Image-To-Video Generation

Konstantinos, Elyanne, Lisanne, Taiki, Wojciech

University of Amsterdam

Goals

Main: Improve video quality from single-image diffusion-based video generation by applying and combining post-processing methods.

Background

Motivation

- Real estate agencies often work with only a few high-quality interior photos. Generating video tours from these photos can reduce production time and cost.

Single-view vs Multi-view

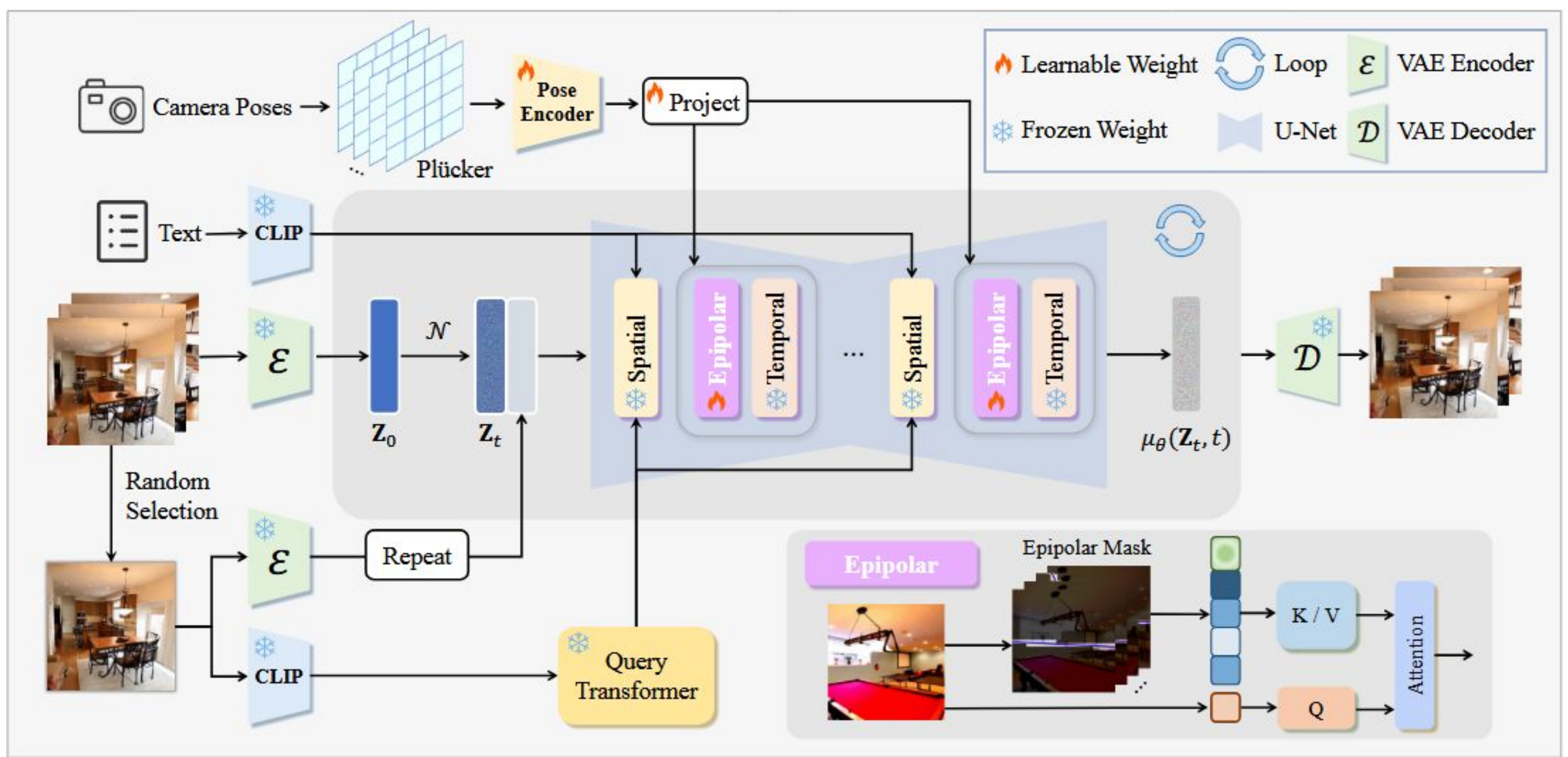
- Multi-view models require several images and camera parameters. We only had two PNGs per scene and no metadata, which made these approaches often impractical.

Limitations of the multi-view models

- Explored models: PixelSplat [1], NoPoSplat [2], LLFF [3], MVSplat [4], and LVSM [5].
  - ❌ Requires camera poses [1, 4, 5] and dense coverage of a scene [3]
  - ❌ COLMAP to estimate camera poses (exterinsic) failed on our data[1, 2, 3, 4, 5]
  - ❌ Outputs were often blurry or empty [1, 4, 5]

➤Changed to single-image diffusion-based generation + post-processing.

Explored Models



CamI2V Architecture [1]

CamI2V [1]

- Camera-guided image-to-video diffusion model
- Uses random camera paths to simulate smooth motion
- Takes a single image and generates realistic video output

Post Processing Methods:

FastDVDnet [2]

- Denoising model using temporal feature fusion
- Reduces flickering and improves visual consistency

Upscale-A-Video [3]

- Super-resolution model for video upscaling
- Enhances sharpness and detail of generated frames

[1] CAMI2V: Camera-Controlled Image-To-Video Diffusion Model, Guangcong Zhen et al. 2025, ICLR 2025
[2] FastDVDnet: Towards Real-Time Deep Video Denoising Without Flow Estimation, Matias Tassano et al. 2020, CVPR 2020
[3] Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution, Shangchen Zhou et al. 2024, CVPR 2024

Experimental Setup

Comparison

- For each of the 6 selected images, 3 distinct video movements are generated, resulting in 18 videos. Each video undergoes the following post-processing, leading to 72 processed videos:
  a. Denoising only
  b. Upscaling only
  c. Denoising → Upscaling
  d. Upscaling → Denoising

Metrics

- Peak Signal-to-Noise Ratio (PSNR):
  - Measures the fidelity of generated frames (pixel-level accuracy). Higher = better.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

- Structural Similarity Index (SSIM):
  - Assesses perceptual similarity, considering luminance, contrast, and structure. Closer to 1 = better.

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

Results & Discussion

- Denoising yielded the highest PSNR (27.338) by reducing pixel-level noise, which PSNR directly measures.
- The combined method achieved the best SSIM (0.917) by enhancing structural coherence and detail which SSIM measures.
- Limited overall PSNR improvement due to generative artifacts which persist beyond the post-processing fixes.
- Post-processing order is important: Upscaling→Denoising surpassed Denoising→Upscaling in SSIM (0.917 vs 0.902) and PSNR (27.280 vs 26.979), likely because upscaling after denoising can introduce new artifacts, whereas denoising last effectively cleans the upscaled image.

	CamI2V	CamI2V + Denoising	CamI2V + Upscaling	CamI2V + Denoising → Upscaling	CamI2V + Upscaling → Denoising
PSNR (↑)	26.267 (±0.709)	**27.338 (±0.954)**	26.889 (±0.901)	26.979 (±0.953)	27.280 (±0.970)
SSIM (↑)	0.816 (±0.012)	0.853 (±0.013)	0.896 (±0.009)	0.902 (±0.001)	**0.917 (±0.008)**