Enhancing Motion in Text-to-Video Generation with Decomposed Encoding and Conditioning

Penghui Ruan[1,2], Pichao Wang[3], Divya Saxena[1], Jiannong Cao[1], Yuhui Shi[2]
[1]The Hong Kong Polytechnic University
[2]Southern University of Science and Technology
[3]Amazon
Accepted at NeurIPS 2024

November 11, 2024

Motivation

- We observed that current text-to-video generation models struggle with accurately understanding and generating motions.

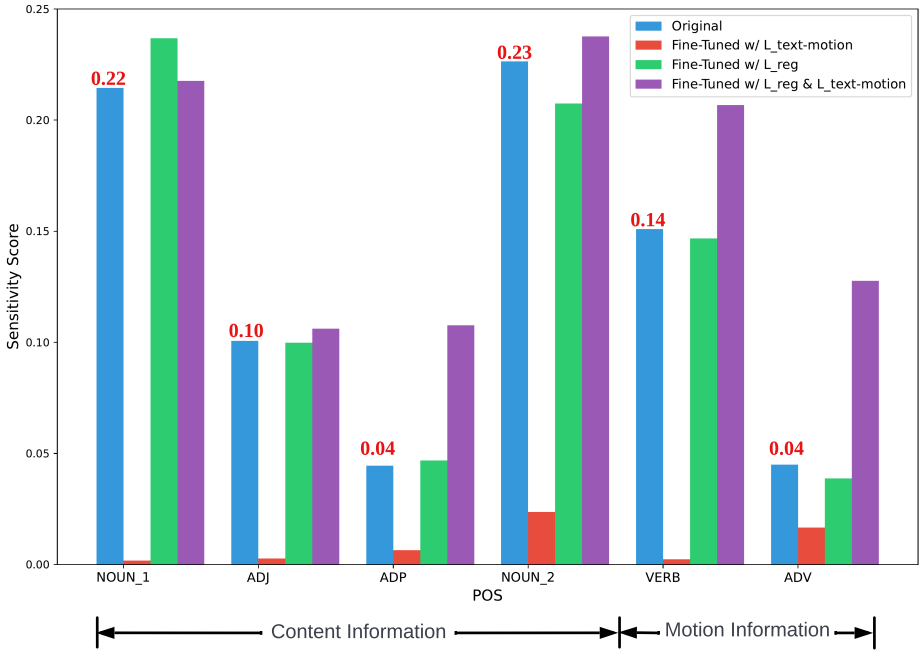
A man is practicing Taichi.

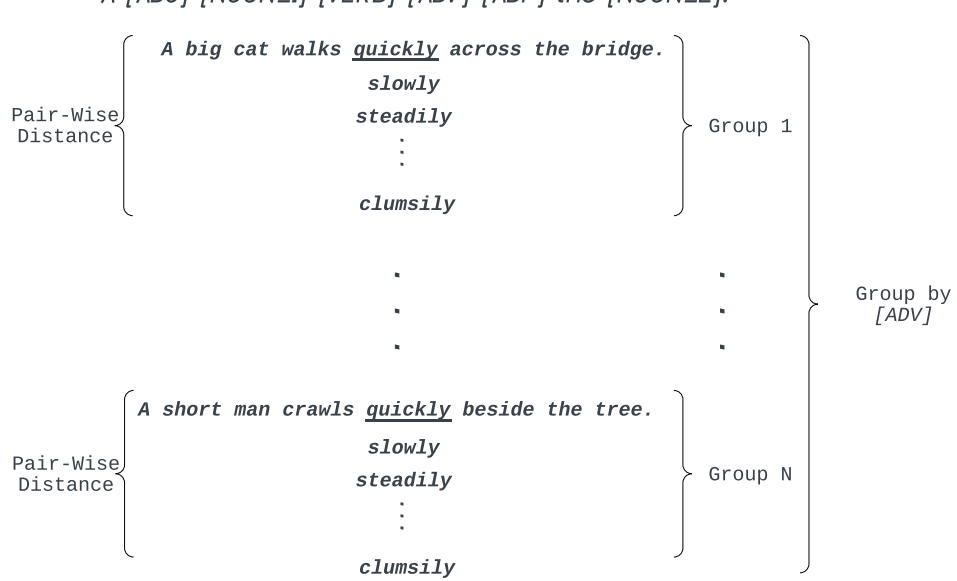
A goat is climbing the mountain.

Two men are wrestling.

Insufficient Text Encoding

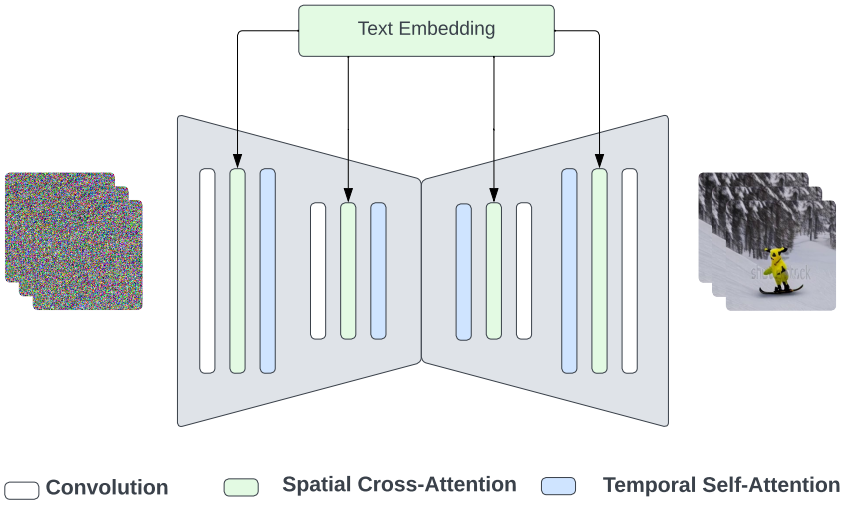
- The text encoding is significantly **biased towards nouns and objects**, with **insufficient consideration of motion information**.



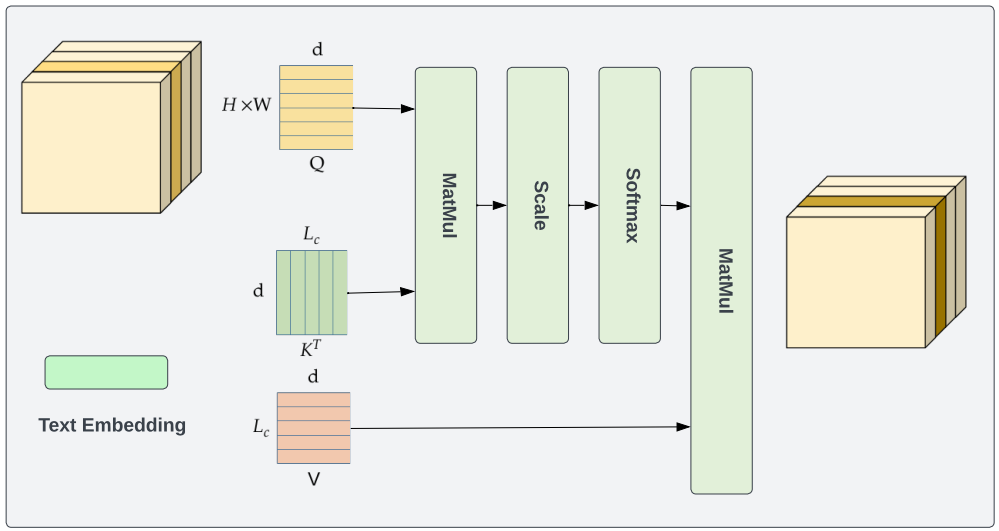


Insufficient Text Conditioning

- Current text conditioning mechanisms operate **only in the spatial dimension**, whereas **motion** is an essential element intertwined with **both space and time**.



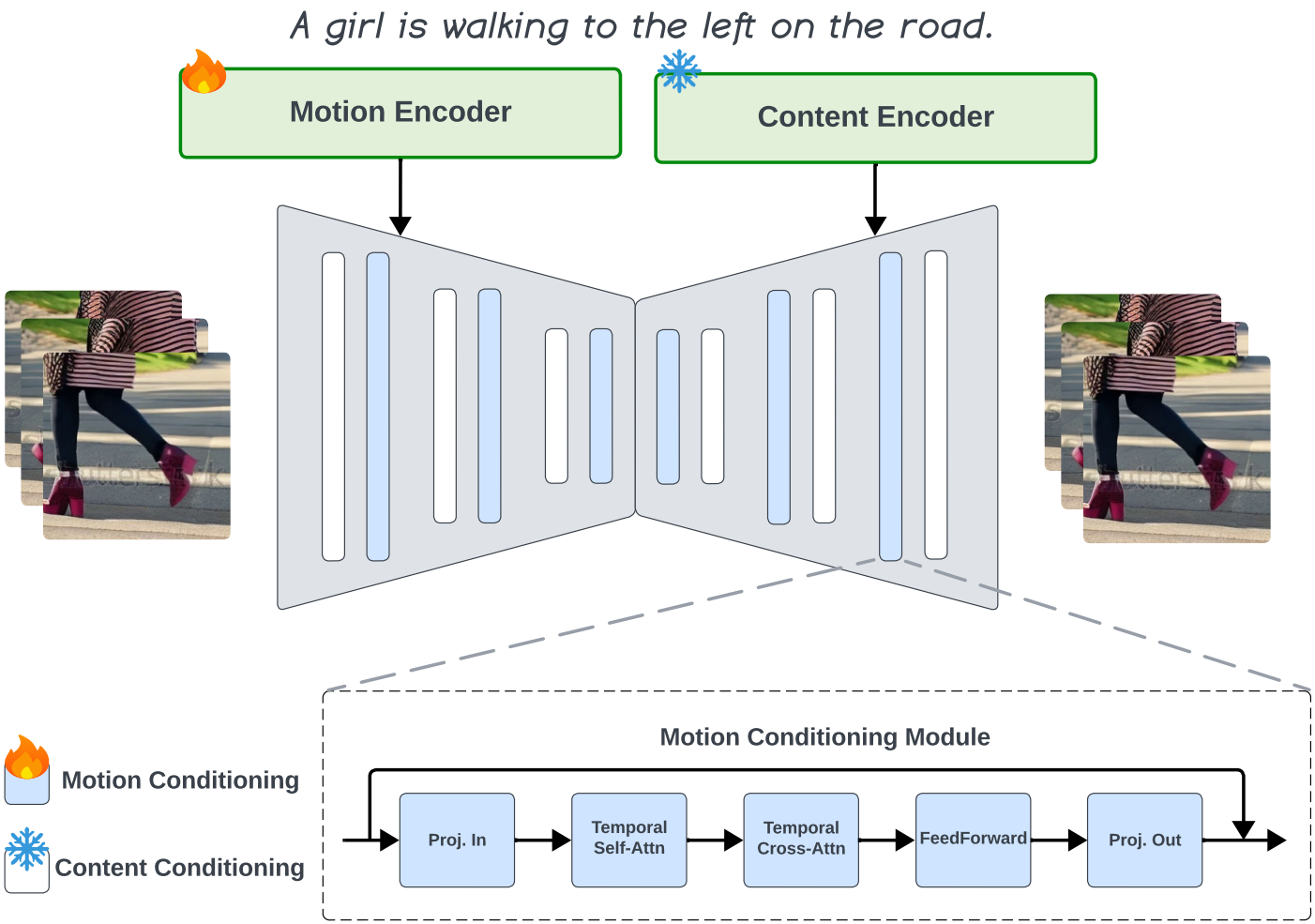
Existing Text-to-Video Generation Model Architecture.



Spatial Cross-Attention Mechanism.

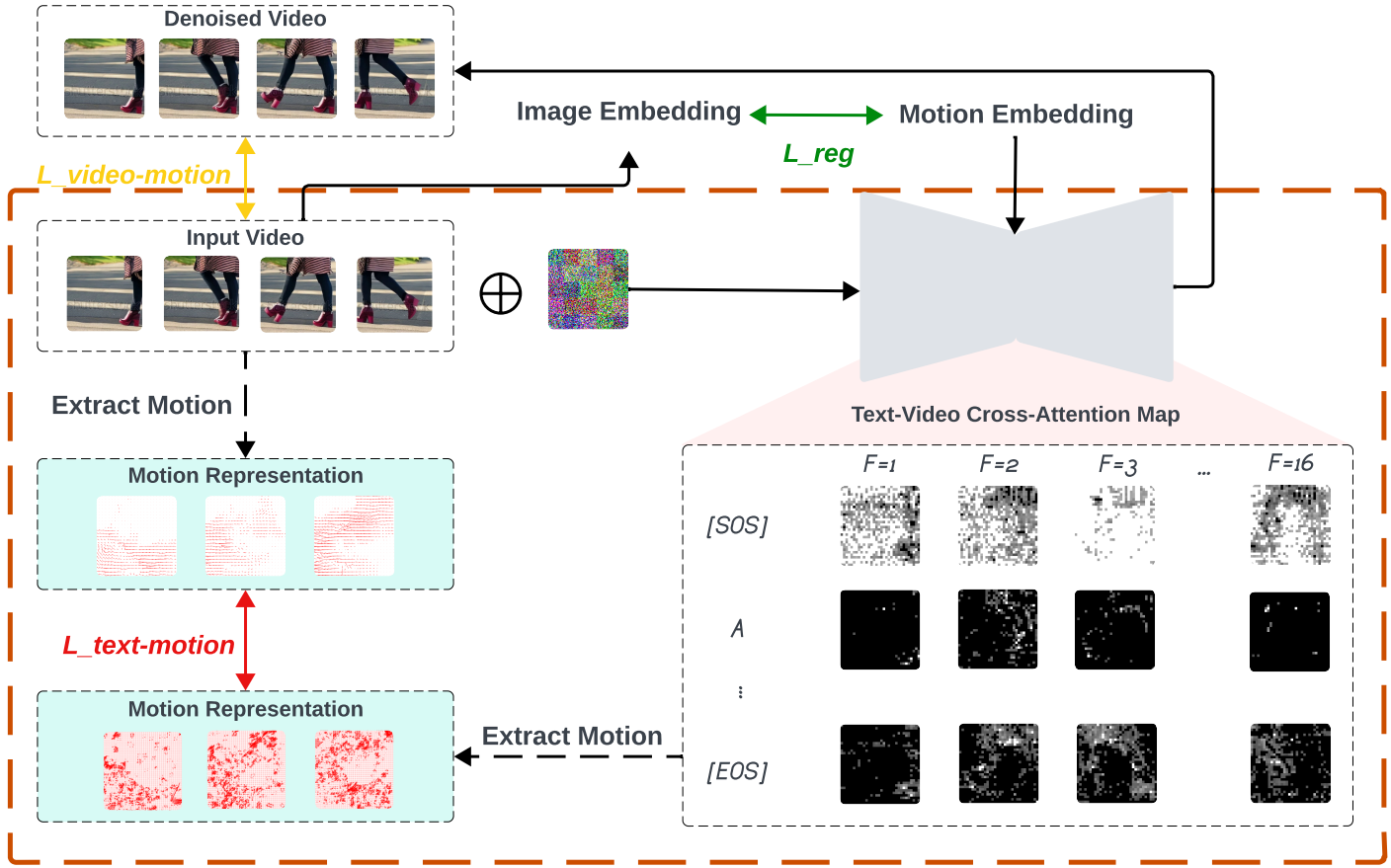
Purposed Solution: DEcomposed MOtion (DEMO)

- Decompose the text encoding and text conditioning into separated **content** and **motion** dimensions.



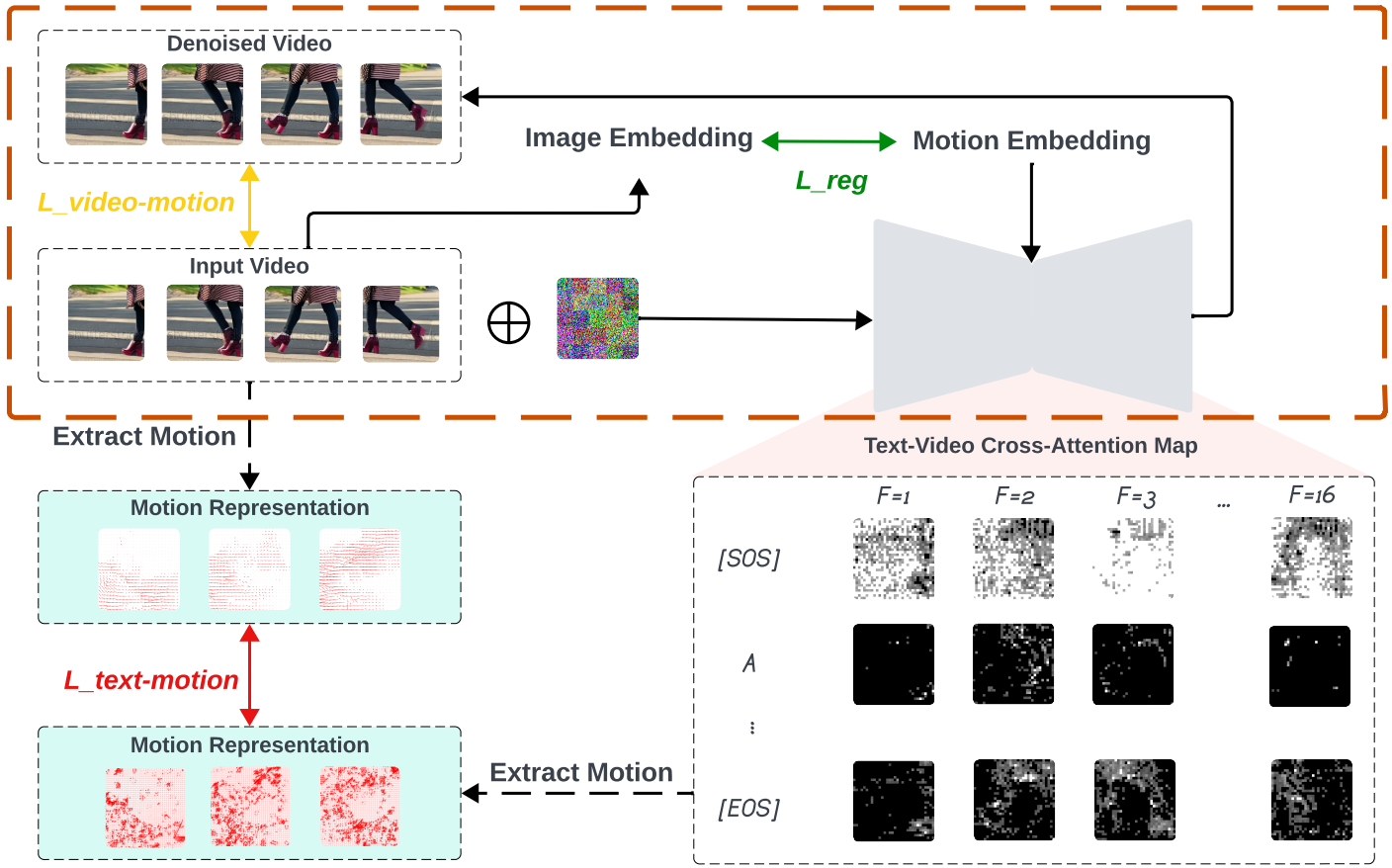
Purposed Solution

- Enhance the text encoding with $\mathcal{L}_{\text{text-motion}}$ (text-motion supervision) and $\mathcal{L}_{\text{reg}}$ (regularization).



Purposed Solution

- Enhance the motion generation with $\mathcal{L}_{\text{video-motion}}$ (video-motion supervision).



Results of Quantitative Evaluation

Table 1: Results of zero-shot T2V generation on MSR-VTT (Evaluation protocol comparison can be found in the appendix).

Model	FID (↓)	FVD (↓)	CLIPSIM (↑)
MagicVideo [74]	-	1290	-
Make-A-Video [46]	13.17	-	0.3049
Show-1 [70]	13.08	538	0.3072
Video LDM [4]	-	-	0.2929
LaVie [59]	-	-	0.2949
PYoCo [14]	10.21-9.73	-	-
VideoFactory [58]	-	-	0.3005
EMU VIDEO [45]	-	-	-
SVD [3]	-	-	-
ModelScopeT2V[3] [56]	14.89	557	0.2941
ModelScopeT2V fine-tuned	13.80	536	0.2932
DEMO	**11.77**	**422**	**0.2965**

Table 2: Results of zero-shot T2V generation on UCF-101 (Evaluation protocol comparison can be found in the appendix).

Model	IS (↑)	FVD (↓)
MagicVideo [74]	-	655.00
Make-A-Video [46]	33.00	367.23
Show-1 [70]	35.42	394.46
Video LDM [4]	33.45	550.61
LaVie [59]	-	526.30
PYoCo [14]	47.76	355.19
VideoFactory [58]	-	410.00
EMU VIDEO [45]	42.70	606.20
SVD [3]	-	242.02
ModelScopeT2V [56]	**37.55**	628.17
ModelScopeT2V fine-tuned	37.21	612.53
DEMO	36.35	**547.31**

Table 3: Results of T2V generation on WebVid-10M (Val).

Model	FID (↓)	FVD (↓)	CLIPSIM (↑)
ModelScopeT2V	11.14	508	0.2986
ModelScopeT2V fine-tuned	10.53	461	0.2952
DEMO	**9.86**	**351**	**0.3083**

Results of Quantitative Evaluation

Table 4: Results of zero-shot T2V generation on EvalCrafter.

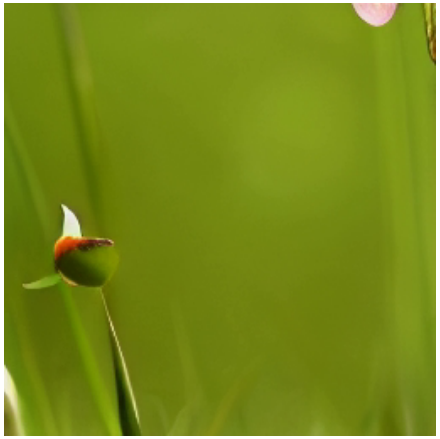
Model	Video Quality			Motion Quality		
	VQA_A (↑)	VQA_T (↑)	IS (↑)	Action Score (↑)	Motion AC-Score (↑)	Flow Score (↑)
ModelScopeT2V	15.12	**16.88**	14.60	75.88	44	2.51
ModelScopeT2V fine-tuned	15.89	16.39	14.92	74.23	40	2.72
DEMO w/o $\mathcal{L}_{\text{video-motion}}$	18.78	15.12	17.13	76.20	48	3.11
DEMO	**19.28**	15.65	**17.57**	**78.22**	**58**	**4.89**

Table 5: Results of zero-shot T2V generation on VBench.

Model	Motion Dynamics (↑)	Human Action (↑)	Temporal Flickering (↑)	Motion Smoothness(↑)
ModelScopeT2V	62.50	90.40	96.02	96.19
ModelScopeT2V fine-tuned	63.75	90.40	**96.35**	**96.38**
DEMO	**68.90**	**90.60**	94.63	96.09

Results of Qualitative Evaluation

ModelScopeT2V | LaVie | VideoCrafter2 | **DEMO**

Slow motion flower petals fall from a blossom, landing softly on the ground.

ModelScopeT2V | LaVie | VideoCrafter2 | **DEMO**




An old man with white hair is shown speaking.

Thank You

Paper, code, and model are available at
https://pr-ryan.github.io/DEMO-project/

THANKS FOR YOUR ATTENTION.