MotionCharacter: Fine-Grained Motion Controllable Human Video Generation

Haopeng Fang[1*], Di Qiu[2†], Binjie Mao[2], He Tang[1†]

[1]School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China
[2]Meituan, Beijing, China
haopengfang@hust.edu.cn, {qiudihk, binjiemao}@gmail.com, hetang@hust.edu.cn

Abstract

Recent advancements in personalized Text-to-Video (T2V) generation have made significant strides in synthesizing character-specific content. However, these methods face a critical limitation: the inability to perform fine-grained control over motion intensity. This limitation stems from an inherent entanglement of action semantics and their corresponding magnitudes within coarse textual descriptions, hindering the generation of nuanced human videos and limiting their applicability in scenarios demanding high precision, such as animating virtual avatars or synthesizing subtle micro-expressions. Furthermore, existing approaches often struggle to preserve high identity fidelity when other attributes are modified. To address these challenges, we introduce *MotionCharacter*, a framework for high-fidelity human video generation with precise motion control. At its core, *MotionCharacter* explicitly decouples motion into two independently controllable components: action type and motion intensity. This is achieved through two key technical contributions: (1) a Motion Control Module that leverages textual phrases to specify the action type and a quantifiable metric derived from optical flow to modulate its intensity, guided by a region-aware loss that localizes motion to relevant subject areas; and (2) an ID Content Insertion Module coupled with an ID-Consistency loss to ensure robust identity preservation during dynamic motions. To facilitate training for such fine-grained control, we also curate Human-Motion, a new large-scale dataset with detailed annotations for both motion and facial features. Extensive experiments demonstrate that *MotionCharacter* achieves substantial improvements over existing methods. Our framework excels in generating videos that are not only identity-consistent but also precisely adhere to specified motion types and intensities.

Project Page — https://motioncharacter.github.io/

Introduction

High-quality, personalized, and controllable human video generation has gained significant traction, with applications spanning social media, film production, virtual avatars, and personalized content creation. Recent pioneering advancements in text-driven video generation models (Ho et al. 2022b; Guo et al. 2024; Ho et al. 2022a; Wang et al. 2023a, 2024b; Zhou et al. 2022; Chen et al. 2024) have driven substantial progress in this field. In particular, subject-driven Text-to-Video (T2V) generation models (Jiang et al. 2024; Wei et al. 2024; Ma et al. 2024; Wu et al. 2024; He et al. 2024) have made strides in producing high-quality videos that faithfully depict specific individuals.

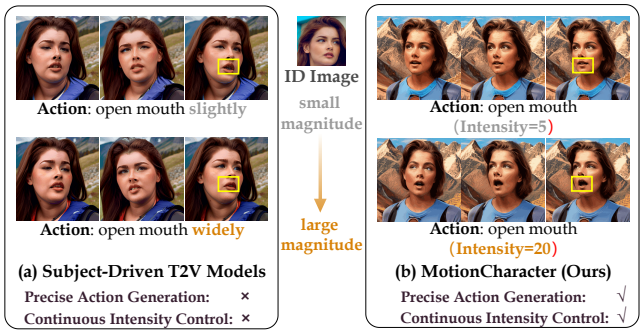

Figure 1: Comparison of subject-driven T2V methods and our proposed *MotionCharacter* framework. Existing approaches specify coarse actions (e.g., "open mouth") but struggle to capture nuanced motion magnitude (e.g., "slightly" vs. "widely"). In contrast, *MotionCharacter* decouples action type and motion intensity, enabling fine-grained, continuous control over human motion while preserving subject fidelity.

However, existing subject-driven T2V models still face a significant limitation: a lack of fine-grained control over motion. This deficiency curtails their applicability in scenarios demanding high precision, such as animating digital humans with precise expressions, synthesizing subtle micro-expressions for psychological analysis, or creating realistic virtual avatars that respond dynamically to user input.

The core of this problem lies in the entanglement of action semantics and motion intensity within textual prompts. As shown in Fig. 1(a), current subject-driven T2V models (He et al. 2024; Guo et al. 2024) allow users to specify human actions using action phrases like "open mouth". However, these phrases provide only a coarse description and fall short in precisely controlling motion intensity. Even when employing more nuanced phrases (e.g., "open mouth slightly",

*Work was done during an internship at Meituan.
†Corresponding authors.

"open mouth widely"), the generated results often lack the intended subtleties because language captures actions in a discrete manner, whereas motion intensity is inherently continuous. This forces the model to guess the intended magnitude, leading to unpredictable and uncontrollable results.

To break this entanglement and enable fine-grained controllability, we introduce *MotionCharacter*, a framework that explicitly decouples motion into two components: action type and motion intensity. To achieve this, we first propose a Motion Control Module. This module uses simple action phrases (e.g., "smile") to define the type of motion, while a separate, continuous signal derived from optical flow modulates its intensity. This allows a user to specify an action with text and then fine-tune its magnitude precisely, for instance, via a simple slider. To further enhance motion dynamics and the quality of facial transitions, this module is enhanced by a region-aware loss that directs the model's attention to critical facial regions like the lips and eyes, preventing distortion during motion.

Moreover, as motions become more dynamic and expressive, maintaining the subject's identity becomes a significant challenge. To address this, we introduce an ID Preservation Module that injects detailed facial features into the generation process. This is reinforced by a powerful ID consistency loss to ensure that the character's core identity remains stable and detailed, even during large and complex movements.

Finally, while several human video datasets (Zhu et al. 2022; Yu et al. 2023) exist, they predominantly focus on capturing basic emotional expressions or broad action categories, lacking the granular annotations needed for fine-grained motion control. Also, these datasets are limited in scale and contain unfiltered multi-subject and low-quality scenes. This limitation creates a significant bottleneck in training models capable of nuanced motion generation. To address this gap, we introduce Human-Motion, a large-scale dataset of 106,292 video clips. We use Large Multimodal Models (LMMs) and optical flow estimation to generate detailed annotations for motion type and intensity, aiming to facilitate research on fine-grained human video generation.

Through extensive experiments, we present qualitative, quantitative, and user study results that validate the performance of our method in terms of identity consistency and adherence to motion instructions. In summary, our contributions are as follows:

1. We propose *MotionCharacter*, a framework for personalized human video generation that decouples motion control into action type and intensity, enabling fine-grained and continuous motion adjustments.
2. We introduce a Motion Control Module that leverages text for action semantics and a continuous signal for intensity, enhanced by a region-aware loss to ensure high-quality dynamics. We also propose an ID Preservation Module with a dedicated ID-Consistency loss to maintain robust identity fidelity.
3. We curate Human-Motion, a large-scale and high-quality dataset of 106,292 video clips with detailed annotations for motion and identity, created specifically to facilitate research on controllable human video generation.

Related Work

Text-to-Video Diffusion Model. Recent advancements in diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021; Rombach et al. 2022) have established them as a leading approach in Text-to-Video generation. The Video Diffusion Model (Ho et al. 2022b) pioneered a space-time-factored U-Net for unconditional video generation in pixel space. Building on this, AnimateDiff (Guo et al. 2024) integrated a motion module into the Stable Diffusion framework to generate videos from text prompts. Later works, such as Imagen Video (Ho et al. 2022a) and Make-a-Video (Singer et al. 2022), adopted sequential models for pixel-space T2V generation. To overcome challenges with high-dimensional video data, latent diffusion models (Blattmann et al. 2023b) operate in the latent space of an auto-encoder, inspiring further developments including ModelScope (Wang et al. 2023a), LAVIE (Wang et al. 2024b), MagicVideo (Zhou et al. 2022), and VideoCrafter (Chen et al. 2023, 2024).

Subject-Driven Image and Video Generation. Subject-driven generation aims to synthesize content while preserving specific subject characteristics. In the image domain, recent works like IP-Adapter (Ye et al. 2023), PhotoMaker (Li et al. 2024), and InstantID (Wang et al. 2024a) have achieved efficient identity preservation through embedding-based approaches, eliminating the need for subject-specific training required by methods like DreamBooth (Ruiz et al. 2023). For video generation, works such as VideoBooth (Jiang et al. 2024), DreamVideo (Wei et al. 2024), and CustomCrafter (Wu et al. 2024) have explored learning-based frameworks to combine visual identity with motion dynamics. While ID-Animator (He et al. 2024) demonstrates zero-shot capabilities, it lacks fine-grained control over motion intensity. Our *MotionCharacter* addresses these limitations by enabling control of both appearance and motion without requiring retraining during inference.

Methodology

Overall Pipeline. 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 framework named *MotionCharacter* which accurately reflects identity information, captures fine-grained motion and maintains smooth visual transitions. Our framework is shown in Fig. 2. 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:

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

Our methodology is built upon the central principle of disentanglement, designed to establish an orthogonal control space for human video generation. We recognize that to robustly control motion, one must first guarantee the stability of identity, as intense dynamics often corrupt subject-specific features. Therefore, our framework is architected around two synergistic pillars that address this challenge sequentially. First, our ID-Preserving Optimization strategy establishes a stable identity foundation. Second, building upon this foundation, our Motion Control Enhancement

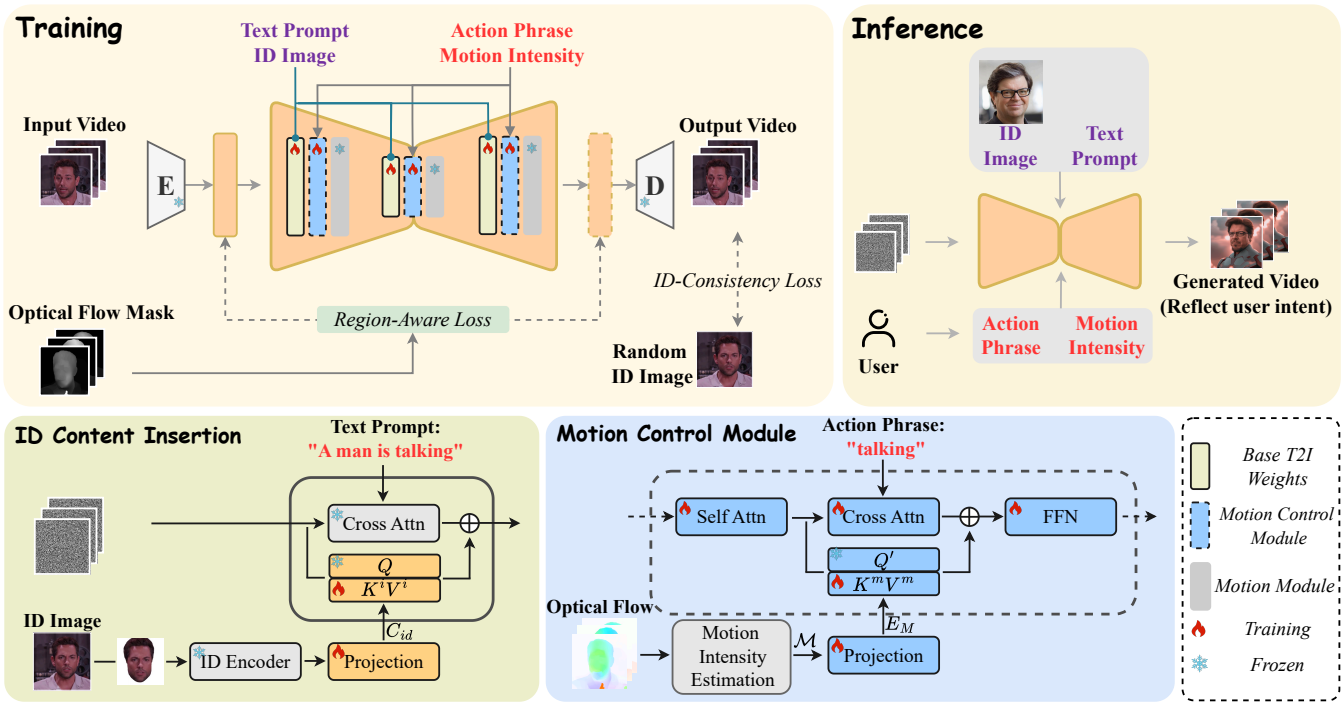

Figure 2: Framework overview. Our proposed framework comprises three core components: the ID Content Insertion Module, the Motion Control Module, and a composite loss function. The loss function incorporates a Region-Aware Loss to ensure high motion fidelity and an ID-Consistency Loss to maintain alignment 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.

module operates to independently manipulate action type and intensity. Furthermore, to enable robust training for such a disentangled framework, we introduce the Human-Motion dataset, specifically curated with dual-track annotations for motion semantics and intensity.

ID-Preserving Optimization

ID Content Insertion. Since the adopted pretrained Text-to-Video (T2V) diffusion model (Guo et al. 2024) lacks identity-preserving capabilities, we first intend to introduce an ID Content Insertion Module into the backbone to emphasize identity-specific regions and reduce irrelevant background interference. As illustrated in Fig. 2, the module 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 (Radford et al. 2021) and a face recognition model ArcFace (Deng et al. 2019) 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:

$$C_{id} = \text{Proj}(\text{Attn}(E_{arc}W'_q, EW'_k, EW'_v)), \quad (2)$$

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. This fusion allows the detailed ArcFace features to selectively attend to the most relevant contextual information from CLIP embeddings. 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 (Ye et al. 2023; Wang et al. 2024a), 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:

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

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 $K^i = C_{id}W^i_k$ and $V^i = C_{id}W^i_v$ 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. To enforce identity preservation at a semantic level, we introduce an ID-Consistency loss that complements the standard pixel-wise MSE objective. The MSE loss, while crucial for visual fidelity, is fundamentally

agnostic to high-level concepts like identity. Our proposed loss addresses this by penalizing deviations directly within a learned identity feature space. Specifically, it minimizes the feature-space distance between the reference identity and the generated frames, ensuring that the model maintains the subject's core characteristics even during complex motion.

In practice, at a specific diffusion step $\hat{t}$, the diffusion model can estimate the noise-free latent $\hat{z}_0$ from a noisy latent $z_{\hat{t}}$ via the 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 pre-trained face recognition backbone (Deng et al. 2019), $\phi(X_i^f)$ and $\phi(I)$ represent the normalized face embedding of each generated frame i and the reference identity image I, respectively. N is the total number of frames.

Motion Control Enhancement

In subject-driven T2V models, achieving precise motion control remains a significant challenge. While previous works (He et al. 2024) have successfully generated identity-specific videos, they often fail to capture fine-grained motion dynamics, resulting in limited responsiveness to subtle motion cues. To address this limitation, we propose a spatial-aware motion control module that explicitly incorporates motion intensity information to enhance controllability. Moreover, a region-aware loss is employed to enhance spatial coherence and realism in dynamic regions such as face.

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 use LMM (Zhu et al. 2023) to automatically generate overall descriptions and action phrases, enriching the dataset with motion-related information and serving as the primary text description $\mathcal{P}$ and action phrase $\mathcal{A}$, as defined in Eq. 1. Then the action phrase is fed to CLIP text encoder (Radford et al. 2021) to obtain the action embedding E_A which captures the semantic intent of the motion.

Motion Intensity Estimation. The magnitude of motion intensity is challenging to define directly, especially considering the diversity of movements and their varying characteristics. To address this, we employ optical flow estimation to capture motion intensity. Optical flow captures pixel-wise motion between adjacent frames, offering a fine-grained measure of movement that directly reflects the motion of objects in the video. 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 (Teed and Deng 2020) 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 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 then defined as the average foreground optical flow value across all frames:

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

where $\mathcal{M}$ is a scalar representing the motion intensity and N is the number of frames. Subsequently, the motion intensity $\mathcal{M}$ is passed through a motion intensity estimator, utilizing a multi-layer perceptron (MLP) to generate a motion intensity embedding E_M aligned with the dimensionality of the action embedding E_A.

Motion Condition Injection. As illustrated in Fig. 2, two parallel cross attention modules are adopted in the motion control module to insert the action embedding E_A and motion intensity 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 Content Insertion 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 intensity embedding E_M, respectively. The parameter α balances the influence of motion intensity within the combined attention output Z'' and is set to 1 by default.

Although both our approach and Stable Video Diffusion (SVD) (Blattmann et al. 2023a) leverage optical flow to characterize motion, a key distinction lies in our explicit decoupling of motion cues. SVD uses motion bucket parameter, which is integrated via additional time embeddings, to coarsely control overall motion magnitude. This results in a global and non-localized representation that may obscure subtle variations (e.g., faces). In contrast, our dual-branch motion condition injection strategy separates semantic guidance (what motion is desired) from quantitative strength (how intense it should be). The use of parallel cross-attention to fuse these decoupled cues is a key mechanism that enables predictable and fine-grained control, allowing our model to capture subtle motion nuances with high fidelity.

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 $M_{i,\text{norm}}$:

$$M_{i,\text{norm}} = \text{clip}\left(\frac{f_{i,fg}(x, y)}{255} + \delta,\ 1.0,\ 1.0 + \delta\right), \quad (9)$$

where $\text{clip}(\cdot,\ a,\ b)$ restricts the values to $[a, b]$ and δ is a scalar offset that modulates the spatial weighting of the optical flow mask. This mask assigns higher weights to regions with significant motion, ensuring that both the primary subject (e.g., faces) and high-dynamic background regions, especially when the text prompt describes a moving background, receive adequate attention. Note that although the optical flow mask assigns non-zero weights to some background regions, it ensures that regions with significant motion from both the subject and background are appropriately emphasized in the loss computation. In contrast to segmentation, depth maps and other similar static scene cues that may overlook such background dynamics, the optical flow mask retains comprehensive motion information across the scene. Then the region-aware loss $\mathcal{L}_R$ across all N frames can be compactly defined as:

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

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.

Training Paradigm

Our training paradigm is tailored for learning fine-grained, disentangled control. It comprises three integral parts: (1) a dataset with dual-track annotations for motion type and intensity, (2) a mixed image-video strategy for robust generalization and zero-motion calibration, and (3) a loss function that synergistically optimizes for both identity stability and motion fidelity.

Human-Motion Dataset. While foundational, current human video datasets like CelebV-HQ (Zhu et al. 2022) and CelebV-Text (Yu et al. 2023) lack the fine-grained motion annotations and quality controls needed for training controllable models. Their coarse emotional categories and data artifacts limit the learning of nuanced dynamics. To enable disentangled motion generation, we introduce Human-Motion, comprising 106,292 video clips from various public and private sources. This collection includes VFHQ (Xie et al. 2022) (1,843 clips), CelebV-Text (Yu et al. 2023) (52,072 clips), CelebV-HQ (Zhu et al. 2022) (31,004 clips), AAHQ (Liu et al. 2021) (17,619 clips), and a private dataset (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.

Human-Motion is distinguished by its dual-track annotation schema, which provides parallel labels for both motion semantics and intensity. Specifically, to enrich the dataset with motion-related information, we used LMM (Zhu et al. 2023) 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. It provides a crucial "zero-intensity calibration" for our motion control module. Training on these static sequences enables the model to learn a robust "motionless" state representation, which anchors the zero-point of the continuous motion intensity spectrum. This calibration contributes to generating smooth, predictable transitions from complete stillness to dynamic action. Moreover, 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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_R + \lambda_{id} \cdot \mathcal{L}_{id}, \quad (11)$$

where $\mathcal{L}_R$ and $\mathcal{L}_{id}$ synergistically guide the optimization. $\mathcal{L}_R$ ensures motion fidelity by focusing on dynamic regions, while $\mathcal{L}_{id}$ preserves identity consistency across the sequence. The hyperparameter λ_{id} mediates the trade-off between these two critical objectives.

Experiments

Experiment Setup

Implementation Details. Our implementation is built upon a large-scale pre-trained Video Diffusion Model (VDM) (Guo et al. 2024). 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 high-quality 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 (Loshchilov 2017) optimizer with a learning rate of 1×10^{-5} and trained the model for 12,000 total steps. For the validation, 16-frame video sequences were generated at a 512×512 resolution, applying a standard 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 the Training Paradigm subsection. For evaluation, we benchmark our method on the Unsplash-50 test set (Gal et al. 2024), a standard benchmark for ID fidelity containing 50 diverse portraits. To rigorously test performance, we paired each portrait with 140 distinct prompts generated via GPT-4 (Achiam et al. 2023), yielding a challenging evaluation suite of 7,000 pairs.

Evaluation Metrics. We assess the quality and consistency of generated videos using six key metrics. The Dover Score (Wu et al. 2023) assesses overall video quality, considering technical and aesthetic factors. Motion Smoothness (Huang et al. 2024) evaluates the continuity of movement between frames, while Dynamic Degree (Huang et al. 2024) indicates the extent of motion diversity in the video. CLIP-I (Hessel et al. 2021; Xiao et al. 2024) measures visual similarity to the reference using the CLIP encoder (Radford et al. 2021), and CLIP-T (Hessel et al. 2021; Xiao et al. 2024) evaluates the alignment between the video content and the text description. To assess identity preservation, we calculate Face Similarity (Xiao et al. 2024), which measures the resemblance between the facial features in the reference image and the generated video.

Comparison with Baselines

We employ four well-known methods in ID-preserving generation task for comparison, i.e., IPA-PlusFace (Ye et al. 2023), IPA-FaceID-Portrait (Ye et al. 2023), IPA-FaceID-PlusV2 (Ye et al. 2023) and ID-Animator (He et al. 2024). They all adopt AnimateDiff (Guo et al. 2024) as the base Text-to-Video generation model.

Qualitative Comparisons. We conduct qualitative comparisons using six diverse identities and prompts generated by GPT-4 (Achiam et al. 2023). As shown in Fig. 3, we test a static scenario (null action phrase) to assess baseline ID fidelity and a dynamic scenario ("smiling") to evaluate motion control. The visual results underscore the superiority of *MotionCharacter*. Methods like IPA-FaceID-Portrait exhibit noticeable identity degradation, while others like IPA-PlusFace and IPA-FaceID-PlusV2 suffer from temporal flickering during motion. This indicates their ID representations are not robust enough to withstand dynamic deformations. Furthermore, even the highly competitive ID-Animator fails to bind attributes correctly, omitting "glasses" from the generated character. This suggests a critical entanglement between its motion and appearance generation modules. In contrast, our method, powered by a dedicated ID-Preserving module and a disentangled motion architecture, successfully maintains identity and all specified attributes while executing the action.

Quantitative Comparisons. The quantitative results in Table 1 further substantiate our findings and reveal a critical challenge in existing methods: the inherent trade-off between identity preservation and motion dynamics. Specifically, IPA-FaceID-PlusV2 achieves the highest Face Similarity (0.617) but at the cost of producing nearly static videos, evidenced by its extremely low Dynamic Degree (0.085). Conversely, ID-Animator attains a high Dynamic Degree (0.433) but with significantly compromised identity fidelity and weaker text alignment (CLIP-T of 0.204), which confirms our qualitative observations. This demonstrates that current methods are forced to sacrifice either motion or identity. In contrast, *MotionCharacter* breaks this trade-off. Our method achieves state-of-the-art scores in five out of six metrics, including overall quality (Dover), motion attributes, and content consistency (CLIP-I, CLIP-T). Crucially, we attain a Face Similarity score (0.609) that is highly competitive with the top score, while simultaneously achieving the highest Dynamic Degree (0.449). This result is a powerful validation of our core hypothesis: by explicitly disentangling motion control from identity representation, our framework can generate highly dynamic and expressive videos without compromising identity fidelity. The negligible 1.3% difference in Face Similarity is a small price for a staggering 428% improvement in Dynamic Degree com-

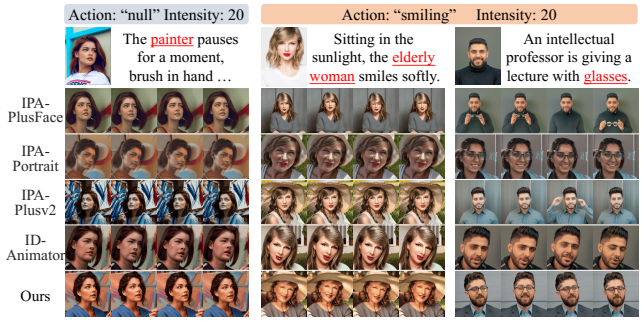

Figure 3: **Qualitative Comparison.** Comparison of our method with other approaches across diverse prompts and unseen reference images, encompassing various identities (male, female, celebrity, non-celebrity). Each column represents a unique identity and action phrase, with motion intensity fixed at 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 prompt to guide generation.

Method \ Metrics	Dover Score↑	Motion Smooth.↑	Dynamic Degree↑	CLIP Image↑	CLIP Text↑	Face Sim.↑
IPA-PlusFace	0.797	0.985	0.325	0.587	0.218	0.480
IPA-FaceID-Portrait	0.849	0.984	0.191	0.545	0.223	0.531
IPA-FaceID-PlusV2	0.813	0.987	0.085	0.575	0.217	**0.617**
ID-Animator	0.857	0.979	0.433	0.607	0.204	0.546
Ours	**0.869**	**0.998**	**0.449**	**0.633**	**0.227**	0.609

Table 1: Quantitative comparison with state-of-the-art methods. Higher values (↑) indicate better performance. **Bold** and underlined numbers denote the best and second-best results, respectively. All methods use an empty action phrase with motion intensity set to 20 for fair comparison.

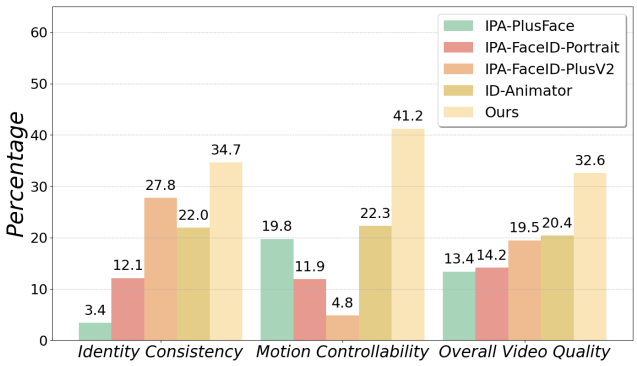

Figure 4: User study results comparing our method with baselines across three evaluation criteria: identity consistency, motion controllability, and overall video quality.

pared to the leading ID-preserving method.

User Study. Recognizing that CLIP scores (Hessel et al. 2021; Xiao et al. 2024) may not fully align with human perception (Molad et al. 2023; Wang et al. 2023b), we conducted a user study to validate our quantitative findings. The study involved 10 expert raters from enterprises and academic labs, each with expertise in generative models, evaluating 100 videos across three key dimensions: Identity Consistency, Motion Controllability, and Overall Video Quality. For each video, participants selected the best performing method among all approaches, with videos presented in randomized order to prevent bias. This resulted in 3,000 total ratings (10 raters $\times$ 100 videos $\times$ 3 criteria). As shown in Fig. 4, our method consistently received the highest preference across all evaluation criteria. Notably, the slight discrepancy between Face Similarity metrics and human perception of identity preservation can be attributed to varying levels of motion richness. While methods like ID-Animator and IPA achieve high similarity scores through strict identity constraints that limit facial dynamics, our approach generates more varied and expressive motions while maintaining perceptually consistent identity. This human evaluation validates our quantitative results, confirming that our decoupled control framework achieves a superior balance between identity preservation and motion expressiveness.

Ablation Study

Region-Aware Loss. As shown in Table 2, integrating only the Region-Aware Loss ($\mathcal{L}_R$) yields substantial improvements in motion-related metrics compared to the vanilla baseline. Specifically, it boosts the Dover Score by a significant +0.059 and increases the Dynamic Degree by +0.064, while also enhancing Motion Smoothness. This quantitatively confirms that by forcing the model to focus its objective on high-motion areas, $\mathcal{L}_R$ is highly effective at improving the clarity and quality of dynamic movements. Notably, its impact on Face Similarity is minimal, underscoring its specific role in refining motion rather than identity.

ID-Consistency Loss. The ID-Consistency Loss ($\mathcal{L}_{id}$) is designed to anchor the subject's identity. When added alone, it produces a massive gain of +0.104 in Face Similarity, proving its efficacy in robust identity preservation. However, the most compelling finding is the synergistic effect when both losses are combined. The full model not only achieves the best scores across all metrics but also surpasses the performance of individual components in their respective domains. For instance, the final Face Similarity (0.609) is higher than with $\mathcal{L}_{id}$ alone (0.588), and the Dynamic Degree (0.449) is greater than with $\mathcal{L}_R$ alone (0.419). This powerful synergy validates our core hypothesis: a stable identity foundation provided by $\mathcal{L}_{id}$ enables $\mathcal{L}_R$ to sculpt more expressive and high-fidelity motion, demonstrating that both losses are critical and complementary for achieving state-of-the-art results.

Loss		Metrics					
$\mathcal{L}_R$	$\mathcal{L}_{id}$	Dover Score↑	Motion Smoothness↑	Dynamic Degree↑	CLIP Image↑	CLIP Text↑	Face Similarity↑
✗	✗	0.801	0.978	0.355	0.599	0.219	0.484
✗	✓	0.810	0.983	0.359	0.627	0.220	0.588
✓	✗	0.860	0.995	0.419	0.611	0.224	0.500
✓	✓	**0.869**	**0.998**	**0.449**	**0.633**	**0.227**	**0.609**

Table 2: Ablation study of the Region-Aware Loss $\mathcal{L}_R$ and the ID-Consistency Loss $\mathcal{L}_{id}$.

Module	Metrics					
MCM	Dover Score↑	Motion Smoothness↑	Dynamic Degree↑	CLIP Image↑	CLIP Text↑	Face Similarity↑
✗	0.805	0.978	0.245	0.563	0.204	0.601
✓	**0.869**	**0.998**	**0.449**	**0.633**	**0.227**	**0.609**

Table 3: Ablation study of Motion Control Module (MCM).

Motion Control Module. Table 3 demonstrates the substantial impact of our Motion Control Module (MCM). Its introduction leads to consistent improvements across all metrics, with particularly notable gains in Dynamic Degree (83.3% increase from 0.245 to 0.449). This dramatic improvement in motion expressiveness, achieved while maintaining face similarity (0.609), validates our module's ability to enhance motion control without compromising identity preservation.

Conclusions

In this paper, we presented *MotionCharacter*, a high-fidelity human video generation framework that successfully addresses the fundamental challenge of achieving fine-grained motion control while ensuring strict identity preservation. The key to our approach is a disentangled architecture where two parallel mechanisms operate. The first mechanism anchors the subject's appearance using a dedicated ID Content Insertion Module supervised by an ID-Consistency Loss. The second enables precise control over action and intensity via the Motion Control Module, with a Region-Aware Loss ensuring clarity in dynamic areas. This entire framework is empowered by our Human-Motion dataset, whose fine-grained annotations are crucial for learning such disentangled representations. Exhaustive experiments validate that *MotionCharacter* not only outperforms existing methods in overall quality but also effectively breaks the critical trade-off between motion dynamics and identity fidelity.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant U25A20403), the Natural Science Foundation of Hubei Province of China (No. 2024AFB545). We also thank Meituan for their support.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7310–7320.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Gal, R.; Lichter, O.; Richardson, E.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2024. Lcm-lookahead for encoder-based text-to-image personalization. In *European Conference on Computer Vision*, 322–340. Springer.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *International Conference on Learning Representations*.

He, X.; Liu, Q.; Qian, S.; Wang, X.; Hu, T.; Cao, K.; Yan, K.; Zhou, M.; and Zhang, J. 2024. ID-Animator: Zero-Shot Identity-Preserving Human Video Generation. *arXiv preprint arXiv:2404.15275*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Jiang, Y.; Wu, T.; Yang, S.; Si, C.; Lin, D.; Qiao, Y.; Loy, C. C.; and Liu, Z. 2024. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6689–6700.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8640–8650.

Liu, M.; Li, Q.; Qin, Z.; Zhang, G.; Wan, P.; and Zheng, W. 2021. Blendgan: Implicitly gan blending for arbitrary stylized face generation. *Advances in Neural Information Processing Systems*, 34: 29710–29722.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, Z.; Zhou, D.; Yeh, C.-H.; Wang, X.-S.; Li, X.; Yang, H.; Dong, Z.; Keutzer, K.; and Feng, J. 2024. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*.

Molad, E.; Horwitz, E.; Valevski, D.; Acha, A. R.; Matias, Y.; Pritch, Y.; Leviathan, Y.; and Hoshen, Y. 2023. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Teed, Z.; and 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*, 402–419. Springer.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; Chen, A.; Li, H.; Tang, X.; and Hu, Y. 2024a. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Wang, W.; Jiang, Y.; Xie, K.; Liu, Z.; Chen, H.; Cao, Y.; Wang, X.; and Shen, C. 2023b. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2024b. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 1–20.

Wei, Y.; Zhang, S.; Qing, Z.; Yuan, H.; Liu, Z.; Liu, Y.; Zhang, Y.; Zhou, J.; and Shan, H. 2024. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6537–6549.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20144–20154.

Wu, T.; Zhang, Y.; Wang, X.; Zhou, X.; Zheng, G.; Qi, Z.; Shan, Y.; and Li, X. 2024. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *arXiv preprint arXiv:2408.13239*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2024. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 1–20.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 657–666.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, J.; Zhu, H.; Jiang, L.; Loy, C. C.; Cai, W.; and Wu, W. 2023. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14805–14814.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A large-scale video facial attributes dataset. In *European conference on computer vision*, 650–667. Springer.