ICCV HONOLULU HAWAII
OCT 19-23, 2025

EfficientMT: Efficient Temporal Adaptation for Motion Transfer in Text-to-Video Diffusion Models

Yufei Cai[1,2], Hu Han[1,2], Yuxiang Wei[3], Shiguang Shan[1,2], Xilin Chen[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Harbin Institute of Technology

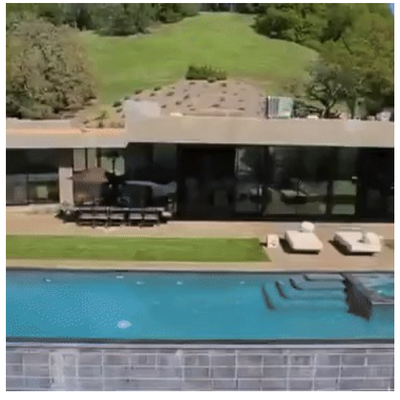
Paper: https://arxiv.org/abs/2503.19369 Code: https://github.com/PrototypeNx/EfficientMT

Video Motion Transfer

Reference Video

Woman walking in the desert

Astronaut walking on the beach

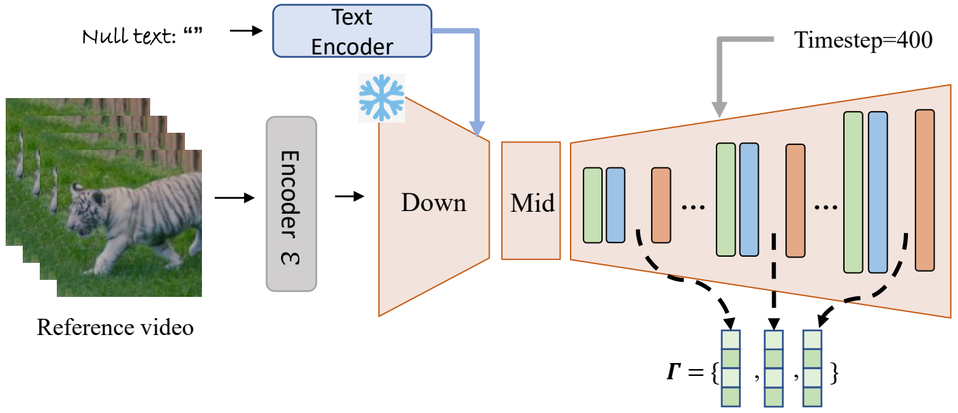


A starry night sky above a quiet mountain range

A rocky cliff overlooking a vast ocean horizon

EfficientMT

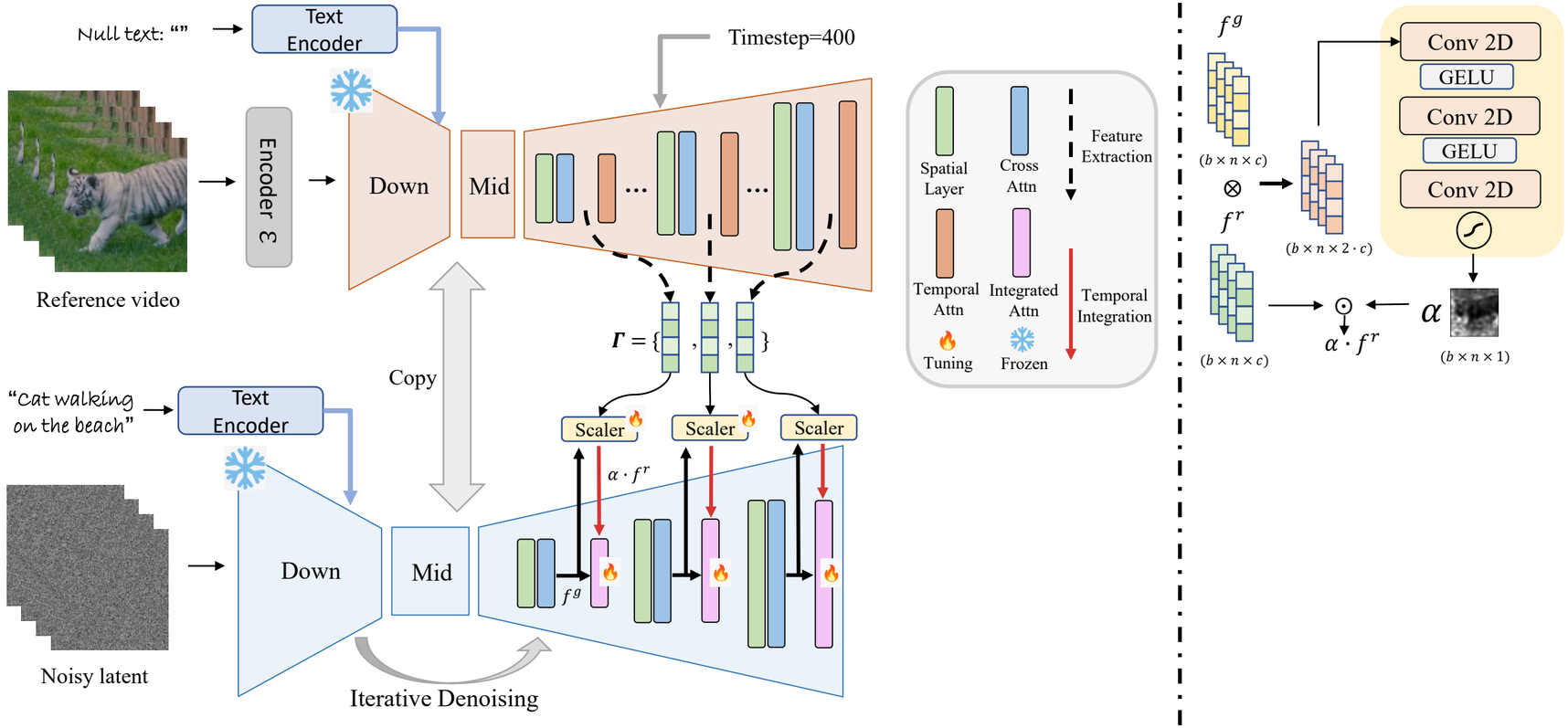
Repurpose T2V backbone as a motion feature extractor



$$\Gamma = \hat{\epsilon}(x, t, \tau_\theta(y^r)) = \{f\}_{upblocks}^{temporal}$$

EfficientMT

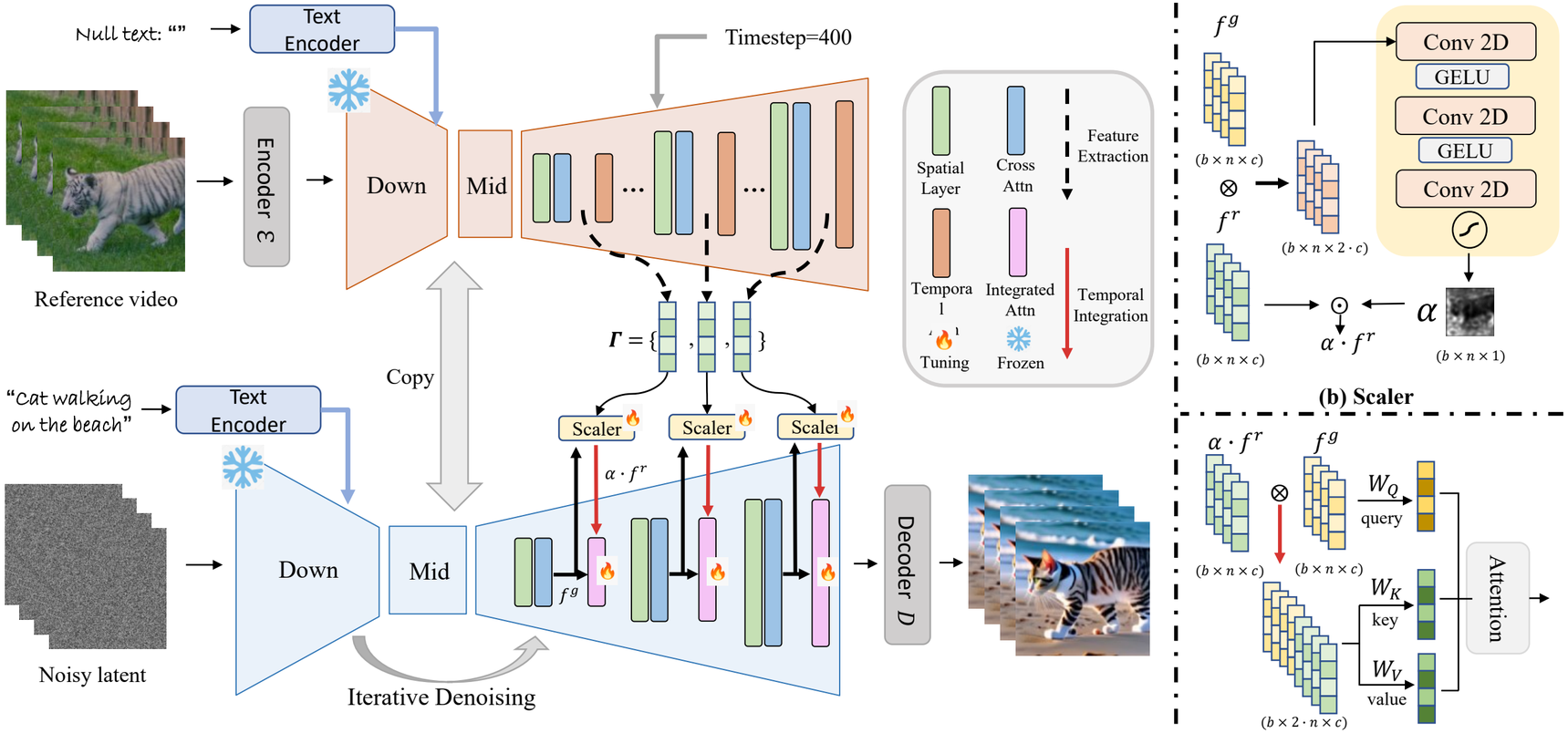
Introduce the scaler to predict a fine-grained scale map for the reference features to distill motion information



$$\alpha = \mathcal{S}(f^r \otimes f^g) \in \mathbb{R}^{h \times w \times n \times 1}, \quad \hat{f}^r = \alpha \cdot f^r$$

EfficientMT

Inject the motion features via Temporal Integrated Attention



$$f^{int} = f^g \otimes \hat{f}^r \in \mathbb{R}^{h \times w \times 2 \cdot n \times c} \qquad Q = W_Q \cdot f^g, K = W_K \cdot f^{int}, V = W_V \cdot f^{int}$$

Results

AnimateDiff

Reference Video

Lion walks in the desert

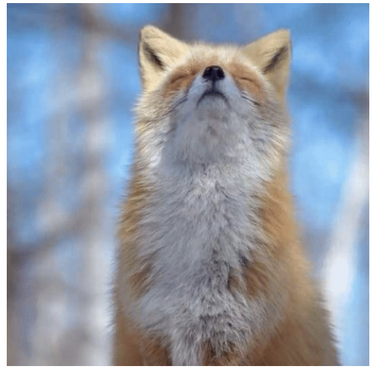
Cat walks on the sunlit beach

Reference Video

robot walks in the room

hunter walks in the forest

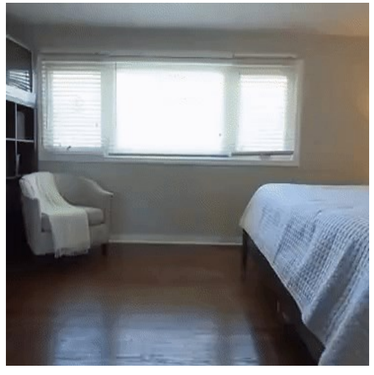
Reference Video



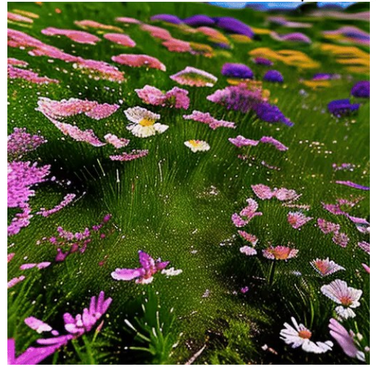
tiger moves its head in the forest

lion moves its head in the wild

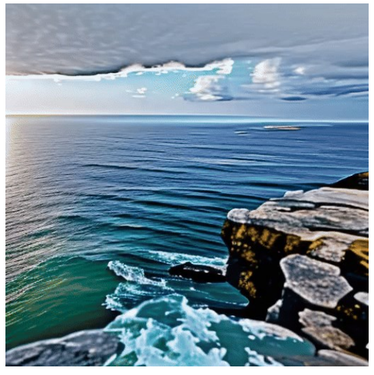
Reference Video



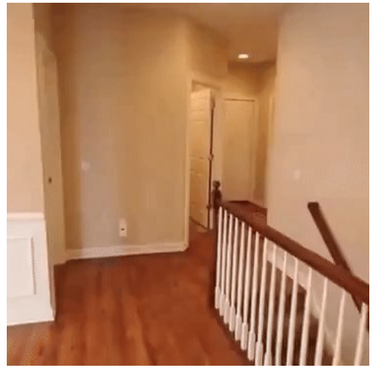
a vibrant meadow filled with wild flowers in spring



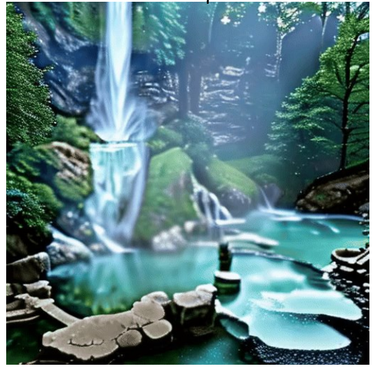
a rocky cliff overlooking a vast ocean horizon



Reference Video



a misty waterfall cascading into a clear pool below

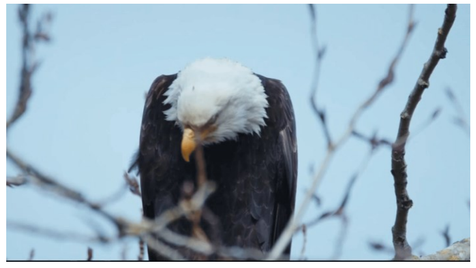


a volcanic landscape with steaming vents and rugged terra

Results

VideoCrafter2

Reference Video

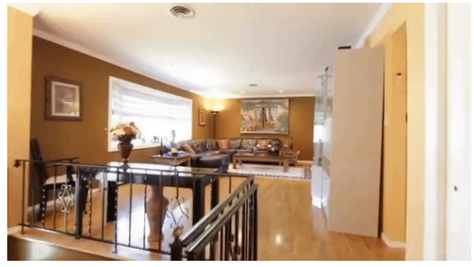


parrot raises head in the snow

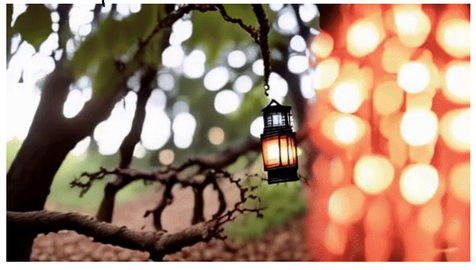
seagull raises head in the desert

tiger jumps in the city

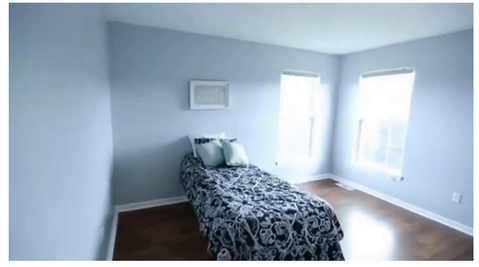
wolf jumps in the desert



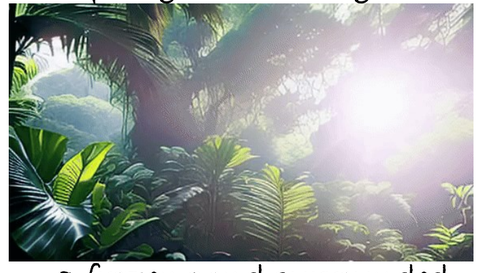
a glowing lantern hanging from a tree branch



a canyon with layered rock formations under sunlight



a tropical jungle with dense foliage and sunlight



a frozen pond surrounded by frost covered trees

Results

Quantitative Comparison

Method	Temporal Consistency (↑)	Text Alignment (↑)	Motion Fidelity (↑)	Time Cost (↓)
ControlVideo [38]	0.9213	0.2483	0.5533	80s
VideoComposer [29]	0.9192	0.2635	0.6356	18s
Control-A-Video [4]	0.9172	0.2438	0.6225	25s
MotionDirector [39]	**0.9327**	0.2525	0.8361	473s
MotionClone [14]	0.9108	<u>0.2637</u>	**0.8569**	190s
MOFT [34]	0.9283	0.2581	0.7698	<u>127s</u>
Ours (AnimateDiff)	<u>0.9291</u>	**0.2712**	<u>0.8470</u>	**16s**
DMT [35]	0.9275	0.2479	0.6974	<u>203s</u>
MotionInversion [28]	<u>0.9329</u>	<u>0.2558</u>	**0.7373**	418s
Ours (VideoCrafter2)	**0.9456**	**0.2677**	<u>0.7116</u>	**21s**

Ablation Study

Method	TC (↑)	TA (↑)	MF (↑)
w/o Scaler	0.9244	0.2638	0.8135
w/o Data Filter	0.9237	0.2649	0.8278
Inject Upblock.1	0.9013	**0.2789**	0.6374
Inject Upblock.1,2	0.9198	0.2755	0.7236
Ours	**0.9291**	0.2712	**0.8470**

User Study

Method	VC	MD	MC	MI
Temporal Consistency	75.3	48.4	55.9	57.0
Text Alignment	80.6	64.5	57.0	61.3
Motion Fidelity	75.3	78.4	49.5	58.1

Take home message

- We propose EfficientMT, an efficient video motion transfer framework that adapts a pretrained T2V model into an end-to-end motion transfer model using only a small set of synthesized data samples.
- We reuse the backbone for motion feature extraction and introduce temporal integration mechanism combined with a scaler to distill motion guidance into the generation process.
- Extensive experiments show that our EfficientMT achieves a faster generation process while exhibiting robust motion transfer capabilities.

Thank you for watching!