Decomposing Text into Motion and Appearance for Training-Free Human Video Generation

Sungho Bae* Sanghwa Hong* Sangbum Lee* Boyeong Im*

Seoul National University

{sunghobae, hongsw5911, sblee99, boyng707}@snu.ac.kr

Abstract

Contemporary text-to-video systems often conflate appearance and motion cues, yielding videos that omit fine-grained visual details or exhibit unrealistic actions. To address this issue, we present a controllable text-to-human video generation framework that explicitly decomposes motion and appearance in the textual domain. Given a natural language prompt, a large language model first splits the sentence into a motion phrase and an appearance phrase. A text-to-motion (T2M) model converts the motion phrase into a 2D pose sequence, while a text-to-image (T2I) model renders a reference image that captures fine-grained visual attributes. These two modality-specific cues serve as modular conditions for a pre-trained image-to-video (I2V) generator, which synthesizes a temporally coherent video in which the reference appearance follows the motion trajectory.

This factorized pipeline offers three key advantages: ***(1) interpretability and fine-grained control****, as motion and appearance can be manipulated independently;* ***(2) architectural modularity****, providing plug-and-play compatibility with off-the-shelf T2M, T2I, or I2V backbone models without requiring retraining; and* ***(3) improved fidelity****, resulting in more detailed frames and more realistic motion than single-stream baselines. Extensive experiments show that our method outperforms strong alternatives in both qualitative studies and quantitative metrics (FVD, FaceSim-Arc, FC, and AX-NDCG). These results demonstrate the effectiveness of semantic decomposition in enabling controllable, high-fidelity human video synthesis without additional training.*

1. Introduction

Text-to-video (T2V) generation is an emerging task in generative AI that aims to synthesize temporally coherent video sequences from natural language descriptions. This task requires interpreting the semantics of the input text and translating it into a sequence of visual frames that accurately reflect the described content. Recognizing that videos are fundamentally composed of temporally coherent image sequences, early studies in T2V have explored the potential of diffusion-based architectures to generate high-fidelity videos from text prompts [23–25, 45, 54]. These early methods predominantly addressed simple object movements or static scene changes. In contrast, recent studies have focused on generating temporally coherent videos that reflect complex human behaviors and fine-grained appearance details [31, 34, 36, 51].

Despite notable progress in T2V generation [38, 44, 50], current models often struggle to faithfully reflect the full intent of user instructions. This is especially evident when the prompts contain both appearance (e.g., clothing, physical traits) and motion (e.g., running, jumping). For example, a prompt such as "a man in a red jacket is sprinting" requires an accurate depiction of both the outfit and dynamic human motion. However, existing models frequently focus on one aspect at the expense of the other, or fail to capture either correctly. As a result, the generated motions are often static or misaligned with the intended action, and fine-grained appearance details are frequently omitted.

To address these challenges, we propose an approach that explicitly disentangles motion and appearance in the textual modality, offering both interpretability and controllability over each factor in the video generation process. By decomposing textual instructions into appearance- and motion-specific components, our framework employs dedicated pre-trained modules to process each modality independently: a text-to-image (T2I) model synthesizes a reference image from the appearance prompt, while a text-to-motion (T2M) model generates a motion sequence from the motion prompt. These outputs are subsequently integrated as modular conditions into a pre-trained image-to-video (I2V) generator. This explicit factorization enables finer control over visual attributes and temporal dynamics, resulting in improved appearance fidelity and more realis-

*Equal contribution

tic human motion in the generated video. In addition, our framework enables the flexible reuse of existing pre-trained models, reducing training cost while maintaining high generation quality. This paradigm not only enhances the fidelity in T2V generation, but also opens up new possibilities for digital human creation and multimodal content editing.

Our key contributions can be summarized as follows:

- We propose a novel training-free text-to-human video generation pipeline based on semantic decomposition as shown in Figure 1, where an input text prompt is explicitly split into two distinct components: motion and appearance prompts. This separation enables a more accurate reflection of human characteristics by allowing the model to interpret and utilize motion and appearance attributes independently.
- The decomposition enhances controllability and interpretability by enabling the model to process spatial (appearance) and temporal (motion) features separately. This reduces interference between visual factors and facilitates the generation of semantically consistent and diverse human videos.
- Our framework adopts a modular two-stage pipeline that leverages pre-trained modules to extract motion and appearance representations independently. This design promotes both the efficiency and reusability of existing high-quality generative models.
- The decomposed motion and appearance features are then fused in an I2V module to synthesize coherent video sequences, ensuring that both dynamic motion and human appearance remain aligned with the original text prompt.

2. Related Works

2.1. Text to video generation

T2V generation has rapidly progressed with the emergence of diffusion-based generative models, which significantly outperform earlier GAN[21] or autoregressive methods in both visual fidelity and temporal coherence.

Early diffusion-based models such as CogVideo [25] and ModelScope-T2V [45] directly conditioned the generative process on CLIP-based text embeddings [37], employing spatial-temporal architectures to synthesize coherent short video clips. ModelScope-T2V, in particular, introduced a two-stage diffusion pipeline with text encoding and spatio-temporal attention modules, achieving high realism in short-duration sequences. Video-LDM [12] extended the latent diffusion framework to 3D video space by incorporating temporal convolution and cross-frame conditioning, which enabled scalable training on high-resolution videos. VideoCrafter2 [16] overcomes the lack of large-scale, high-quality video data by disentangling appearance and motion at the feature level to attain superior picture quality and temporal coherence without high-quality video supervision. These models exemplify a broader trend toward modularized frameworks, where pretrained components are recombined for efficiency and flexibility.

Despite these advancements, several limitations remain. Current models struggle with accurately capturing complex object interactions, long-term motion dynamics, and scene transitions, especially in crowded or ambiguous settings [30]. Even state-of-the-art systems like Sora [13] occasionally exhibit intermittent object disappearance and physically implausible motion, indicating a disconnect between the linguistic understanding embedded in large language models and the rendering capabilities of visual diffusion backbones. Efforts such as FlowZero [33], VideoDrafter [32], and SceneScape [20] aim to address these gaps by integrating LLM-guided prompt understanding, depth-aware generation, and dynamic scene decomposition. However, achieving physically consistent, semantically grounded, and temporally coherent video generation remains a key open challenge in the T2V domain.

2.2. Text to human video generation

Early work on text-conditioned human video synthesis mapped linguistic cues to explicit pose trajectories and then rendered photorealistic frames. SignSynth [40] introduces a two-stage gloss-to-pose and pose-to-video pipeline for sign-language production that learns continuous 3D signing motions from weakly labeled glosses and renders them with a GAN. Follow Your Pose [36] presents a pose-guided diffusion framework that extracts reference skeletons from human videos and fuses them with text embeddings in a diffusion backbone, producing anatomically plausible movements without manual pose annotation. Text2Performer [27] decomposes the VQ-VAE latent space into appearance and pose codebooks and introduces a continuous VQ-diffuser with motion-aware masking, thereby generating high-resolution human videos from text while preserving identity and producing temporally coherent motion.

In addition, HumanSD [28] fine-tunes stable diffusion with skeleton heat maps to strengthen pose adherence without catastrophic forgetting, while ID-Animator [22] injects a reference-image adapter into AnimateDiff to achieve zero-shot, identity-preserving animation from a single portrait. SignLLM [19] treats pose prediction as multilingual sequence generation, directly emitting 3D key points from text with a large transformer and a priority-learning channel. Despite these advances, current models still struggle with fine-grained appearance retention during large articulations, scalable high-quality motion synthesis, and physically consistent multi-person interactions. Bridging these gaps remains a central challenge for next-generation text-to-human video generators.

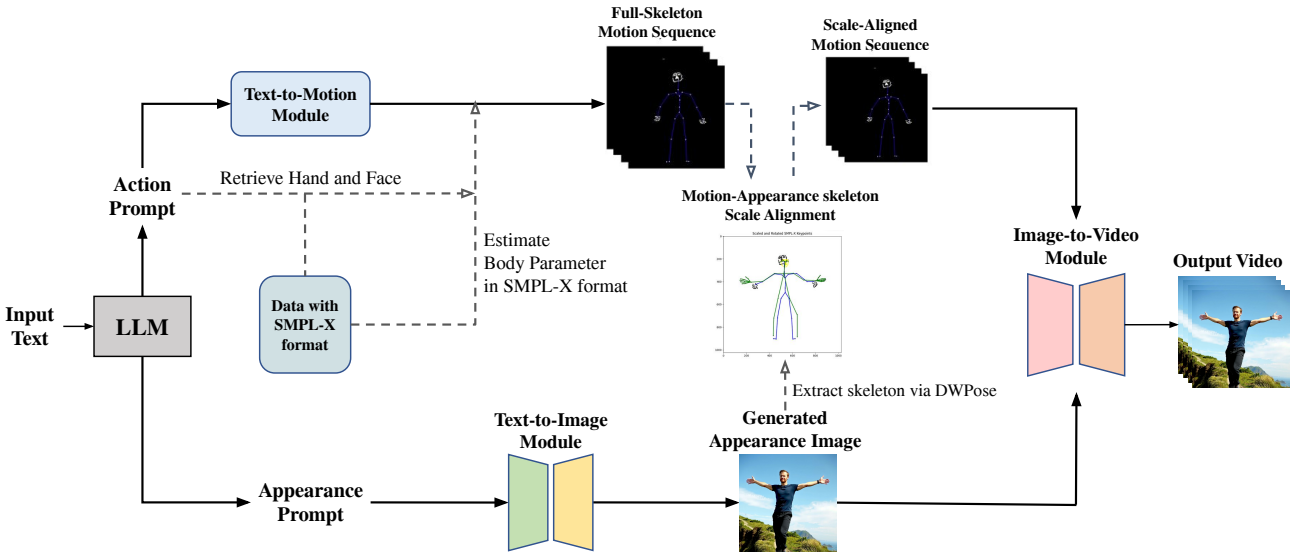

Figure 1. **Overview of our controllable human video generation framework.** Given an input text, a large language model (LLM) decomposes it into an appearance prompt and an action prompt. The appearance prompt is fed into a T2I module to synthesize a reference image, from which a pose skeleton is extracted via DWpose. The action prompt is used by a T2M module to generate a full-body motion sequence in SMPL-X format. To ensure visual consistency between motion and appearance, a scale alignment process adjusts the skeletons. Finally, the appearance image and aligned motion sequence are fused in an image-to-video module to produce the output video.

2.3. Pose to human video generation

Recent pose-conditioned generators extend latent diffusion backbones to improve identity fidelity and temporal range. DreamPose [29] adapts stable diffusion with a CLIP–VAE adapter plus subject-specific fine-tuning, yielding realistic cloth dynamics yet requiring per-identity retraining. Animate anyone [26] fused stable diffusion with ReferenceNet and a lightweight pose guider, enabling arbitrary character images to be animated into temporally consistent, high-fidelity videos without per-identity tuning. MagicAnimate [48] inflates a 2D diffuser into a spatio-temporal model with an appearance encoder and sliding-window fusion, while MagicPose [15] disentangles appearance and skeletal control through multi-source self-attention for zero-shot pose-and-expression retargeting.

A complementary target of recent research is the scalability and editability of the generated frames. MimicMotion [52] introduces confidence-aware pose maps and progressive latent fusion to stitch arbitrary-length action videos, whereas MotionFollower [41] frames motion editing as score-guided diffusion with two compact controllers, drastically reducing memory cost. UniAnimate [46] unifies reference image, pose stack, and noisy video in a single 3D UNet augmented by first-frame conditioning and state-space temporal layers, enabling minute-long, appearance-consistent synthesis. Unlike these pose-guided approaches, our method dispenses with reference skeletons altogether and generates accurate human videos from text alone.

3. Methodology

This section describes our method for generating a video aligned with a given text prompt $\mathbf{U}$. The pipeline consists of four main stages: (1) decomposition of the input prompt $\mathbf{U}$ using a large language model $\mathcal{L}$, (2) generation of a motion sequence $\{\mathbf{Q}_t\}_{t=1}^{T}$ based on the motion-related text $\mathbf{U}_{\text{mot}}$, (3) generation of a reference image $\mathbf{I}$ from the appearance-related text $\mathbf{U}_{\text{app}}$, and (4) final video generation $\mathbf{V}$ conditioned on the generated skeletons $\{\mathbf{K}_t\}_{t=1}^{T}$ and image $\mathbf{I}$.

3.1. Prompt decomposition using LLM

Given a natural language prompt $\mathbf{U}$ that describes a person or object performing an action (e.g., "a man in a red coat jumps and spins"), our method first decomposes the prompt into two distinct components: motion-related text $\mathbf{U}_{\text{mot}}$ and appearance-related text $\mathbf{U}_{\text{app}}$. This decomposition is performed using a large language model $\mathcal{L}$ that is prompted to identify and separate phrases corresponding to physical movement (e.g., "jumps and spins," mapped to $\mathbf{U}_{\text{mot}}$) and those corresponding to appearance or attributes (e.g., "a man in a red coat," mapped to $\mathbf{U}_{\text{app}}$). This structured decomposition enables specialized downstream components to focus on their respective generation tasks [10].

Table 1. Notation.

Symbol	Description
$\mathbf{U}$	Text prompt
$\mathcal{L}$	Large Language Model (LLM)
$\mathbf{U}_{\text{mot}}$	Motion-related text prompt
$\mathbf{U}_{\text{app}}$	Appearance-related text prompt
J	Number of body joints
T	Number of frames
$\mathcal{M}$	Text-to-motion model, $\mathcal{M} : \mathcal{T} \to (\mathbb{R}^{J\times 3})^T$
$\mathbf{Q}_t \in \mathbb{R}^{J\times 3}$	3D skeleton at frame t
$\mathcal{S}$	Parametric body model, e.g., SMPL-X
$\boldsymbol{\theta}_t$	Full-body SMPL-X parameter vector at frame t
$\boldsymbol{\theta}_{\text{body},t}$	Body-specific SMPL-X parameters at frame t
$\boldsymbol{\psi}_{\text{hand},t}$	Retrieved hand SMPL-X parameters at frame t
$\boldsymbol{\psi}_{\text{face},t}$	Retrieved face SMPL-X parameters at frame t
$\mathcal{E}_{\text{CLIP}}$	CLIP text encoder model
$\mathcal{D}_{\text{MX}}$	Motion-X reference dataset
Π	3D→2D projection operator
$\mathbf{K}_t \in \mathbb{R}^{J\times 2}$	2D skeleton at frame t
$\mathcal{I}$	Text-to-image model, $\mathcal{I} : \mathcal{T} \to \mathbb{R}^{H\times W\times 3}$
$\mathbf{I}$	Reference image
$\mathcal{V}$	Video generator
$\mathbf{V}$	Output video, $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^T$

$$\big(\mathbf{U}_{\text{mot}},\, \mathbf{U}_{\text{app}}\big) = \mathcal{L}(\mathbf{U}) \tag{1}$$

Instructions. We design our text instructions for the LLM with three components: task specification, supporting details, and a strict output format. We leverage the LLaMA-3.1-8B-Instruct model to perform sentence decomposition via in-context learning. The model receives a task instruction followed by several annotated examples. A full description of the instruction is presented in the Appendix A.

We retain the full original contents in the appearance prompt to provide the T2I module with rich contextual grounding. This helps prevent ambiguity, such as those arising from homonyms or underspecified entities. To ensure the generation of a full-body image, we enforce a fixed prompt format that begins with *"Full-body image of [original contents]"*.

In our preliminary experiments, we observed that the LLM, guided by a well-defined instruction and a few representative examples, reliably followed the constraints without hallucinating appearance-related content. However, when the appearance prompt was passed to the T2I module without details and additional contextual cues—such as motion or background—it occasionally led to unrealistic generations (e.g., humans appearing unclothed). To mitigate this, we chose to always include motion-related content (e.g., body posture or activity) alongside appearance descriptions and permitted the model to infer plausible and commonly expected attributes only when necessary, strictly based on the given context.

In-context Learning. Following the task instructions, we present the LLM with several in-context examples to reinforce the intended prompt structure and reduce ambiguity. We use the LLaMA-3.1-8B-Instruct model, which is specifically optimized for instruction-following tasks, to perform sentence decomposition via in-context learning. Its alignment with human-style prompts allows it to robustly follow task descriptions and adhere to structured output formats. The complete set of examples used during inference is also provided in the Appendix A.

By exposing the model to concrete demonstrations, we encourage it to generalize the expected formatting patterns and semantic cues necessary for the task. Prior works in T2V generation, including OpenSora[53] and VideoCrafter[16], have shown that in-context learning, when supported with sufficient contextual signals, can significantly enhance the quality of generation within DiT-based architectures[49].

3.2. Appearance image generation

Simultaneously, the appearance-related text $\mathbf{U}_{\text{app}}$ is fed into a T2I generation model $\mathcal{I}$ (e.g. Stable Diffusion, FLUX) to produce a single image $\mathbf{I}$ that visually represents the object or character described in the prompt [2, 6, 8]. The generated image preserves key visual cues including clothing, colors, and object identity. This image serves as a static visual reference for the appearance of the character or object throughout the video.

$$\mathbf{I} = \mathcal{I}\big(\mathbf{U}_{\text{app}}\big) \tag{2}$$

3.3. Motion sequence generation

The motion-related text $\mathbf{U}_{\text{mot}}$ is passed to a T2M model $\mathcal{M}$ (e.g. T2M-GPT or MotionDiffuse) that translates the description into a temporal sequence of 3D human skeletons $\{\mathbf{Q}_t\}_{t=1}^T$ [1, 3, 7]. Each skeleton frame consists of the 3D coordinates (X, Y, Z) of the J key body joints. For further processing, this 3D motion is projected into a 2D viewpoint space using an estimated affine transformation $\mathbf{P}$ described on subsection 3.6, yielding a time-series of 2D key-point skeletons $\{\mathbf{K}_t\}_{t=1}^T$ suitable for image-based rendering.

$$\{\mathbf{Q}_t\}_{t=1}^T = \mathcal{M}\big(\mathbf{U}_{\text{mot}}\big),\ \mathbf{K}_t = \mathbf{P}\big(\mathbf{Q}_t\big),\ t = 1, \ldots, T. \tag{3}$$

3.4. Hand and Face Motion Retrieval

The base T2M model $\mathbf{Q}^{Body}$ primarily generates expressive body motion, but it only produces coordinate values for the torso and major body joints, completely neglecting detailed, high-fidelity hand and facial articulation. To address this, we leverage a large-scale, high-quality motion dataset, *Motion-X*, which contains rich hand and face

SMPL-X parameters. Instead of generating these complex motions from scratch, we employ a retrieval-based strategy.

First, we encode the input motion description $\mathbf{U}_{\text{mot}}$ into a feature vector using a pre-trained CLIP text encoder. This vector is then compared against a pre-computed database of CLIP embeddings for all textual descriptions within the Motion-X dataset. We identify the Motion-X entry with the highest cosine similarity to our input prompt. From this best-matching data point, we retrieve its corresponding hand and facial SMPL-X parameter sequences, denoted as

$$\{\psi_{\text{hand},t}\}_{t=1}^{T} \quad \text{and} \quad \{\psi_{\text{face},t}\}_{t=1}^{T},$$

respectively. To ensure temporal synchronization with the generated body sequence of length T, these retrieved sequences are resampled using linear interpolation, and then used as the definitive hand and facial articulations in our final output.

3.5. Parametric Body Model Fitting for Coherent Merging

To create a cohesive full-body motion, we must merge the generated body skeleton sequence $\{\mathbf{Q}_t^{Body}\}_{t=1}^{T}$ with the retrieved hand and face parameters. A direct fusion is challenging, as the data representations are incompatible (i.e., absolute joint coordinates vs. parametric rotations). We considered converting the retrieved hand/face parameters into joint coordinates and attaching them to the body skeleton, but this often leads to anatomical inconsistencies (improper joint angles, disconnected limbs).

Therefore, we adopt a more robust strategy: convert the body skeleton sequence into the same parametric SMPL-X format as the retrieved data. This conversion is achieved via an optimization process known as Inverse Kinematics (IK) using the SMPL-X model [43]. We seek the body-specific parameters $\{\theta_{\text{body},t}\}_{t=1}^{T}$ (global orientation, body pose, translation) that best reconstruct the target skeleton:

$$\{\theta^*_{\text{body},t}\}_{t=1}^{T} = \underset{\{\theta_{\text{body},t}\}_{t=1}^{T}}{\operatorname{argmin}} \sum_{t=1}^{T} \left\| S(\theta_{\text{body},t})_{\text{joints}} - \mathbf{Q}_t \right\|_2^2, \tag{4}$$

where $S(\theta_{\text{body},t})_{\text{joints}}$ denotes the 3D joint locations produced by SMPL-X given parameters $\theta_{\text{body},t}$. Finally, to ensure a natural and fluid motion, a Gaussian filter is applied temporally across this sequence to mitigate high-frequency jitter.

3.6. Motion–image skeleton alignment

When the 3-D *motion skeleton* decoded from the motion prompt is not geometrically consistent with the 2-D *appearance skeleton* detected in the image, the video generator either hallucinates implausible backgrounds or fails to convey the intended action. To enforce consistency, we align the two skeletons by estimating an affine camera that maps each 3-D joint to its 2-D counterpart and then selecting the best-matching frame of the generated motion.

Let J be the number of body joints and T the number of generated 3D motion frames. For frame $t \in \{1, \dots, T\}$, we denote the homogeneous 3-D joint matrix by $\tilde{\mathbf{Q}}_t = [\tilde{\mathbf{q}}_{1,t}, \dots, \tilde{\mathbf{q}}_{J,t}]^\top \in \mathbb{R}^{J\times 4}$ with $\tilde{\mathbf{q}}_{j,t} = [X_{j,t}, Y_{j,t}, Z_{j,t}, 1]^\top$, and the 2-D appearance skeleton by $\mathbf{Y} = \{\mathbf{y}_j\}_{j=1}^{J} \in \mathbb{R}^{J\times 2}$ with $\mathbf{y}_j = [x_j, y_j]^\top$.

Affine matrix estimation via direct linear transform. For a given frame t we seek $\mathbf{P} \in \mathbb{R}^{2\times 4}$ such that $\mathbf{y}_j \approx \mathbf{P}\,\tilde{\mathbf{q}}_{j,t}$. Stacking the J correspondences yields

$$\underbrace{\begin{bmatrix} \tilde{\mathbf{q}}_{1,t}^\top & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{q}}_{1,t}^\top \\ \vdots & \vdots \\ \tilde{\mathbf{q}}_{J,t}^\top & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{q}}_{J,t}^\top \end{bmatrix}}_{\mathbf{D}\in\mathbb{R}^{2J\times 8}} \mathbf{p} = \underbrace{\begin{bmatrix} x_1 \\ y_1 \\ \vdots \\ x_J \\ y_J \end{bmatrix}}_{\mathbf{b}}, \tag{5}$$

where $\mathbf{p} = \text{vec}(\mathbf{P})$ is the row-major vectorization of $\mathbf{P}$. The reprojection loss is given as below:

$$\mathcal{L}(\mathbf{P}) = \sum_{j=1}^{J} \left\| \mathbf{P}\,\tilde{\mathbf{q}}_{j,t} - \mathbf{y}_j \right\|_2^2 = \|\mathbf{D}\mathbf{p} - \mathbf{b}\|_2^2. \tag{6}$$

Minimising (6) gives the ordinary least-squares solution $\mathbf{p}^\star = (\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{b}$, reshaped to $\mathbf{P}^\star = \text{reshape}(\mathbf{p}^\star, 2, 4)$.

Similarity decomposition. With the estimated affine matrix $\mathbf{P}^\star = [\,\mathbf{S} \mid \mathbf{t}\,]$ with $\mathbf{S} \in \mathbb{R}^{2\times 3}$, we factorize $\mathbf{S}$ into a single isotropic scale s, a row-orthonormal rotation $\mathbf{R} \in \text{SO}(2,3)$, and obtain

$$s = \tfrac{1}{2}\|\mathbf{S}\|_F, \quad \mathbf{R} = (\mathbf{S}/s)\left[(\mathbf{S}/s)(\mathbf{S}/s)^\top\right]^{-1/2}.$$

The 2-D prediction becomes $\hat{\mathbf{y}}_j = s\,\mathbf{R}\,\mathbf{X}_{j,t} + \mathbf{t}$, preserving the limb ratios of the human body.

Best-frame selection. Applying the above to every generated motion frame produces parameter sets $\{(s_t, \mathbf{R}_t, \mathbf{t}_t)\}_{t=1}^{T}$, we choose the best frame as the one with the smallest mean reprojection error with

$$t^\star = \arg\min_t \frac{1}{J} \sum_{j=1}^{J} \left\| \hat{\mathbf{y}}_{j,t} - \mathbf{y}_j \right\|_2, \tag{7}$$

where $\hat{\mathbf{y}}_{j,t} = s_t\mathbf{R}_t\mathbf{X}_{j,t} + \mathbf{t}_t$.

Because all operations—least-squares solve, polar decomposition, and error evaluation—are closed-form and differentiable, the alignment module provides an efficient and numerically stable alignment block.

3.7. Video generation with motion conditioning

In the final stage, we use a video generation model $\mathcal{V}$ that conditions on both the reference image $\mathbf{I}$ and the sequence of 2D skeletons $\{\mathbf{K}_t\}_{t=1}^{T}$. The model learns to generate a video $\mathbf{V}$ in which the appearance from the reference image is animated to follow the poses defined by the 2D skeleton sequence [4, 5, 9, 11]. The output is a temporally coherent and visually consistent video that aligns with both the motion and appearance semantics of the original text prompt.

$$\mathbf{V} = \mathcal{V}\Big(\mathbf{I},\, \{\mathbf{K}_t\}_{t=1}^{T}\Big) \tag{8}$$

4. Experiments

4.1. Setup

We use a modular zero-shot pipeline composed of four pretrained components: LLaMA- 3.1-8B-Instruct for prompt decomposition, MotionDiffuse and T2M-GPT for T2M, FLUX and Stable Diffusion XL (SDXL) for T2I, and Animate anyone[26] for I2V synthesis.

The Animate anyone model is used with 512×512 resolution and 30 sampling steps. The sequence length (`-L`) is dynamically set based on the generated motion skeleton, typically ranging from 32 to 128 frames. All experiments are conducted on 4 × NVIDIA RTX 3090 GPUs. $N = 50, C = 10$

4.2. Evaluation metrics

We evaluate the quality of the generated videos using three metrics: Fréchet Video Distance (FVD), CLIPScore, and Mean Per-Joint Position Error (MPJPE).

Fréchet Video Distance (FVD) [42]. FVD measures the distributional similarity between real and generated videos in a deep video feature space. Features are extracted using a pretrained I3D [14] network over full video sequences, and the Fréchet distance is computed as:

$$\mathrm{FVD} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}), \tag{9}$$

where $\boldsymbol{\mu}_r$, $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$ are the means and covariances of the real and generated video feature distributions. Lower FVD indicates better spatio-temporal coherence.

FaceSim-Arc [17]. FaceSim-Arc measures the cosine similarity between the face embeddings of a cropped reference image and those of a generated image. For video generation, we compute the similarity between each frame and the reference image, then average the scores across frames to obtain a video-level score. The final metric is the mean similarity across all videos.

$$\text{FaceSim-Arc} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\cos\left(f(x_{i,t}), f(r_i)\right)\right) \tag{10}$$

N denotes the number of videos, and T_i is the number of frames in the i-th generated video. $x_{i,t}$ is the t-th generated frame of the i-th video, and r_i is the corresponding reference image $f(\cdot)$ denotes the face embedding function (e.g., ArcFace), and $\cos(\cdot,\cdot)$ is the cosine similarity between two embedding vectors. A higher score indicates greater identity similarity between the generated face and the reference image, implying better identity preservation in the generated outputs.

Frame Consistency (FC) [18]. Frame Consistency (FC) evaluates the temporal stability of a generated video by measuring the semantic similarity between consecutive frames. This metric is designed to quantify artifacts such as flickering or unnatural content shifts. It is calculated by averaging the cosine similarity between the CLIP image embeddings of all adjacent frame pairs within a video sequence. The formula is as follows:

$$\mathrm{FC} = \frac{1}{T-1}\sum_{t=2}^{T}\cos\left(\mathbf{E}_\mathbf{I}(F_t), \mathbf{E}_\mathbf{I}(F_{t-1})\right), \tag{11}$$

where T is the total number of frames in the video, F_t is the frame at timestamp t, $\mathbf{E}_\mathbf{I}(\cdot)$ is the CLIP image embedding function, and $\cos(\cdot,\cdot)$ denotes the cosine similarity. The score ranges from -1 to 1. A value closer to 1 indicates that adjacent frames are highly similar in the embedding space, implying a temporally coherent video with smooth transitions.

AX-NDCG@k. AX-NDCG@k evaluates alignment quality between generated videos and text prompts using X-CLIP [35] embeddings and retrieval-based ranking [47]. Given N video-text pairs $\{(\mathbf{V_1}, \mathbf{U_1}), \ldots, (\mathbf{V_N}, \mathbf{U_n})\}$, we extract $\mathbf{E_V}, \mathbf{E_U} \in \mathbb{R}^{N\times d}$ using X-CLIP. Rows $\mathbf{e}_v^{(i)}$, $\mathbf{e}_u^{(j)}$ denote the i-th video/text embedding.

Step 1: Similarity Matrix.

$$\mathbf{S} = \mathbf{E}_V\mathbf{E}_U^\top, \quad S_{ij} = \langle \mathbf{e}_v^{(i)}, \mathbf{e}_u^{(j)} \rangle$$

Step 2: Ranking. For each V_i, sort $\{S_{ij}\}_{j=1}^{N}$ descending. Let rank_i be the rank of ground-truth $\mathbf{U}_i$.

Step 3: NDCG@k. Binary relevance:

$$\mathrm{rel}_{ij} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{DCG}_i@k = \begin{cases} \frac{1}{\log_2(\mathrm{rank}_i+1)} & \text{if } \mathrm{rank}_i \le k \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{NDCG}_i@k = \mathrm{DCG}_i@k$$

Step 4: Final Metric.

$$\text{XCLIPScore-NDCG@}k = \frac{1}{N}\sum_{i=1}^{N}\mathrm{NDCG}_i@k$$

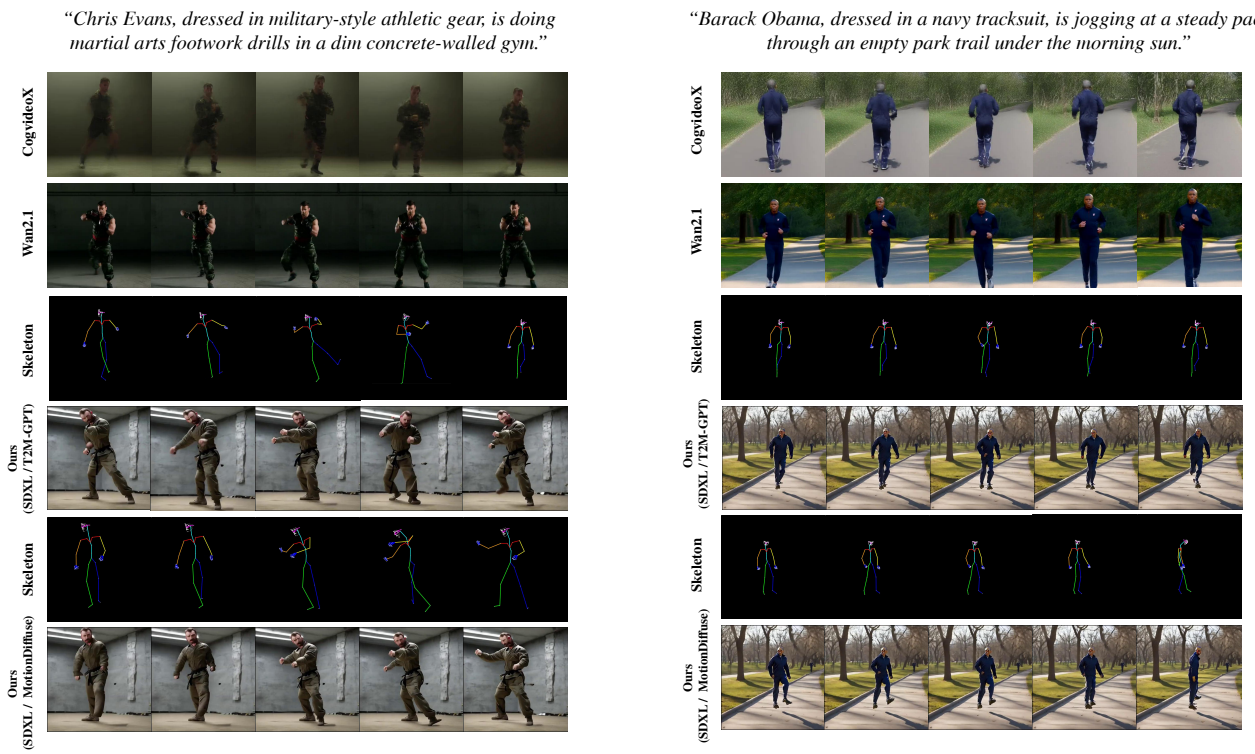

Figure 2. **Qualitative comparison across different combinations of T2M modules, and baseline models** For each prompt, we show results from CogVideoX and Wan2.1 baselines, followed by our method with two T2M variants (T2M-GPT and MotionDiffuse). The third and fifth row of the figure visualizes the 2D skeletons used as motion conditions. In this figure, reference appearance images are generated using SDXL, and motion sequences are produced by the corresponding T2M modules. Each video is synthesized by combining these two conditions via AnimateAnyone. For each model, five representative frames are sampled across the video, demonstrating consistency in identity and motion fidelity.

Step 5: Stochastic Averaging. X-CLIP samples fixed frames (e.g., 8/16), introducing variance. Repeat Steps 1–4 with C random seeds:

$$\text{AX-NDCG@}k = \frac{1}{C}\sum_{c=1}^{C}\text{XCLIPScore-NDCG@}k^{(c)}$$

AX-Hit@k. AX-Hit@k extends XCLIPScore-Hit@k with multiple seeds. Hit@k for each V_i:

$$\text{Hit}_i\text{@}k = \begin{cases} 1 & \text{if } \text{rank}_i \le k \\ 0 & \text{otherwise} \end{cases}$$

$$\text{XCLIPScore-Hit@}k = \frac{1}{N}\sum_{i=1}^{N}\text{Hit}_i\text{@}k$$

4.3. Qualitative results

Figure 2 presents a qualitative comparison across different combinations of T2M and T2I modules, evaluated on two example prompts. For each prompt, we show baseline results from CogVideoX and Wan2.1, followed by outputs from our method using two different T2M backbones: T2M-GPT and MotionDiffuse. For each generated video, we show five representative frames covering the temporal progression of the sequence.

The skeleton rows visualize the 2D pose sequences used as motion conditions. Compared to MotionDiffuse, T2M-GPT tends to produce more anatomically plausible and smoother motion, while MotionDiffuse better reflects the semantic meaning of the prompt. In terms of appearance, SDXL provides diverse and contextually rich reference images, which help preserve clothing style and facial consistency, though it occasionally suffers from spatial distortion in complex motion cases.

These visual comparisons highlight the complementary strengths of different T2M backbones and demonstrate the modularity of our generation pipeline.

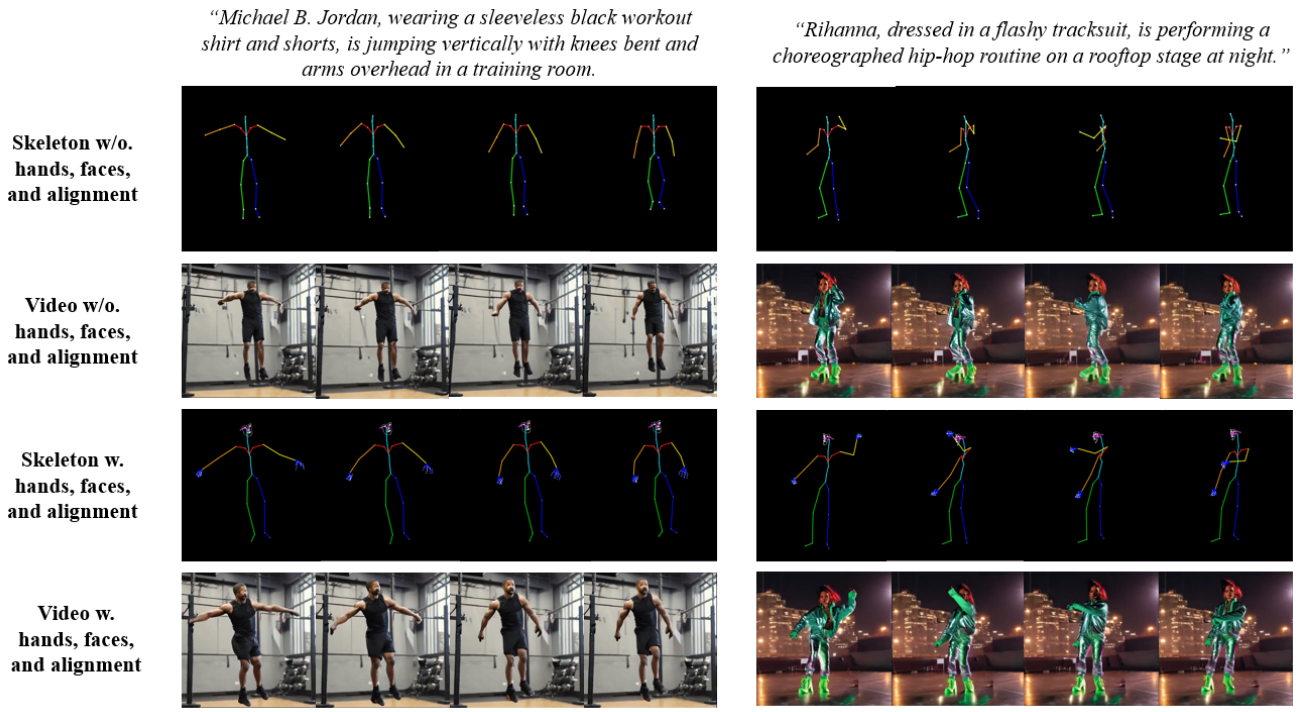

Figure 3. **Effect of skeleton alignment, face/hands modeling on video generation quality.** We compare generations with and without skeleton alignment and extended keypoints (hands and face). Each column corresponds to a sample video generated from the input text (shown at the top). Top two rows of each example show the skeletons and generated frames using a basic body-only skeleton without alignment. These results often exhibit misaligned motion and unnatural body proportions in video synthesis. In contrast, the bottom two rows show the results when applying our proposed alignment method and incorporating hands and face keypoints. We observe significantly improved motion naturalness, stable facial features, and better alignment with the reference image pose. This highlights the importance of precise skeleton alignment and full-body joint modeling for realistic text-to-video generation.

4.4. Quantitative results

For the quantitative evaluation, we employed the FVD metric using reference statistics computed from the UCF101 dataset [39], which comprises diverse human actions and varied background scenes. The I3D network [14] was used as the backbone feature extractor. Following the protocol used in the reference statistics, all generated videos were uniformly downsampled to a spatial resolution of 256×256 pixels and temporally cropped to fixed-length sequences of 16 frames sampled at equal intervals. In the case of FaceSim-Arc[17], every frame is extracted from each video, and facial embeddings are obtained using the pretrained ArcFace-based *buffalo-l* model from the InsightFace library. A reference embedding is similarly extracted from the corresponding reference image. Cosine similarity is then computed between the reference embedding and the embeddings of all frames in which a face is successfully detected. For each model, we report the final performance as the mean FaceSim-Arc score averaged over all evaluated video samples. This metric reflects how consistently the generated face resembles the identity in the reference image throughout the video.

Table 2. **Metric evaluation results** We cross-combine two text-to-motion and two text-to-image modules. Lower is better for FVD; higher is better for FaceSim-Arc and FC.

#	Methods		FVD↓	FaceSim-Arc↑	FC↑
-	CogVideoX		240.18	**0.2531**	0.9633
-	Wan2.1		**225.76**	0.0773	0.9682
A	MotionDiffuse	SDXL	<u>238.67</u>	0.1128	**0.9908**
B	MotionDiffuse	FLUX	281.89	0.0733	0.9896
C	T2M—GPT	SDXL	248.49	<u>0.1201</u>	<u>0.9898</u>
D	T2M—GPT	FLUX	281.89	0.0635	0.9878

As shown in Table 2, the proposed method demonstrates comparable or improved performance over state-of-the-art (SOTA) T2V models without requiring any additional training.

In terms of FVD, which measures the distributional similarity between generated and real videos (lower is better), the combination of MotionDiffuse and SDXL (Model A) achieves an FVD score of 238.67, outperforming CogVideoX (240.18) and closely approaching Wan2.1 (225.76). This indicates that the generated videos from the

Table 3. Comparison of AX-NDCG and AX-HIT metrics

#	Model	AX-NDCG@1	AX-NDCG@3	AX-NDCG@5	AX-NDCG@10	AX-HIT@1	AX-HIT@3	AX-HIT@5	AX-HIT@10
-	CogVideo	0.328	0.5544	0.6002	0.6445	0.328	0.708	**0.820**	**0.954**
-	WAN	0.322	0.4679	0.5210	0.5723	0.322	0.576	0.704	0.862
A	T2MGPT & SDXL	0.458	0.6074	**0.6500**	**0.6879**	0.458	**0.716**	0.818	0.934
B	T2MGPT & FLUX	0.304	0.4211	0.4772	0.5303	0.304	0.512	0.648	0.808
C	MotionDiffuse & SDXL	**0.464**	**0.6131**	0.6468	0.6859	**0.464**	0.714	0.796	0.914
D	MotionDiffuse & FLUX	0.322	0.4752	0.5240	0.5655	0.322	0.588	0.708	0.836

proposed configuration better approximate the real video distribution compared to the existing T2V models.

For the FaceSim-Arc metric, which evaluates the perceptual similarity of facial identity (higher is better), the combination of T2M–GPT and SDXL (Model C) shows the highest value of 0.1201, suggesting that the generated facial features are more consistent with real identities than those from CogVideoX (0.2531) or Wan2.1 (0.0773). Wan showed a lower FaceSim score, which was attributed to generating videos of individuals different from the reference image. Although CogVideo achieved a higher FaceSim score, this was often due to merely enlarging and repeating the reference image across frames, rather than generating genuinely high-quality or identity-consistent video content. These observations suggest that our method preserves the identity and appearance of the reference person at a level comparable to existing models.

In terms of Frame Consistency (FC), where higher scores indicate smoother and temporally coherent frame transitions, all the proposed combinations (Models A, B, C, and D) exhibit consistently high values (above 0.987), on par with or exceeding those of baseline models. This demonstrates that the proposed cross-modal composition strategy can maintain temporal coherence in the generated sequences even without fine-tuning.

Overall, these results validate that our proposed zero-shot combination framework is effective in generating high-quality, identity-preserving, and temporally consistent videos, outperforming or matching SOTA models across multiple evaluation metrics.

5. Conclusion

We have presented a *training-free*, modular pipeline that factorises a text prompt into **appearance** and **motion** streams, couples them with off-the-shelf T2I, T2M and I2V backbones, and enforces geometric consistency through an efficient 2-D/3-D skeleton–alignment block. Without any additional fine-tuning, the framework yields videos that (i) retain fine-grained identity cues, (ii) follow complex motion trajectories, and (iii) achieve competitive or superior scores on FVD, FaceSim-Arc and Frame-Consistency against strong T2V baselines. In addition, the method remains interpretable—each sub-module can be manipulated or replaced independently—highlighting the practical value of semantic decomposition for controllable human video generation.

Limitations. Although our training-free pipeline reaches state-of-the-art fidelity, it still assumes a global affine camera that fails under extreme viewpoints or self-occlusion, and it derives appearance from a single reference frame, causing subtle hand-/face details and identity consistency to fade in very long clips.

Future work. We will replace the 2-D image-to-video stage with a depth-aware diffusion backbone, introduce lightweight joint fine-tuning to better couple motion and appearance streams, and extend the system to multi-actor prompts while compressing the cascade for near–real-time authoring.

References

[1] Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 4

[2] High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[3] Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2023. 4

[4] Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, 2023. 6

[5] Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *ICCV*, 2023. 6

[6] Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 4

[7] T2m-gpt: Generating human motion from textual descriptions. In *CVPR*, 2023. 4

[8] Flux image generation models, 2024. 4

[9] Animatediff: Animate your personalized text-to-image diffusion models. In *ICLR*, 2024. 6

[10] Prompt decomposition: The missing piece to scaling generative ai, 2024. 3

[11] Dispose: Disentangling pose guidance for controllable human image animation. In *CVPR*, 2025. 6

[12] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis.

Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 2

[13] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6, 8

[15] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023. 3

[16] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2, 4

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6, 8

[18] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 6

[19] Sen Fang, Lei Wang, Ce Zheng, Yapeng Tian, and Chen Chen. Signllm: Sign languages production large language models. *arXiv preprint arXiv:2405.10718*, 2024. 2

[20] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36:39897–39914, 2023. 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[22] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 2

[23] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 1

[24] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[25] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1, 2

[26] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3, 6

[27] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22747–22757, 2023. 2

[28] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023. 2

[29] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690, 2023. 3

[30] Wentao Lei, Jinting Wang, Fengji Ma, Guanjie Huang, and Li Liu. A comprehensive survey on human video generation: Challenges, methods, and insights. *arXiv preprint arXiv:2407.08428*, 2024. 2

[31] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, pages 56–74. Springer, 2024. 1

[32] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *CoRR*, 2024. 2

[33] Yu Lu, Linchao Zhu, Hehe Fan, and Yi Yang. Flowzero: Zero-shot text-to-video synthesis with llm-driven dynamic scene syntax. *arXiv preprint arXiv:2311.15813*, 2023. 2

[34] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1430–1440, 2024. 1

[35] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647, 2022. 6

[36] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 1, 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 8

[40] Stephanie Stoll, Simon Hadfield, and Richard Bowden. Signsynth: Data-driven sign language video generation. In *European Conference on Computer Vision*, pages 353–370. Springer, 2020. 2

[41] Shuyuan Tu, Qi Dai, Zihao Zhang, Sicheng Xie, Zhi-Qi Cheng, Chong Luo, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motionfollower: Editing video motion via lightweight score-guided diffusion. *arXiv preprint arXiv:2405.20325*, 2024. 3

[42] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges, 2018. 6

[43] Vikram Voleti, Boris N. Oreshkin, Florent Bocquelet, Félix G. Harvey, Louis-Simon Ménard, and Christopher Pal. Smpl-ik: Learned morphology-aware inverse kinematics for ai driven artistic workflows, 2022. 5

[44] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

[45] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 2

[46] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024. 3

[47] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR, 2013. 6

[48] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3

[49] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18826–18836, 2025. 4

[50] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[51] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 1

[52] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 3

[53] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 4

[54] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1

Appendix

A. Full Prompt used in Decomposition

1. **Task specification:**

> *You are given a descriptive sentence about a person performing an action. The given prompt describes a person who is performing a certain action or activity, potentially with additional details about their appearance or surrounding environment. Your task is to split this sentence into exactly two parts:* **["[Original description with certain format]", "[Motion or action description]"]** *Do not add or assume anything that is not present in the original sentence.*

2. **Supporting details and format:**

> **APPEARANCE / BACKGROUND PROMPT**
> *- Use this fixed output format:* ***"Full-body image of*** **[original contents]"**
> *- Copy all information from the original sentence. But, if the original prompt lacks information about Appearance or Background, you can infer some common attributes based on the given context, like "Example 4" and "Example 5" below.*
>
> **MOTION / ACTION PROMPT**
> *- Extract only the part of the sentence describing movement, physical activity, or body dynamics.*
> *- Start with the first motion-related verb (e.g. jumps, runs, twirls, kicks).*
> *- Include all motion-related details: body posture, gesture, orientation, dynamics, etc.*

3. **In-context Examples:**

> **EXAMPLES**
> **Example 1:**
> Input: *Barack Obama, wearing a navy sleeveless basketball jersey and black shorts, is shooting a basketball into the hoop on an outdoor court with city buildings in the background under clear skies*.
> Output:
> *"Full-body image of Barack Obama, wearing a navy sleeveless basketball jersey and black shorts, shooting a basketball into the hoop on an outdoor court with city buildings in the background under clear skies.", "Shooting a basketball into the hoop."*
> **Example 2:**
> Input: *Elon Musk, in black boxing shorts and red gloves, is throwing a straight punch inside a gym-style boxing ring with ropes and overhead lighting.*
> Output:
> *"Full-body image of Elon Musk, in black boxing shorts and red gloves, throwing a straight punch inside a gym-style boxing ring with ropes and overhead lighting.", "Throwing a straight punch."*
> **Example 3:**
> Input: *A young woman in a gray sports bra and black leggings is practicing kickboxing inside a dimly lit gym.*
> Output:
> *"Full-body image of a young woman in a gray sports bra and black leggings practicing kickboxing inside a dimly lit gym.", "Practicing kickboxing."*
> **Example 4:**
> Input: *A teenage boy is shooting a basketball into the hoop.*
> Output:
> *"Full-body image of a teenage boy wearing a sleeveless basketball jersey and shorts, shooting a basketball into the hoop on an outdoor court", "Shooting a basketball into the hoop."*
> **Example 5:**
> Input: *A man is running.*
> Output:
> *"Full-body image of a man in athletic clothing, running on a jogging trail.", "Running."*