**0.227**	<u>0.609</u>

Table 1. Comparison of different methods across multiple metrics. Higher values indicate better performance, with the best scores in bold and the second best in underline. It is important to note that all methods were configured with an empty action phrase and a motion intensity setting of 20 for a more dynamic effect.

Region-Aware Loss

The fluency of the generated video heavily relies on the spatial coherence and realism of dynamic regions, e.g. the face areas. To achieve this goal, we apply a region-aware loss to force the model to focus more on the high-motion regions. Specifically, we normalize the foreground optical flow $f_{i,fg}(x,y)$ defined in Eq. (6) and calculate the optical flow mask $M_{i,\text{norm}}$:

$$M_{i,\text{norm}} = \text{clip}(f_{i,fg}(x,y)/255 + 0.5, 1.0, 1.5), \tag{9}$$

where $\text{clip}(\cdot, a, b)$ restricts the values into $[a, b]$. The high-motion areas will be assigned a greater value than the low-motion regions. Then the region-aware loss $\mathcal{L}_R$ across all N frames can be compactly defined as:

$$\mathcal{L}_R = \frac{1}{NH'W'} \sum_{i=1}^{N} \sum_{x=1}^{H'} \sum_{y=1}^{W'} M_{i,\text{norm}} \cdot [\epsilon_i(x,y) - \hat{\epsilon}_i(x,y)]^2, \tag{10}$$

where $\epsilon_i(x,y)$ and $\hat{\epsilon}_i(x,y)$ denote the target and predicted noise at location (x,y), respectively. H' and W' correspond to the resolution of latent.

3.4. Training Paradigm

Human-Motion Dataset

To support high-quality video generation, we constructed a diverse dataset named Human-Motion, comprising 106,292 video clips from various public and private sources. This collection includes VFHQ[32] (1,843 clips), CelebV-Text[17] (52,072 clips), CelebV-HQ[33] (31,004 clips), AAHQ[34] (17,619 clips), and a private dataset, Sing Videos (3,752 clips). Each clip in the Human-Motion dataset was rigorously filtered and re-annotated to ensure high-quality identity and motion information across diverse video formats, resolutions, and styles.

To enrich the dataset with motion-related information, we used MiniGPT[30] to automatically generate two types of captions for all videos: overall descriptions and action phrases. The overall descriptions provide a general summary of the video's content, while the action phrases offer specific annotations of facial and body movements present in the clips. These captions serve as the primary text description $\mathcal{P}$ and action phrase $\mathcal{A}$ in our framework.

Image-Video Training Strategy

To improve the model's generalization across different visual styles, we combined image and video data in training. While realistic videos effectively capture human portraits, they struggle with stylized and artistic content, such as anime. To bridge this gap, we incorporated around 17,619 styled portrait images as static 16-frame videos by replicating each image to simulate a motionless sequence with a motion intensity of 0. This approach addresses the challenge of generalizing to stylized portraits by expanding the model's exposure to a wider spectrum of visual characteristics, including variations in texture, color, and artistic exaggeration common in non-realistic styles. By training on both static styled images and dynamic realistic videos, the model learns to preserve identity traits across photorealistic and stylized visual representations, improving its ability to generalize across different visual styles.

Overall Objective

The total learning objective combines the Region-Aware Loss, which captures dynamic motion in high-activity regions, and the ID-Consistency Loss, which ensures identity consistency across frames. This dual objective guides the model to preserve both identity and motion fidelity in the generated videos. The total objective function, $\mathcal{L}_{\text{total}}$, is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_R + \beta \cdot \mathcal{L}_{id}, \tag{11}$$

where $\mathcal{L}_R$ guides the model to capture dynamic motion in high-activity regions, and $\mathcal{L}_{id}$ ensures identity consistency across frames. The hyperparameter β balances the influence of identity preservation against motion fidelity.

4. Experiments

4.1. Experiment Setup

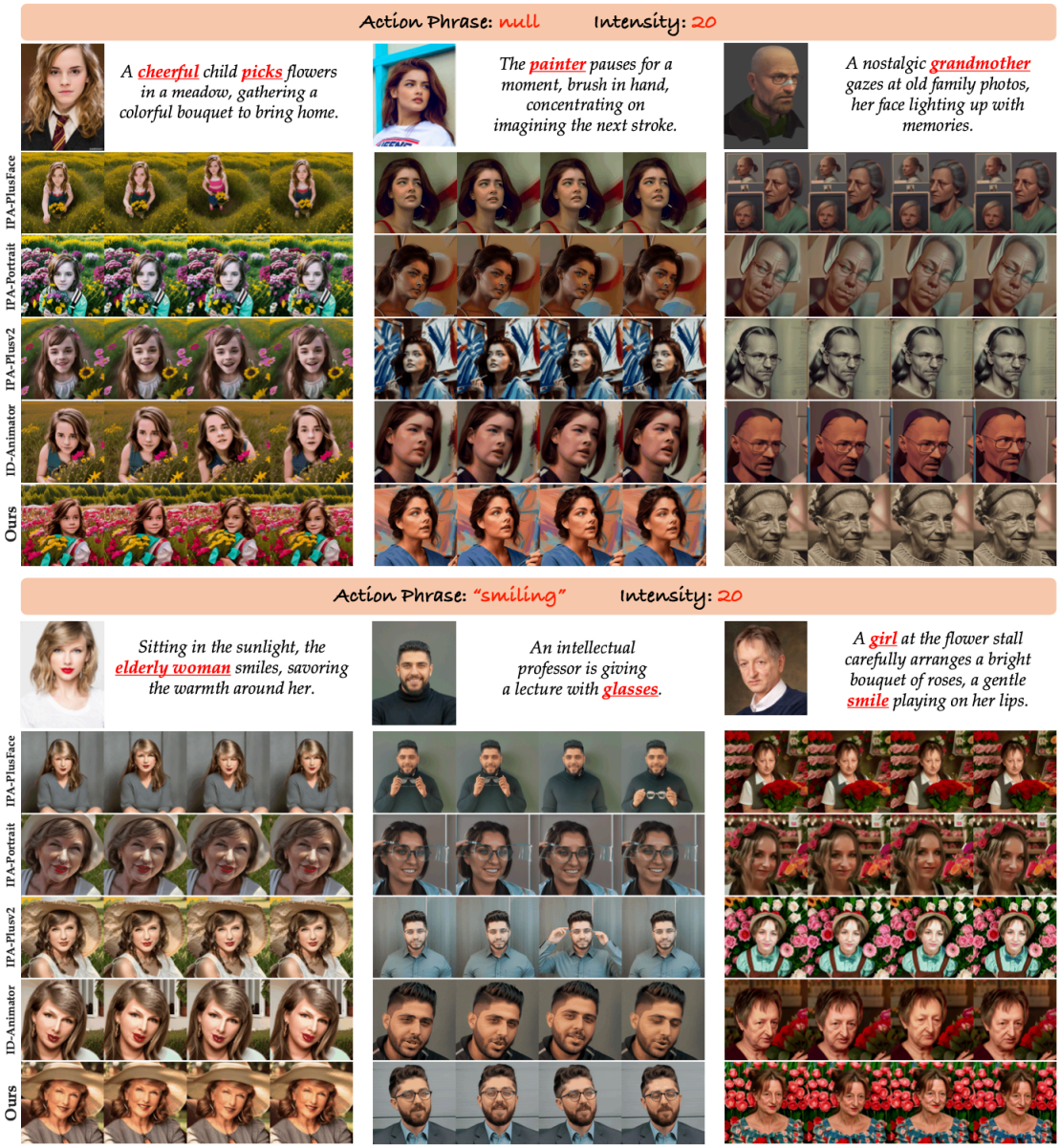

Figure 3. Qualitative Comparison. Comparison of our method with other approaches across diverse prompts and unseen reference images, encompassing various identities (male, female, celebrity, non-celebrity). Each column represents a unique identity and action phrase, with motion intensity fixed at 20 for clarity. "null" indicates a blank action phrase. Key prompt elements are highlighted in underline to emphasize specific actions or descriptors. For other methods, the action phrase and motion intensity are incorporated with the

doi.org/10.32388/S2EWVR

prompt to guide generation. To simplify notation, we abbreviated method names on the far left by omitting the common “FaceID” field, resulting in labels like IPA-Portrait.

Implementation Details

Our implementation is built upon a large-scale pre-trained Video Diffusion Model (VDM)[2]. All experiments were conducted using 8 NVIDIA A100 GPUs (80GB), with the training process taking approximately 24 hours. The batch size was set to 2 for each GPU. The training data consisted of diverse video clips, which were preprocessed to a resolution of 512×512 pixels, with 16 frames sampled per video at a frame rate of 4 frames per second. Data augmentation included random horizontal flipping, resizing, and center cropping to maintain consistent input dimensions. Moreover, text and image dropout rates were set to 0.05, with a 50% probability of dropping the CLIP embeddings E_{clip}. We used the AdamW[35] optimizer with a learning rate of 1×10^{-5} and trained the model for 12,000 steps. For the validation, 16-frame video sequences were generated at a 512×512 resolution, applying a guidance scale of 8.0 and 30 steps.

Datasets

For training, we constructed a dataset of 106,292 video clips from various sources. We detail the composition and annotation process in Sec. 3.4. For evaluation, we used the open-source Unsplash-50 test set[36], which includes 50 portrait images sourced from the Unsplash website. For each reference image, we generated 140 prompts using GPT-4[37], yielding a total of 7,000 prompt-image pairs for calculating evaluation metrics.

Evaluation Metrics

We assess the quality and consistency of generated videos using six key metrics. The Dover Score[38] assesses overall video quality, considering technical and aesthetic factors. Motion Smoothness[39] evaluates the continuity of movement between frames, while Dynamic Degree[39] indicates the extent of motion diversity in the video. CLIP-I[40][41] measures visual similarity to the reference using the CLIP encoder[18], and CLIP-T[40][41] evaluates the alignment between the video content and the text description. To assess identity preservation, we calculate Face Similarity[41], which measures the resemblance between the facial features in the reference image and the generated video.

4.2. Comparison with Baselines

We employ four well-known methods in ID-preserving generation task for comparison, i.e., IPA-PlusFace[25], IPA-FaceID-Portrait[25], IPA-FaceID-PlusV2[25] and ID-Animator[16]. They all adopt AnimateDiff[2] as the base text-to-video generation model.

Qualitative Comparisons

We choose six different individuals covering celebrities and common individuals, and produce corresponding text prompts using GPT-4[37] to present a fair comparison. As illustrated in Fig. 3, the upper videos are generated with "null" action phrase while the bottom videos are generated with "smiling". Obviously, our proposed *MotionCharacter* yields superior results in identity consistency and motion control ability. The frames generated by IPA-FaceID-Portrait are less similar to the reference image, and IPA-PlusFace and IPA-FaceID-PlusV2 cannot maintain the consistency in generated video frames. Compared to ID-Animator, our method shows better capacity to align the prompt and identity information, while ID-Animator fails to present "glasses" in the generated videos.

Quantitative Comparisons

We also perform quantitative comparison with these ID-persevering methods across key metrics, including Dover Score, Motion Smoothness, Dynamic Degree, CLIP-I, and CLIP-T. As the results shown in Table 1, IPA-PlusFace performs well in Motion Smoothness but shows limited motion diversity. While IPA-FaceID-PlusV2 achieves the highest Face Similarity score, it compromises motion diversity. ID-Animator shows limitations in alignment between the video content and the text description. In contrast, our method maintains high identity consistency while producing more dynamic motion.

User Study

We recognize that the CLIP score[40][41] may not consistently align with human perception[42][43]. To address this, we conducted a user study to compare the quality of videos generated by our model against baselines. Participants viewed clips generated with varying action phrases and intensity levels, rating the top method on identity consistency, motion controllability, and overall video quality. We then calculated the percentage distribution across all methods. As shown in Fig. 4, our method was consistently preferred by users. This feedback aligns with our quantitative results, demonstrating its superior balance of identity preservation, motion control, and visual fidelity compared to baselines.

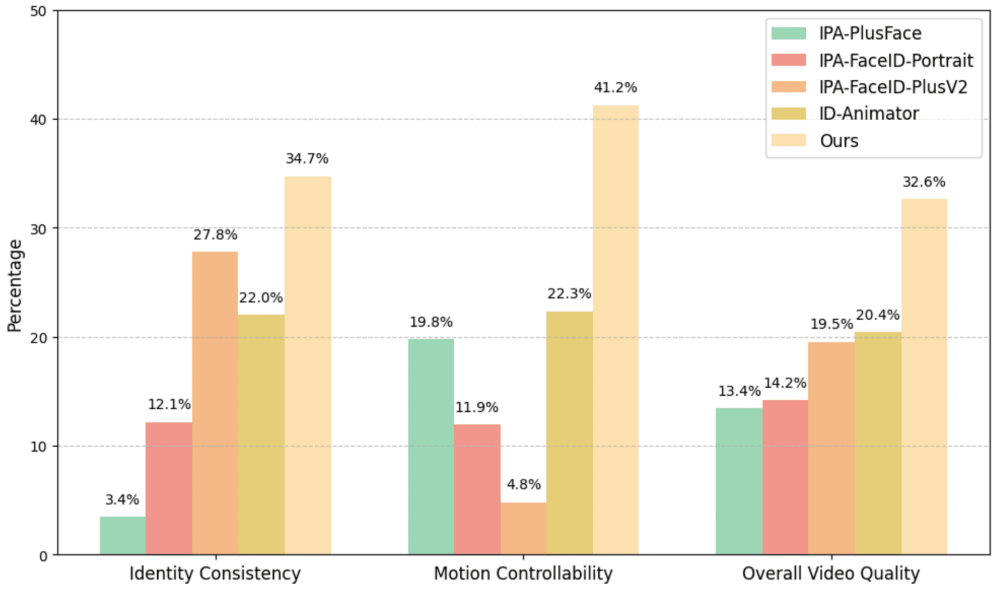

Figure 4. User study results comparing our method with baselines across three evaluation criteria: identity consistency, motion controllability, and overall video quality.

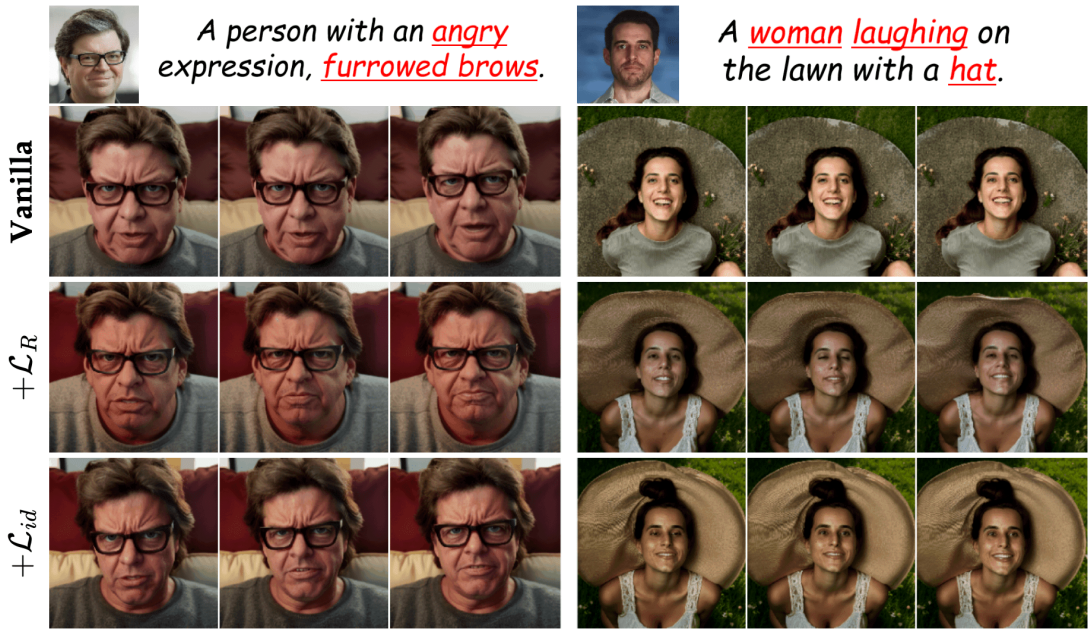

Figure 5. Ablation study on the effects of Region-Aware Loss $\mathcal{L}_R$ and ID-Consistency Loss $\mathcal{L}_{id}$.

doi.org/10.32388/S2EWVR

4.3. Ablation Study

Region-Aware Loss

Fig. 5 illustrates the impact of Region-Aware Loss in addressing motion blur and object distortion. In the ”Vanilla” model (top row) without Region-Aware Loss, high-motion regions like the lower corner of the glasses frame show noticeable distortion, compromising structural integrity. By adding Region-Aware Loss (middle row), the model better preserves structure in dynamic areas, resulting in clearer and more stable motion.

ID-Consistency Loss

The bottom row of Fig. 5 demonstrates the effect of adding ID-Consistency Loss on top of Region-Aware Loss. This combination significantly enhances the retention of identity-specific features, increasing facial similarity to the reference ID image. Furthermore, it better preserves facial detail characteristics (e.g., skin tone), resulting in improved overall facial integrity.

Figure 6. Ablation study on the effects of Motion Control Module.

Motion Control Module

To evaluate the effectiveness of the Motion Control Module (MCM), we conducted an ablation study, holding all training and inference parameters constant while isolating the MCM’s contribution. Specifically, we trained models both with and without the MCM, leaving all other modules unchanged. As

shown in Fig. 6, the model with the MCM generates videos that follow action prompts more precisely, with enhanced clarity and sharper motion details. This improvement demonstrates that the MCM significantly boosts the model's ability to capture nuanced action dynamics, achieving smoother transitions and more lifelike motion fidelity in the generated outputs.

5. Limitations

While our framework achieves significant performance in identity-consistent and controllable Text-to-Video (T2V) generation, it has limitations in handling highly complex or intricate motion sequences, where fine-grained motion dynamics may not be captured effectively. Additionally, the framework's performance is inherently dependent on the capabilities of the underlying T2V base model, which can limit the quality of generated videos. As T2V base models advance, our approach is designed with potential adaptability in mind; future iterations may leverage more powerful video foundation models, such as CogVideo-X[44], to enhance generalization and video fidelity in increasingly demanding scenarios.

6. Conclusions

In this paper, we propose a framework named *MotionCharacter* for human video generation that emphasizes identity consistency and precise motion control. We introduce the ID-Preserving Module, which ensures stable identity representation across frames, enhancing identity fidelity in the generated video. Additionally, we present the Motion Control Module, allowing nuanced adjustments of action phrases and motion intensity for fine-grained motion dynamics. To further improve model's performance, we leverage the region-aware loss, which reinforces fidelity in high-motion regions and achieves improved identity coherence. The Human-Motion dataset with detailed motion annotations enhances our model's adaptability to diverse prompts. Comprehensive evaluations confirm the effectiveness of our method in generating lifelike, personalized videos that accurately capture the specified actions and motion intensities.

Appendix A. Additional Dataset Analysis

To build the Human-Motion dataset, a multi-step pipeline was developed to ensure the collection of high-quality video clips. The detailed process is illustrated in Fig. I.

A.1. Video Sources

Our data sources comprise video clips from diverse origins, including VFHQ[32], CelebV-Text[17], CelebV-HQ[33], AAHQ[34], and a private dataset, Sing Videos. Each clip was carefully filtered and re-annotated to ensure high-quality identity and motion information across various formats, resolutions, and styles.

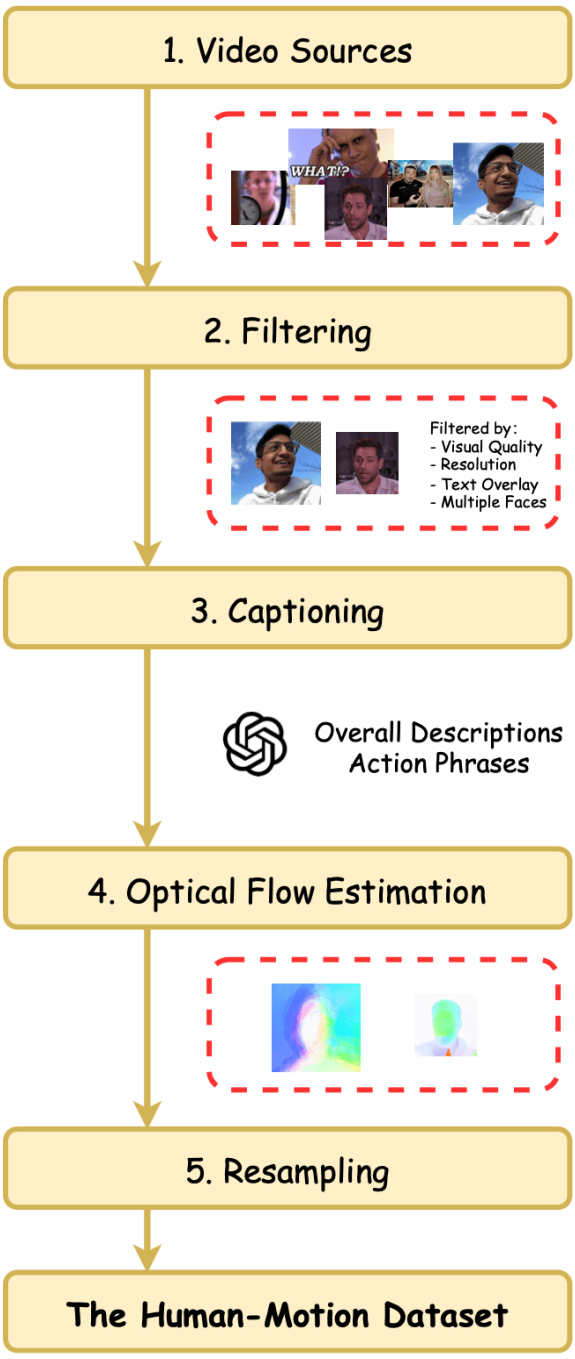

Figure I. The building process of the Human-Motion Dataset.

A.2. Filtering Process

To maintain data quality, a multi-step filtering process was applied:

doi.org/10.32388/S2EWVR

- **Visual Quality Check:** We used CLIP Image Quality Assessment[45] (CLIP-IQA) to evaluate visual quality by sampling one frame per clip, discarding videos with frames of low quality.
- **Resolution Filter:** Videos with resolutions below 512 pixels were removed to uphold visual standards.
- **Text Overlay Detection:** EasyOCR[46] was used to detect excessive subtitles or text overlays, filtering out obstructed frames.
- **Face Detection:** Videos containing multiple faces or low face detection confidence were discarded to ensure each video contains a single, clearly detectable person.

A.3. Captioning

To enrich motion-related data, we utilized MiniGPT[30] to automatically generate two types of captions for each video:

- **Overall Descriptions $\mathcal{P}$:** General summaries of the video content.
- **Action Phrases $\mathcal{A}$:** Detailed annotations of facial and body movements, serving as action phrases $\mathcal{A}$ in our framework.

This dual-captioning strategy enhances the dataset by providing both global context and specific motion dynamics, equipping the model to generate identity-consistent human video clips with controllable action phrases.

A.4. Optical Flow Estimation

Optical flow estimation for video is performed using the RAFT[31] model on consecutive frames to compute motion information. The RAFT model calculates the optical flow field, representing pixel displacements between frames. In this study, we use the extracted optical flow to obtain motion information for each video segment, enabling accurate motion modeling and control. These optical flows are used to compute motion intensity during the training phase and are also utilized in optimizing the loss function.

A.5. Motion Intensity Resampling

To balance the training dataset, we resampled the videos based on their motion intensity values, measured within a range of 0 to 20. Specifically, we adjusted the sampling to ensure that videos across different motion intensity levels are more evenly represented within this range, balancing the

distribution of videos across varying degrees of motion. For videos with motion intensity values exceeding 20 (which constitute a minority within the dataset), we capped their motion intensity at 20. This approach creates a more balanced distribution of motion intensity levels across the dataset.

A.6. The Human-Motion Dataset

The Human-Motion dataset consists of 106,292 video clips sourced from various datasets, including VFHQ[32] (1,843 clips), CelebV-Text[17] (52,072 clips), CelebV-HQ[33] (31,004 clips), AAHQ[34] (17,619 clips), and a private dataset, Human Videos (3,752 clips). Each clip was rigorously filtered and re-annotated to ensure high-quality identity and motion information across diverse formats, resolutions, and styles.

Notes

Project page: https://motioncharacter.github.io/

Haopeng Fang conducted this work during an internship at Meituan.

References

1. a, b *Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ (2022). "Video diffusion models". Advances in Neural Information Processing Systems. **35**: 8633–8646.*
2. a, b, c, d, e *Guo Y, Yang C, Rao A, Liang Z, Wang Y, Qiao Y, Agrawala M, Lin D, Dai B (2023). "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning". arXiv preprint arXiv:2307.04725. arXiv:2307.04725.*
3. a, b *Ho J, Chan W, Saharia C, Whang J, Gao R, Gritsenko A, Kingma DP, Poole B, Norouzi M, Fleet DJ, et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303. 2022.*
4. a, b *Singer U, Polyak A, Hayes T, Yin X, An J, Zhang S, Hu Q, Yang H, Ashual O, Gafni O, et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792. 2022.*
5. a, b *Blattmann A, Rombach R, Ling H, Dockhorn T, Kim SW, Fidler S, Kreis K (2023). "Align your latents: High-resolution video synthesis with latent diffusion models". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22563--22575.*
6. a, b *Wang J, Yuan H, Chen D, Zhang Y, Wang X, Zhang S (2023). "Modelscope text-to-video technical report". arXiv preprint arXiv:2308.06571.*

7. a, b *Wang Y, Chen X, Ma X, Zhou S, Huang Z, Wang Y, Yang C, He Y, Yu J, Yang P, et al. Lavie: High-quality vide o generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103. 2023.*
8. a, b *Zhou D, Wang W, Yan H, Lv W, Zhu Y, Feng J (2022). "Magicvideo: Efficient video generation with latent di ffusion models". arXiv preprint arXiv:2211.11018. arXiv:2211.11018.*
9. a, b *Chen H, Xia M, He Y, Zhang Y, Cun X, Yang S, Xing J, Liu Y, Chen Q, Wang X, et al. Videocrafter1: Open diffu sion models for high-quality video generation. arXiv preprint arXiv:2310.19512. 2023.*
10. a, b *Chen H, Zhang Y, Cun X, Xia M, Wang X, Weng C, Shan Y. "Videocrafter2: Overcoming data limitations for high-quality video diffusion models." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 7310–7320.*
11. ^ *Bao Y, Qiu D, Kang G, Zhang B, Jin B, Wang K, Yan P (2023). "LatentWarp: Consistent Diffusion Latents for Z ero-Shot Video-to-Video Translation". arXiv preprint arXiv:2311.00353. Available from: https://arxiv.org/abs/ 2311.00353.*
12. a, b *Jiang Y, Wu T, Yang S, Si C, Lin D, Qiao Y, Loy CC, Liu Z (2024). "Videobooth: Diffusion-based video genera tion with image prompts". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition. 2024: 6689--6700.*
13. a, b *Wei Y, Zhang S, Qing Z, Yuan H, Liu Z, Liu Y, Zhang Y, Zhou J, Shan H. Dreamvideo: Composing your drea m videos with customized subject and motion. In: Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition. 2024. p. 6537-6549.*
14. a, b *Ma Z, Zhou D, Yeh CH, Wang XS, Li X, Yang H, Dong Z, Keutzer K, Feng J (2024). "Magic-me: Identity-spec ific video customized diffusion". arXiv preprint arXiv:2402.09368. Available from: https://arxiv.org/abs/2402. 09368.*
15. a, b *Wu T, Zhang Y, Wang X, Zhou X, Zheng G, Qi Z, Shan Y, Li X (2024). "Customcrafter: Customized video gen eration with preserving motion and concept composition abilities". arXiv preprint arXiv:2408.13239.*
16. a, b, c, d, e *He X, Liu Q, Qian S, Wang X, Hu T, Cao K, Yan K, Zhou M, Zhang J (2024). "ID-Animator: Zero-Shot I dentity-Preserving Human Video Generation". arXiv preprint arXiv:2404.15275.*
17. a, b, c, d *Yu J, Zhu H, Jiang L, Loy CC, Cai W, Wu W (2023). "Celebv-text: A large-scale facial text-video dataset". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 14805–14814.*
18. a, b, c, d *Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et a l. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR; 2021. p. 8748-8763.*

19. ^Ho J, Jain A, Abbeel P (2020). "Denoising diffusion probabilistic models". Advances in neural information processing systems. **33**: 6840–6851.
20. ^Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B (2020). "Score-based generative modeling through stochastic differential equations". arXiv preprint arXiv:2011.13456.
21. ^Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: CVPR; 2022.
22. ^Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021). "Lora: Low-rank adaptation of large language models". arXiv preprint arXiv:2106.09685. 2021.
23. ^Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano AH, Chechik G, Cohen-Or D (2022). "An image is worth one word: Personalizing text-to-image generation using textual inversion". arXiv preprint arXiv:2208.01618. arXiv:2208.01618.
24. ^Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. p. 22500-22510.
25. a, b, c, d, e, f, g, hYe H, Zhang J, Liu S, Han X, Yang W (2023). "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models". arXiv preprint arXiv:2308.06721.
26. ^Li Z, Cao M, Wang X, Qi Z, Cheng MM, Shan Y. Photomaker: Customizing realistic human photos via stacked id embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 8640-8650.
27. a, bWang Q, Bai X, Wang H, Qin Z, Chen A, Li H, Tang X, Hu Y (2024). "Instantid: Zero-shot identity-preserving generation in seconds". arXiv preprint arXiv:2401.07519.
28. ^Guo Z, Wu Y, Chen Z, Chen L, He Q (2024). "PuLID: Pure and Lightning ID Customization via Contrastive Alignment". arXiv preprint arXiv:2404.16022.
29. a, bDeng J, Guo J, Xue N, Zafeiriou S (2019). "Arcface: Additive angular margin loss for deep face recognition". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pages 4690–4699.
30. a, b, cZhu D, Chen J, Shen X, Li X, Elhoseiny M (2023). "Minigpt-4: Enhancing vision-language understanding with advanced large language models". arXiv preprint arXiv:2304.10592.
31. a, bTeed Z, Deng J (2020). "Raft: Recurrent all-pairs field transforms for optical flow". In: Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part II 16. Springer; 2020. p. 402--419.

32. a, b, cXie L, Wang X, Zhang H, Dong C, Shan Y (2022). "Vfhq: A high-quality dataset and benchmark for video face super-resolution". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 657–666.

33. a, b, cZhu H, Wu W, Zhu W, Jiang L, Tang S, Zhang L, Liu Z, Loy CC (2022). "CelebV-HQ: A large-scale video facial attributes dataset". In: European conference on computer vision. Springer; 2022. p. 650-667.

34. a, b, cLiu M, Li Q, Qin Z, Zhang G, Wan P, Zheng W (2021). "Blendgan: Implicitly gan blending for arbitrary stylized face generation". Advances in Neural Information Processing Systems. **34**: 29710–29722.

35. ^Loshchilov I (2017). "Decoupled weight decay regularization". arXiv preprint arXiv:1711.05101.

36. ^Gal R, Lichter O, Richardson E, Patashnik O, Bermano AH, Chechik G, Cohen-Or D. Lcm-lookahead for encoder-based text-to-image personalization. arXiv preprint arXiv:2404.03620. 2024.

37. a, bAchiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774. 2023.

38. ^Wu H, Zhang E, Liao L, Chen C, Hou JH, Wang A, Sun WS, Yan Q, Lin W. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: International Conference on Computer Vision (ICCV); 2023.

39. a, bHuang Z, He Y, Yu J, Zhang F, Si C, Jiang Y, Zhang Y, Wu T, Jin Q, Chanpaisit N, Wang Y, Chen X, Wang L, Lin D, Qiao Y, Liu Z (2024). "VBench: Comprehensive benchmark suite for video generative models". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

40. a, b, cHessel J, Holtzman A, Forbes M, Le Bras R, Choi Y (2021). "CLIPScore: A Reference-free Evaluation Metric for Image Captioning". In: EMNLP.

41. a, b, c, dXiao G, Yin T, Freeman WT, Durand F, Han S (2024). "Fastcomposer: Tuning-free multi-subject image generation with localized attention". International Journal of Computer Vision. pages 1–20.

42. ^Molad E, Horwitz E, Valevski D, Acha AR, Matias Y, Pritch Y, Leviathan Y, Hoshen Y (2023). "Dreamix: Video diffusion models are general video editors". arXiv preprint arXiv:2302.01329. Available from: https://arxiv.org/abs/2302.01329.

43. ^Wang W, Jiang Y, Xie K, Liu Z, Chen H, Cao Y, Wang X, Shen C (2023). "Zero-shot video editing using off-the-shelf image diffusion models". arXiv preprint arXiv:2303.17599. arXiv:2303.17599.

44. ^Yang Z, Teng J, Zheng W, Ding M, Huang S, Xu J, Yang Y, Hong W, Zhang X, Feng G, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072. 2024.

45. ^Wang J, Chan KC, Loy CC (2023). "Exploring clip for assessing the look and feel of images". Proceedings of the AAAI Conference on Artificial Intelligence. **37** (2): 2555–2563.

46. ^JaidedAI. EasyOCR. 2024. Available from: https://github.com/JaidedAI/EasyOCR.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.