MoSA: Motion-Coherent Human Video Generation via Structure-Appearance Decoupling

Motivation

- Existing video generation models often overemphasize the fidelity of appearance while neglecting the rationality of motion and structural coherence.
- Consequently, their ability to generate complex human movements (such as long-distance dynamic motion, and fine-grained human-environment interactions) is limited, resulting in motion outcomes that are structurally inconsistent.

Motivation

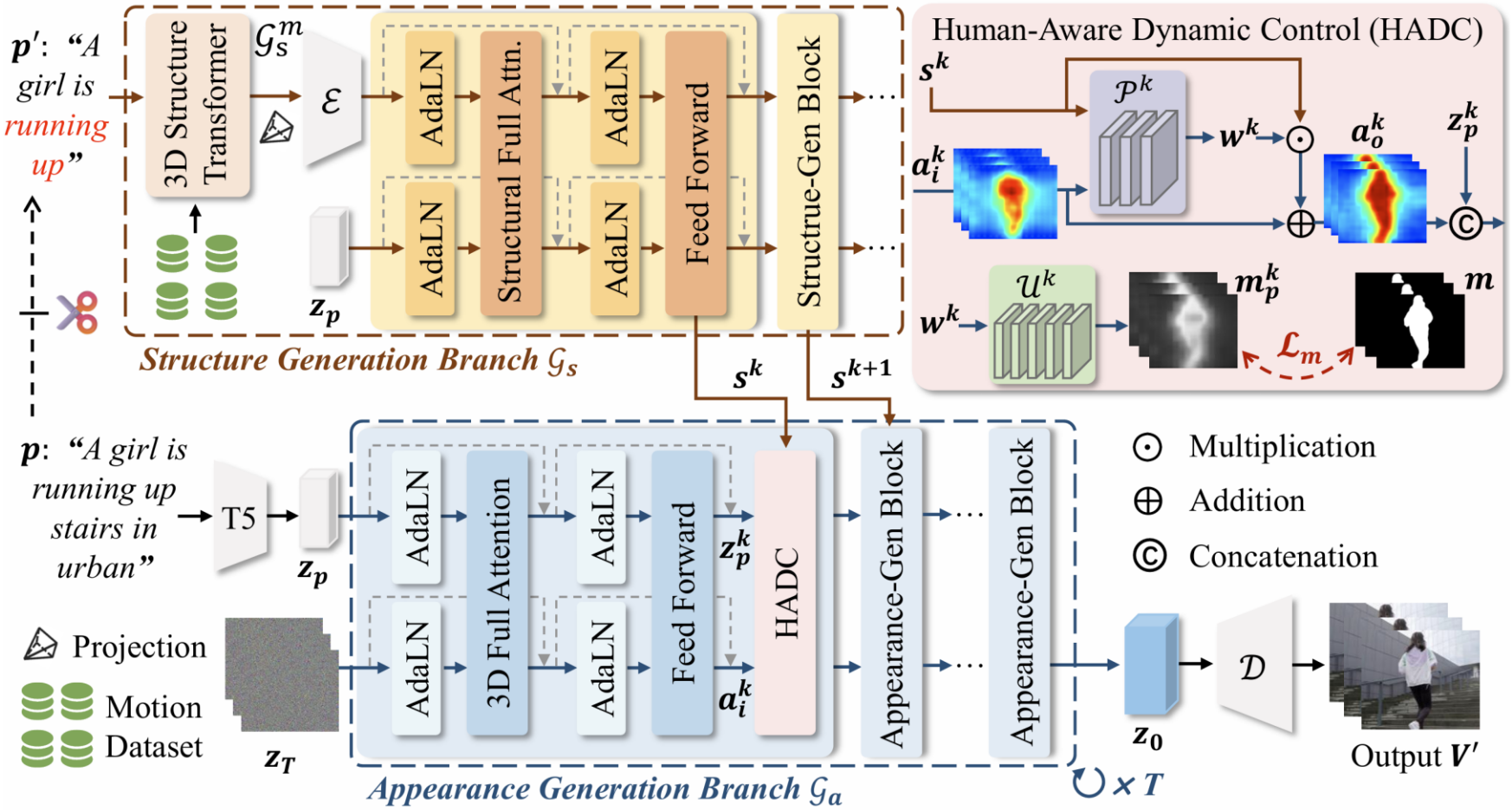
- Most existing human video datasets focus on relatively simple facial or half-body movements, while existing dance datasets are lacking in terms of background diversity and motion complexity.
- These datasets, which primarily focus on simple actions, make it difficult for models trained on such datasets to generate realistic and physically plausible actions.

Motivation

- To address the shortcomings of existing generative models in generating complex human motions, this approach proposes a structure-appearance decoupled generation framework and constructs a novel dataset, MoVid, containing 30K human motion videos along with corresponding text annotations, human skeleton annotations, etc., covering multiple motion types and exhibiting higher motion complexity.

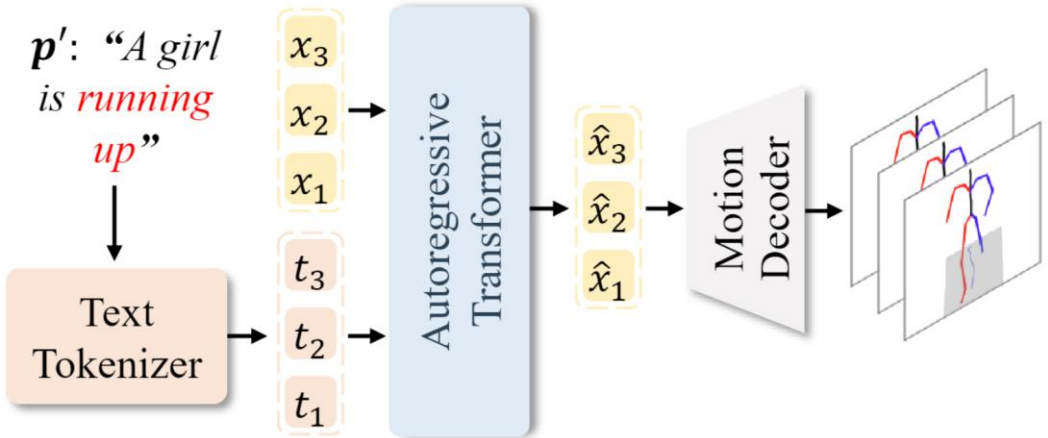
Motivation

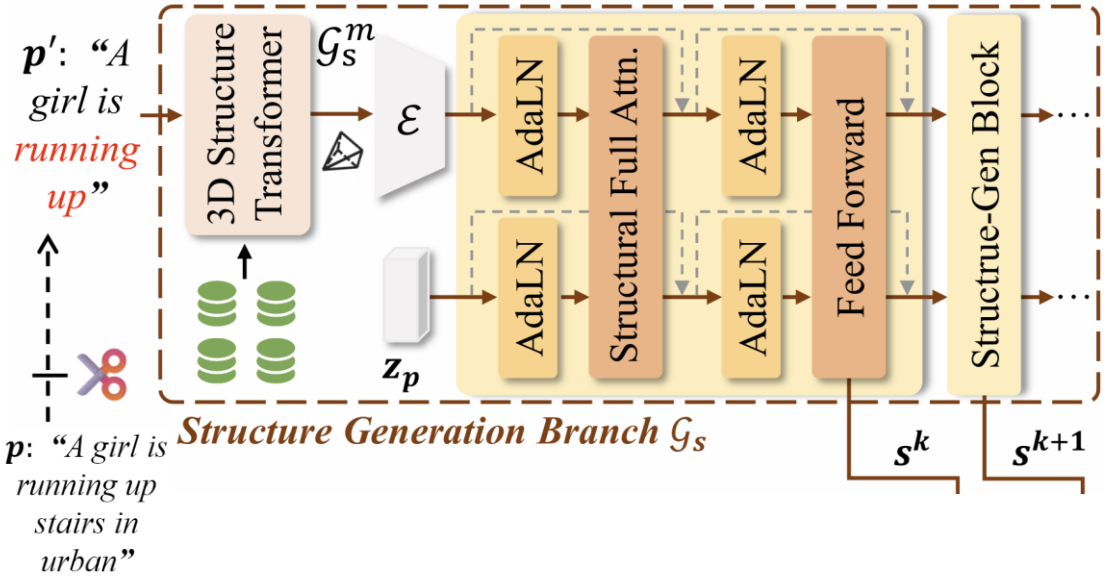
- This research approach decomposes the video generation process into two branches: a structure generation branch and an appearance generation branch, as shown in the figure below.



Method

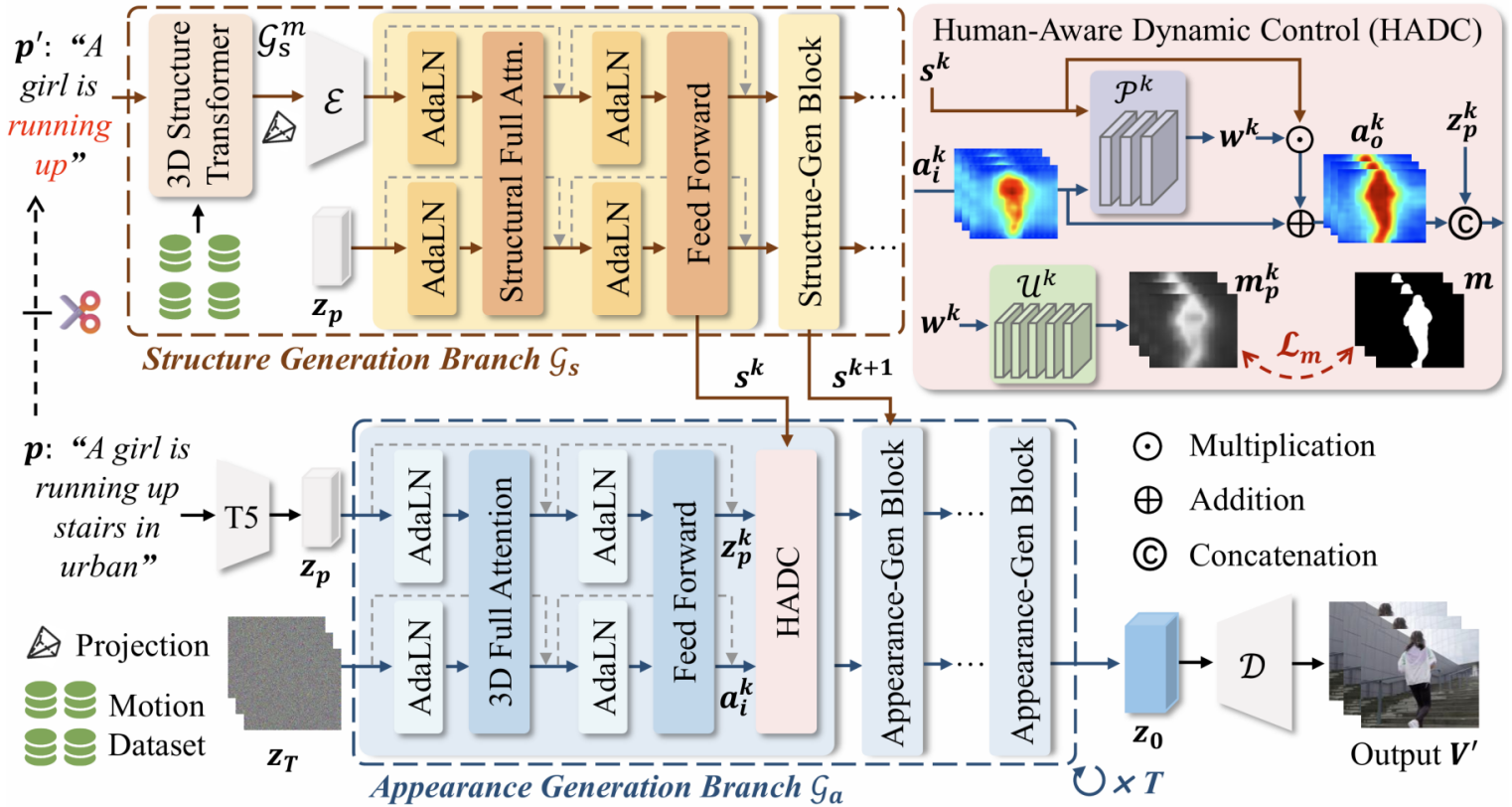
- Given a text cue p, the structure generation branch G_s aims to generate a human motion structure consistent with the motion semantics conveyed by p.
- This approach rephrases the text-driven structure generation task as a 3D keypoint sequence generation task. Based on motion-specific text cues, it generates 3D human keypoints using an autoregressive 3D Structure Transformer and renders them into a 2D human skeleton sequence.
- After obtaining g_s, the scheme introduces a structure generation module based on the DiT architecture, which encodes g_s as a structure control signal $s^{1:N}$ to guide subsequent appearance generation.





Method

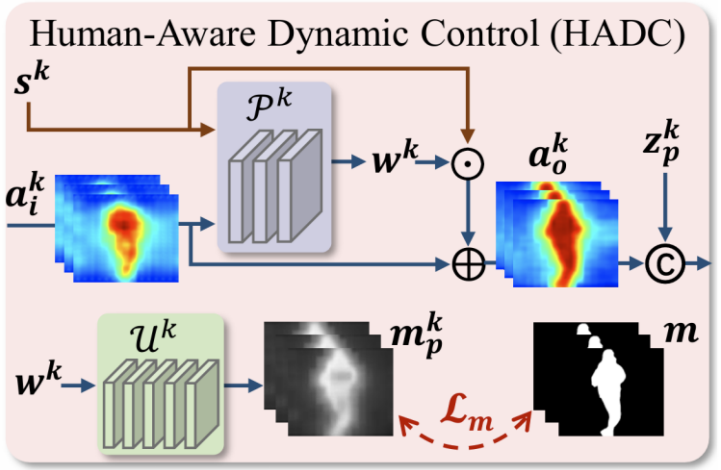
- The appearance generation branch G_a aims to synthesize realistic video content based on the input p and $s^{1:N}$, effectively capturing the appearance of the environment and the features of the human body.
- Considering that the structural guidance $s^{1:N}$ uses a sparse skeleton representation, this scheme introduces a HADC module in this branch to enhance $s^{1:N}$'s fine-grained control capability over human motion structures.



Method

- To ensure the effectiveness of the weights in the HADC module, this scheme designs a learnable network:

$$\mathcal{L}_m = \sum_{k=1}^{N} \left\| \mathcal{U}^k(w^k) - \mathcal{E}(M) \right\|_2^2,$$



- To further enhance the ability to model complex motions, this scheme introduces L_{track}

$$\mathcal{L}_{track} = \frac{1}{\sum_{(t_v,t'_v)}^{\mathcal{S}} e^{\frac{|t_v - t'_v|}{2}}} \sum_{(t_v,t'_v)}^{\mathcal{S}} e^{\frac{|t_v - t'_v|}{2}} \cdot \left\| U'_{t_v \to t'_v} - U_{t_v \to t'_v} \right\|_1,$$

- To address the shortcomings of existing models in human-environment interaction modeling, this scheme proposes a three-dimensional contact constraint L_{cont}

$$\mathcal{L}_{cont}^{f} = \begin{cases} \sum\limits_{h \in H_f^-} \left| SDF(h) - \tau \right|, & \text{if } |H_f^-| > 0, \\ \min\limits_{h \in H_f} SDF(h) - \tau, & \text{otherwise,} \end{cases}$$

- The overall training objective is

$$\mathcal{L} = \mathcal{L}_d + \lambda_m \mathcal{L}_m + \lambda_{track} \mathcal{L}_{track} + \lambda_{cont} \mathcal{L}_{cont}$$

Experiments

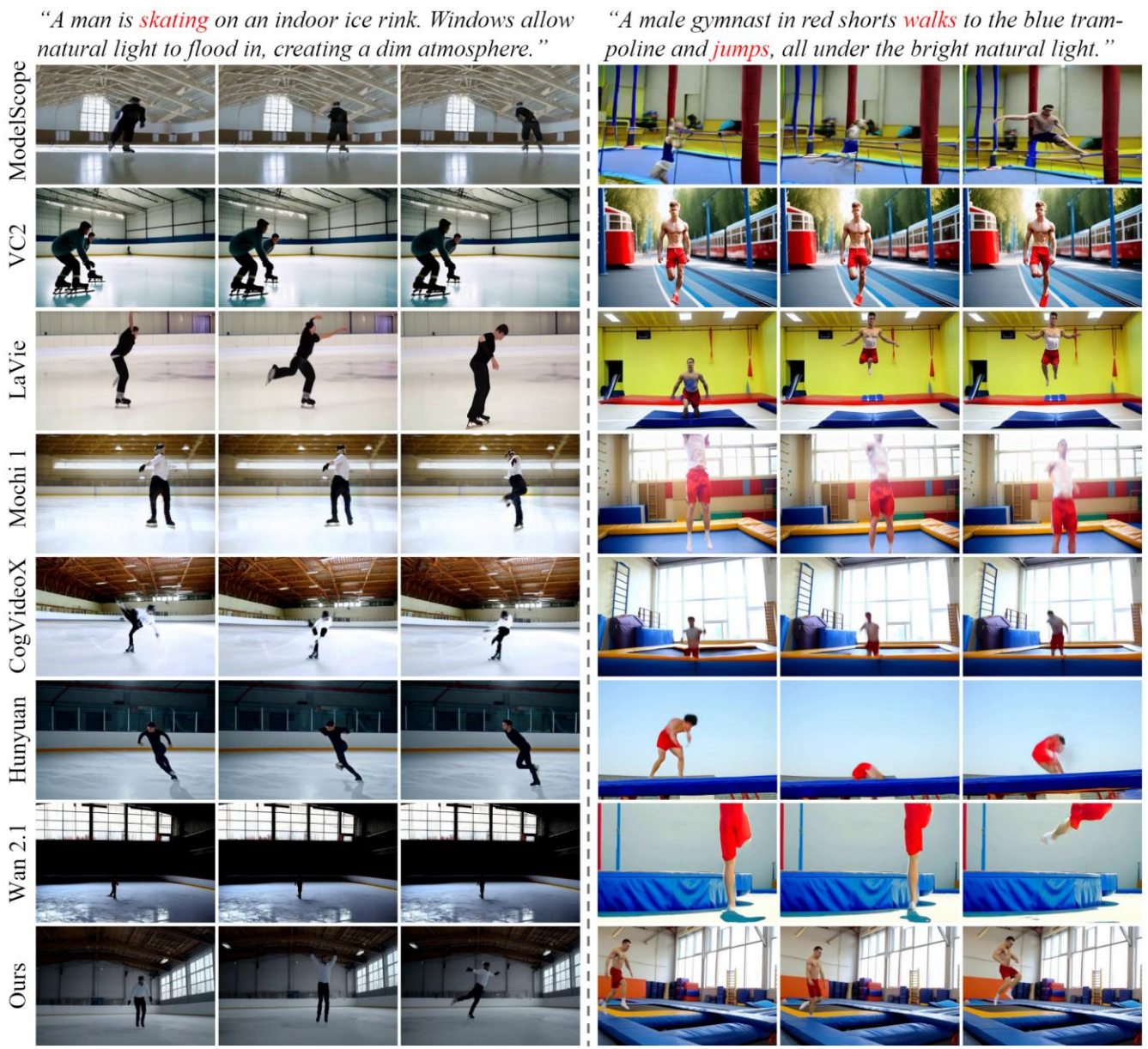
- Quantitative comparison with existing general video generation models

Method	FVD	CLIPSIM	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Imaging Quality
ModelScope	1945	0.2739	90.87%	93.41%	96.22%	48.57%	60.12%
VideoCrafter2	1959	0.2801	93.43%	97.01%	97.31%	35.71%	60.32%
LaVie	1778	0.2895	93.80%	95.51%	97.21%	**53.73%**	62.57%
Mochi 1	1207	0.2903	94.67%	95.32%	97.75%	51.14%	54.65%
CogVideoX	1360	0.2899	93.75%	94.02%	97.78%	51.42%	62.98%
HunyuanVideo	1235	0.2948	94.41%	95.17%	98.95%	50.42%	58.13%
Wan 2.1	1251	0.2951	94.43%	95.55%	98.36%	51.71%	65.21%
Ours	**1093**	**0.3035**	**96.83%**	**97.43%**	**99.25%**	52.86%	**65.43%**

- Compared with existing human video generation models

Method	FVD	CLIPSIM	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Imaging Quality
Animate Anyone	1362	0.2850	94.09%	95.33%	97.23%	41.28%	57.06%
HumanVid	1374	0.2876	95.12%	94.77%	97.42%	42.34%	54.05%
MIMO	1285	0.2904	94.82%	95.49%	97.38%	45.57%	53.83%
StableAnimator	1326	0.2895	95.37%	94.89%	97.88%	42.85%	56.96%
Ours	**1108**	**0.3021**	**97.74%**	**97.37%**	**99.31%**	**52.86%**	**64.68%**

Experiments



Experiments

A man is skating on an indoor ice rink. Windows allow natural light to flood in, creating a dim atmosphere ...

ModelScope

Mochi 1

VideoCrafter

CogVideoX

LaVie

Hunyuan

Wan 2.1

Ours

Experiments

A male gymnast in red shorts walks to the blue trampoline and jumps, all under the bright natural light ...

ModelScope

Mochi 1

VideoCrafter

CogVideoX

LaVie

Hunyuan

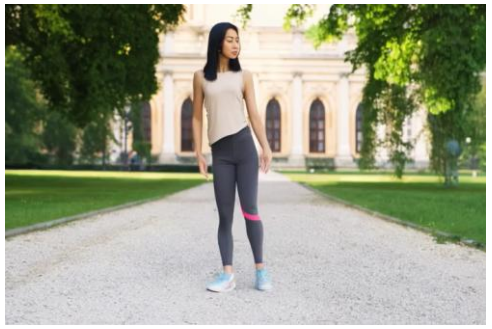
Wan 2.1

Ours

Experiments

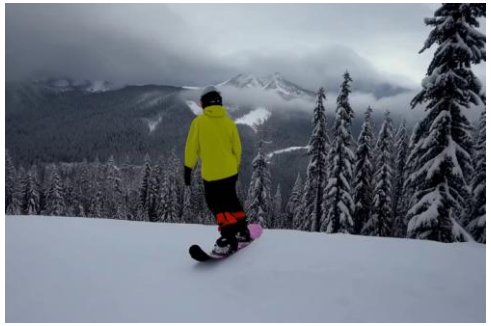
A man stretches on a wooden pier, placing one hand on his foot and the other reaching towards his leg .


A young woman, clad in a sleeveless beige top, stretches her leg on a gravel path in a tranquil park.

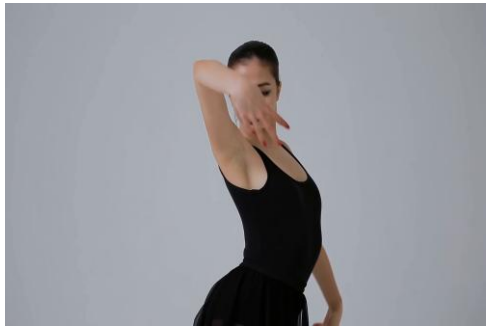



Experiments

Results of human body and background under dynamic camera trajectory

Half-body video results




Multi-person video results


Experiments












Experiments

- Quantitative comparison with existing human video datasets

Dataset	Clips	Resolution	Action Types ↑	Action Complexity ↑	Fine-grained Caption
CelebV-HQ	35K	512×512	Facial type	0.6891	-
CelebV-Text	70K	512×512	Facial type	0.7070	Text
TikTok	340	604×1080	Dancing	0.6816	-
UBC-Fashion	500	720×964	Standing	0.3321	-
IDEA-400	12K	720P	5K	0.5969	-
HumanVid	20K	1080P	7K	0.6669	-
MoVid (Ours)	30K	1080P	17K	1.1124	Text

• Experiments

- Analysis of the effect of the structure-appearance decoupling framework
- HADC Module Performance Analysis
- Analysis of the effect of dense tracking constraints
- 3D Contact Constraint Effect Analysis

Table 3: Effect of the contact constraint.

Constraint $\mathcal{L}_{cont}$	FVD	CLIPSIM
✘	1108	0.3021
Ours	**1093**	**0.3035**

Table 4: Effect of the decoupling framework.

Structure Generation Branch	FVD	CLIPSIM
✘	1262	0.2971
2D Structure Generation	1230	0.2998
Ours	**1093**	**0.3035**

Table 5: Effect of the HADC modules.

HADC Modules	FVD	CLIPSIM
✘	1188	0.2973
w/o $\mathcal{L}_m$	1112	0.3009
Ours	**1093**	**0.3035**

Table 6: Effect of the dense tracking loss $\mathcal{L}_{track}$.

Dense Tracking Loss $\mathcal{L}_{track}$	FVD	CLIPSIM
✘	1172	0.3009
Static Weights	1114	0.3016
Ours	**1093**	**0.3035**

• Experiments

- Analysis of the effect of the structure-appearance decoupling framework

• Experiments

- HADC Module Performance Analysis
- Analysis of the effect of dense tracking constraints

(a) Effect of the HADC modules

(b) Effect of the dense tracking loss

Thanks