Ground-A-Video: Zero-shot Grounded Video Editing Using Text-to-Image Diffusion Models

Hyeonho Jeong, Jong Chul Ye
Graduate School of AI, KAIST, Korea

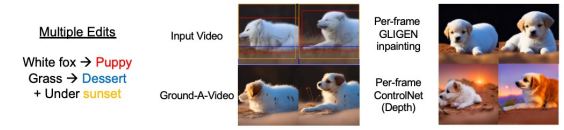
KAIST

Project

ICLR

Problems of Existing Approaches

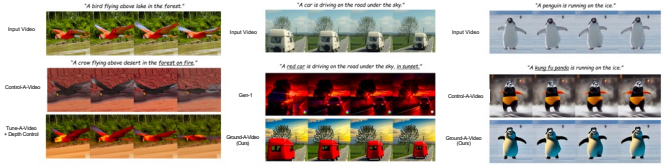
1. Video editing with text-to-image diffusion often fails frame consistency across generated frames.



2. Attention inflation based existing methods exhibit semantic misalignments when faced with the complexities of multi-attribute editing.

Attention Inflation $Q = W^Q \cdot z_t^i, K = W^K \cdot [z_t^1, \ldots, z_t^N], V = W^V \cdot [z_t^1, \ldots, z_t^N],$

Semantic Misalignment



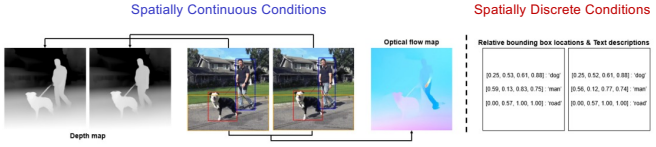
Neglected Edit
Edit on wrong element

Mixed Edit

Preservation Failure

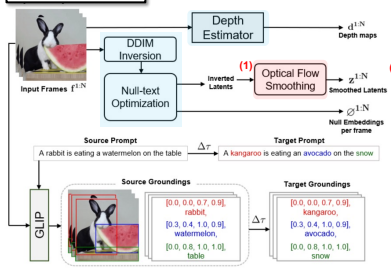
Ground-A-Video: First Framework for Multi-attribute Video Editing

Spatially Discrete Conditions (Groundings) → Disentanglement of Multiple Edits
Spatially Continuous Conditions (Depth Maps, Optical Flows) → Temporal Consistency



Ground-A-Video Pipeline

Input Preparation



→ Per-frame depth map estimation
→ Per-frame DDIM &null-text inversion

(1) **Optical Flow Smoothing:** makes static regions across frames share the same pixel intensities

Algorithm 1 Optical Flow guided Inverted Latents Smoothing

Require: Number of frames N, Source video frames $f^{1:N}$, Timesteps T, Text-Image Diffusion Model (SD), Optical flow estimator (RAFT), Threshold for magnitude difference M_{thres}

$z_T^{1:N} \leftarrow$ DDIM.INV($f^{1:N}$, T, SD) ▷ run inversion (e.g. DDIM inversion)

for all $i = 2, 3, \ldots, N$ **do**

$map_{opt}^i \leftarrow$ RAFT(f^{i-1}, f^i) ▷ obtain optical flow map

$map_{mag}^i \leftarrow$ normalize($\|map_{opt}^i\|$) ▷ compute magnitude map

$map_{mask}^i \leftarrow map_{mag}^i < M_{thres}$ ▷ obtain binary mask denoting static region

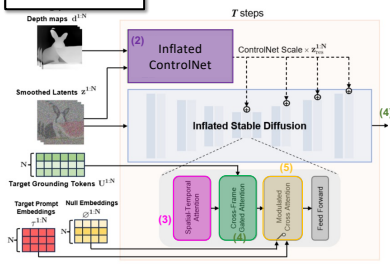
$map_{mask}^i =$ downsample(map_{mask}^i) ▷ downsample to latent-level size

$z_T^i = z_T^{i-1} * map_{mask}^i + z_T^i * (1 - map_{mask}^i)$ ▷ flow guided smoothing

end for

Return $z_T^{1:N}$ ▷ temporally smoothed latents

Denoising Process



(2) **ControlNet 3D (Depth):** ensures structural consistency between input and output frames

(4) **Video Groundings:** decompose multiple edits into tokens
Cross-Frame Gated Attention: Projects the grounding tokens onto latent features.

Video grounding tokens

M: number of objects in a single frame, e_i: text, l_i: bounding box

v_j^i: groundings token for j-th object in i-th frame $\quad v_j^i = \text{MLP}([\underbrace{\xi_{CLIP}(e_j^i)}_{\text{text token}}, \underbrace{\text{Fourier}(l_j^i)}_{\text{layout token}}])$

$U^i = [v_1, \ldots, v_M]$

Cross-Frame Gated Attention

$Q = W^Q \cdot [z_t^i, U^i], \quad K = W^K \cdot [z_t^1, U^1, \ldots, z_t^N, U^N], \quad V = W^V \cdot [z_t^1, U^1, \ldots, z_t^N, U^N]$

(3) **Self-Attention inflation:** enables inter-frame interaction.

Spatial-Temporal Self-Attention

$Q = W^Q \cdot z_t^i, K = W^K \cdot [z_t^1, \ldots, z_t^N], V = W^V \cdot [z_t^1, \ldots, z_t^N],$

z_t^i : latent representation of i-th frame at timestep t

W^Q, W^K, W^V : attention projection matrices

(5) **Modulated Cross Attention:** Handles variations between frame-independently optimized null embeddings.

Classifier-free Guidance

$$\tilde{\epsilon}_\theta(z_t, t, \tau, \varnothing) = w \cdot \underbrace{\epsilon_\theta(z_t, t, \tau)}_{\text{conditional prediction}} + (1 - w) \cdot \underbrace{\epsilon_\theta(z_t, t, \varnothing)}_{\text{unconditional prediction}}$$

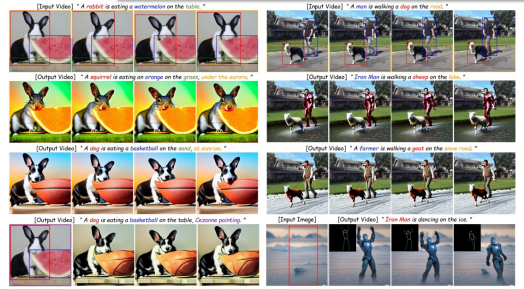
Modulated Cross Attention

$$Q = W^Q \cdot z_t^i, \; K = \begin{cases} W^K \cdot c_t^i & \text{if } cond \\ W^K \cdot [c_t^1, \ldots, c_t^N] & \text{if } uncond \end{cases}, \; V = \begin{cases} W^V \cdot c_t^i & \text{if } cond \\ W^V \cdot [c_t^1, \ldots, c_t^N] & \text{if } uncond \end{cases}$$

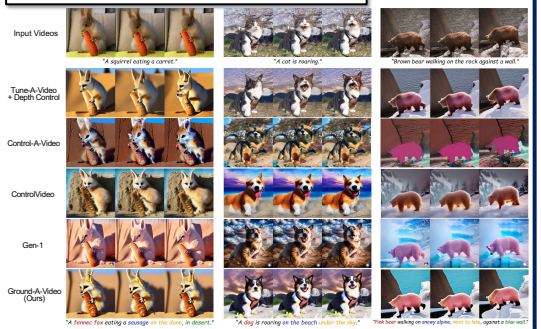
c_t^i : context vector of i-th frame at timestep t

Results

Multi-Attribute Editing, Style Transfer, Pose-to-Video Generation



Qualitative Comparisons (Check out our project page!)



Quantitative Comparisons

Method	CLIP		Human Evaluation		
	Text-Align	Frame-Con	Edit-Acc	Preserve-Acc	Frame-Con
Tune-A-Video (Wu et al, 2022)	0.810	0.959	2.99	3.13	3.05
Control-A-Video (Chen et al, 2023)	0.801	0.955	2.25	2.50	2.88
ControlVideo (Zhang et al, 2023)	0.822	0.963	2.36	2.02	2.08
Gen-1 (Esser et al, 2023)	0.833	0.939	2.41	2.51	2.56
Ground-A-Video (Ours)	**0.837**	**0.970**	**4.13**	**4.24**	**4.01**