# Video completion conditioned by natural language-based descriptions

**Paul Orășan**
**Babeș-Bolyai University**

## WeADL 2023 Workshop

The workshop is organized under the umbrella of WeaMyL, project funded by the EEA and Norway Grants under the number RO-NO-2019-0133. Contract: No 26/2020.

Working together for a green, competitive and inclusive Europe

# Outline

# Text-Guided Video Completion

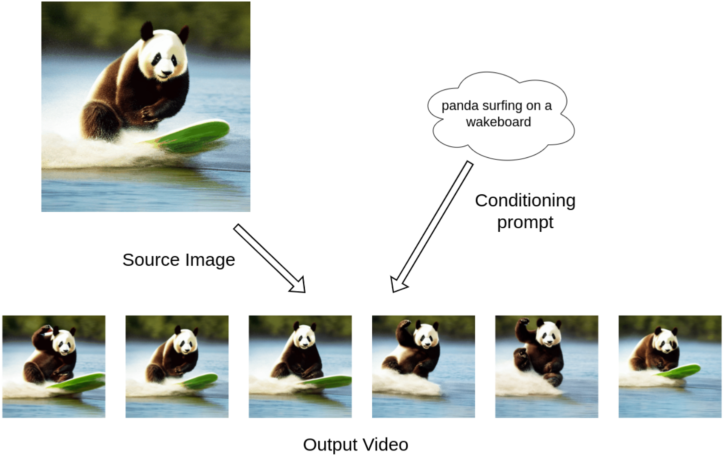



Figure: TVC input and output

# Text-Guided Video Completion

Applications in:

- Entertainment industry
- Gaming industry
- VR/AR
- Education and training

# Text-Video Prediction

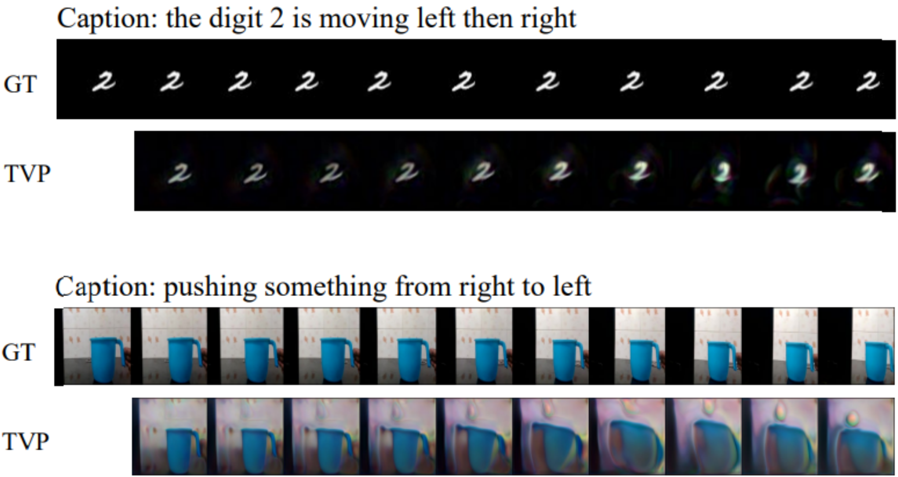



Figure: Samples of Text-Video Prediction [SCZJ22]

# Tell Me What Happened

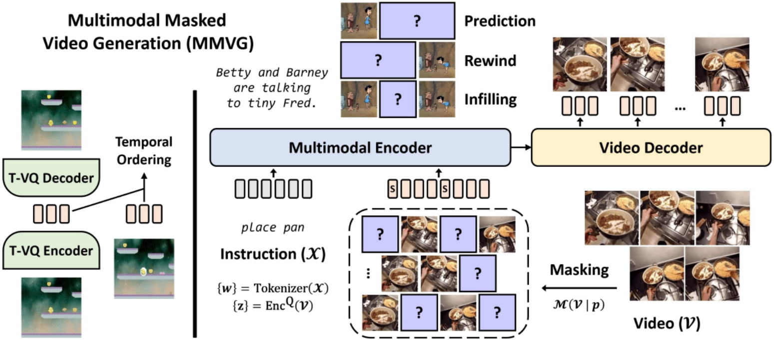



Figure: Multimodal Masked Video Generation (MMVG) architecture overview [FYZ+22]

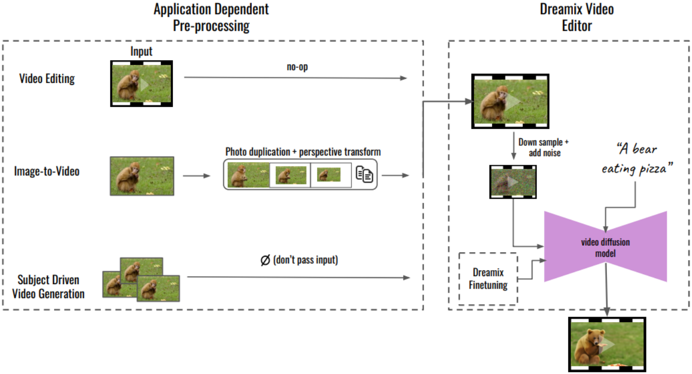



Figure: Dreamix architecture overview [MHV+23]

# Dreamix finetuning

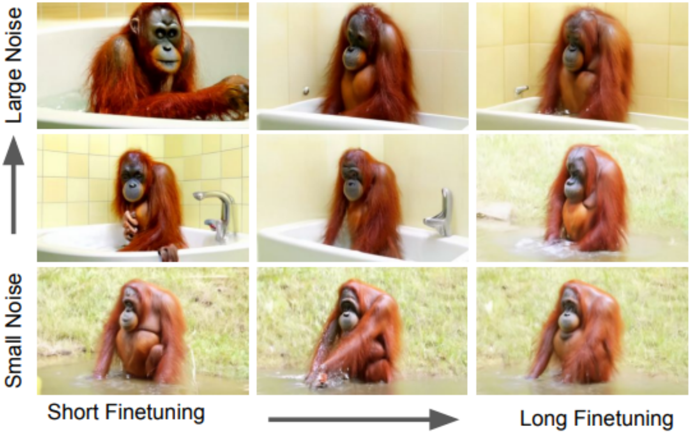
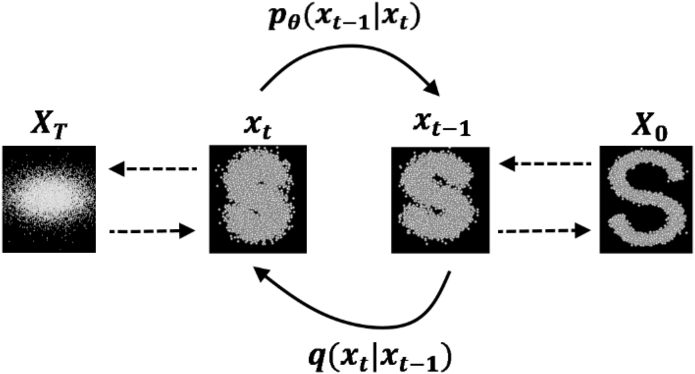



Figure: Dreamix variation [MHV+23]

# New Generative approach

Prafulla Dhariwal and Alexander Quinn Nichol, Diffusion models beat GANs on image synthesis, NeurIPS 2021



# Forward Diffusion

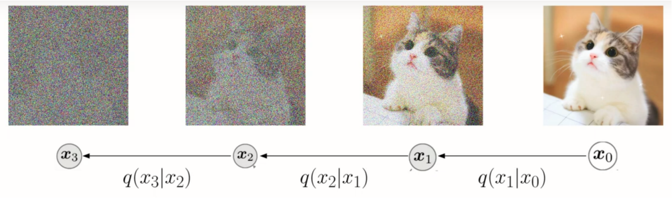



Figure: Forward diffusion process - adding noise each step

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1 - \alpha_{t-1})I) \tag{1}$$

# Backward Diffusion

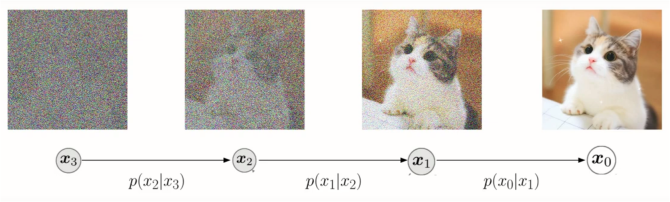
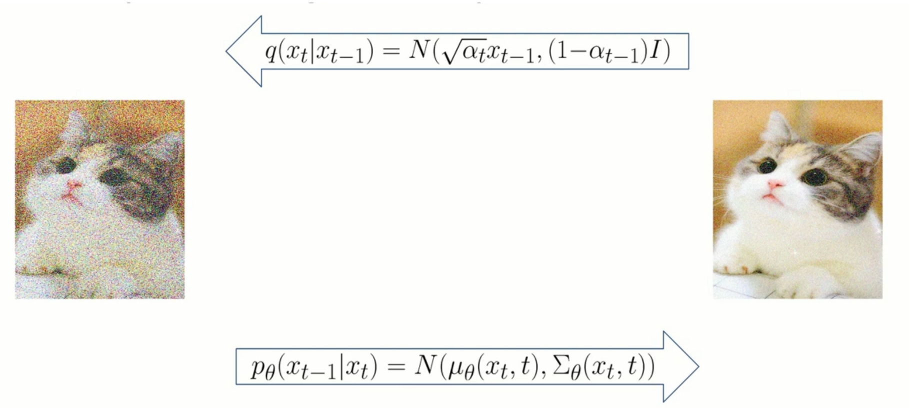


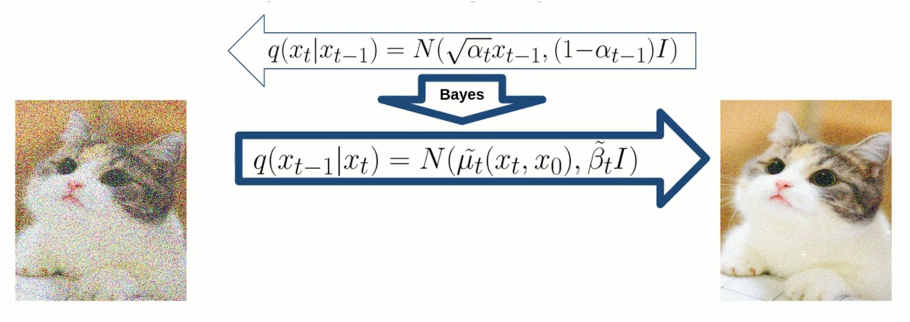

Figure: Backward diffusion process - removing noise each step

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2}$$

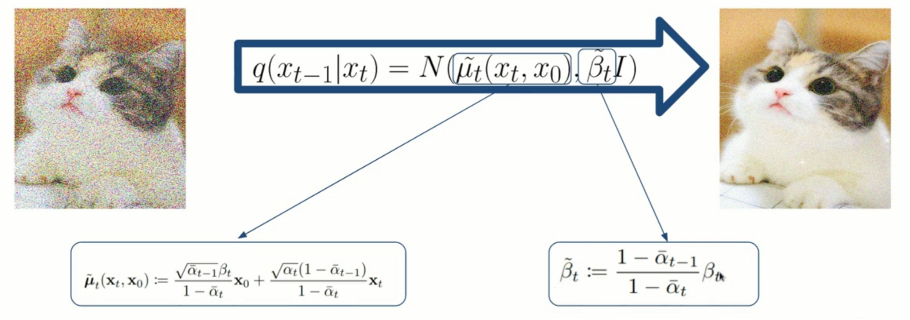
# Learning objective



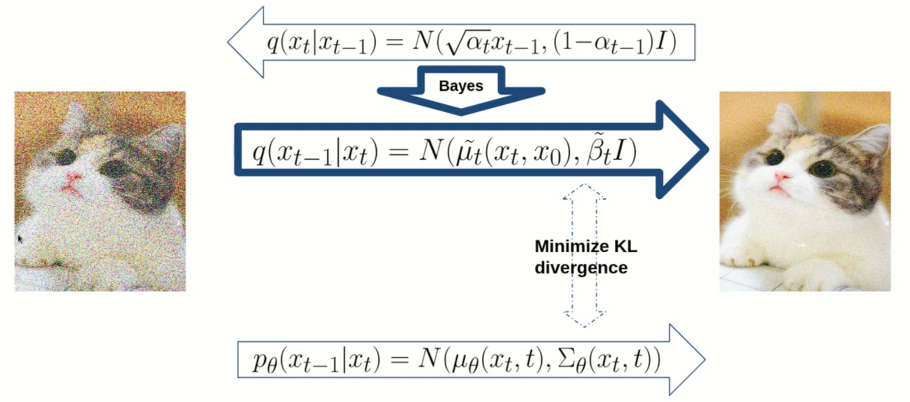
# Learning objective



# Learning objective



# Learning objective



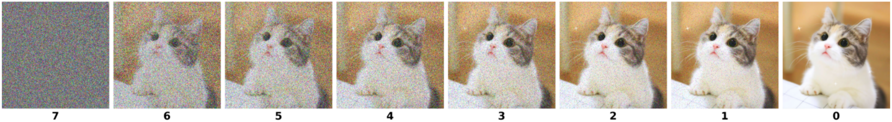

$$L_{t-1} = \mathbb{E}\left[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2\right] \tag{3}$$

# Practical learning goal

$$\boxed{\begin{aligned} \tilde{\mu}_t &= \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x_t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) \\ \mu_\theta(\boldsymbol{x_t}, t) &= \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x_t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x_t}, t)\right) \end{aligned}} \tag{4}$$

$$\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \rightarrow \|\epsilon - \epsilon_\theta(x_t, t)\|^2 \tag{5}$$

# Sampling equation




$$\boxed{\boldsymbol{x_{t-1}}(\boldsymbol{x_t}) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x_t}, t)\right) + \sigma_t \boldsymbol{z}} \tag{6}$$

# Stable Diffusion

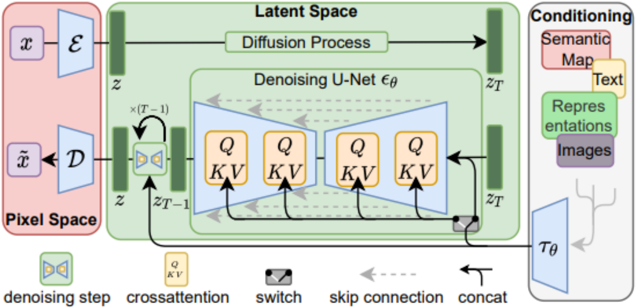
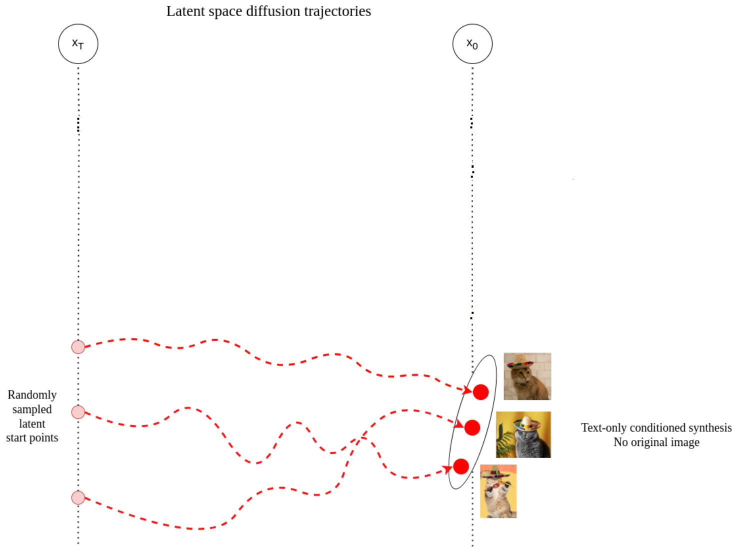
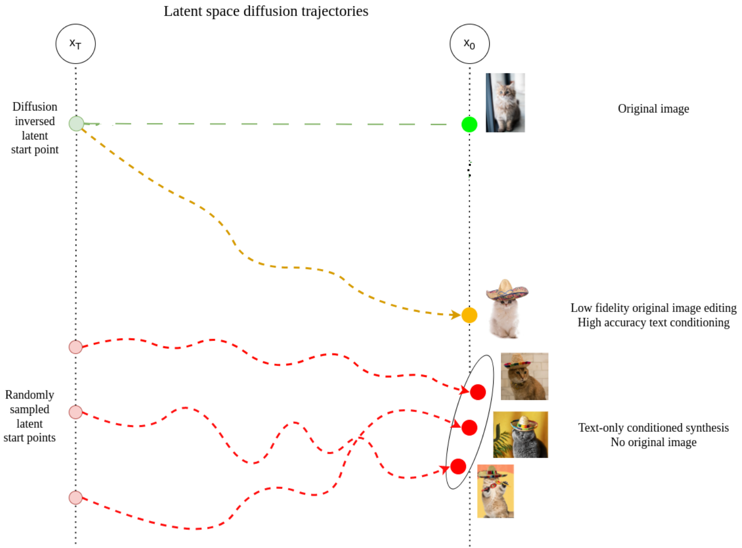



Figure: Stable Diffusion overall architecture [RBL+22]

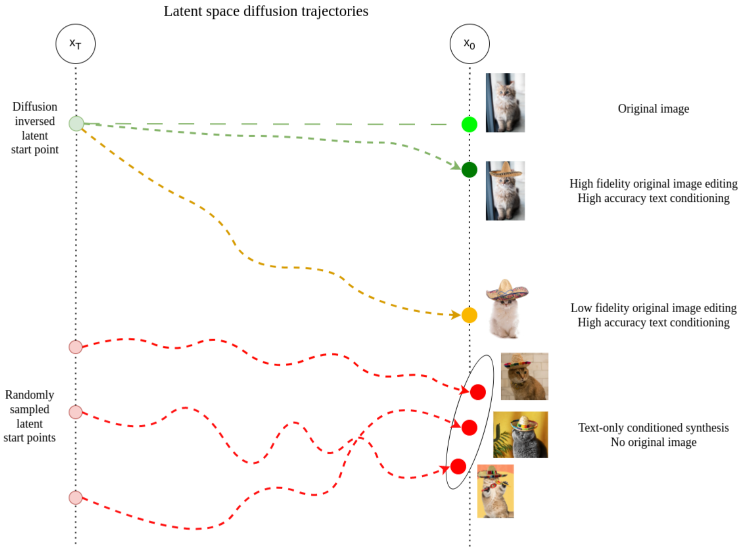
# General synthesis



# Source-image conditioned synthesis



# Our goal



Latent space diffusion trajectories
$x_T$
$x_0$
Diffusion inversed latent start point
Randomly sampled latent start points
Original image
High fidelity original image editing
High accuracy text conditioning
Low fidelity original image editing
High accuracy text conditioning
Text-only conditioned synthesis
No original image

# Null-text Inversion

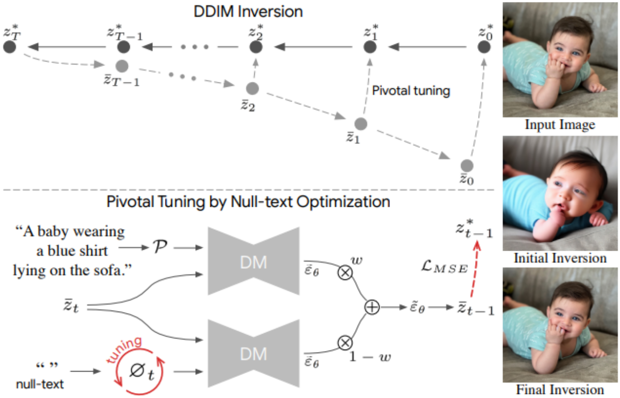



Figure: Null-text Optimization overview [MHA+22]

# Prompt-to-Prompt Image Editing

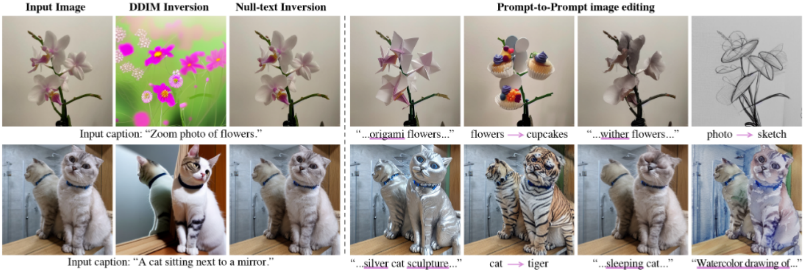
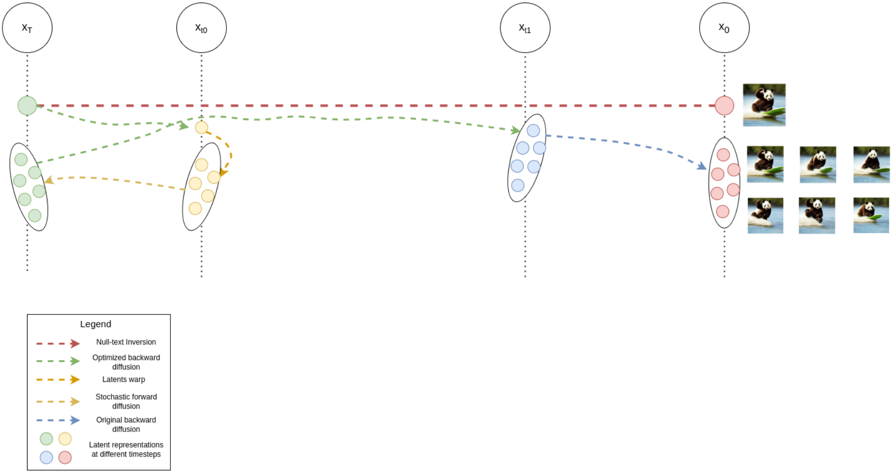



Figure: Naive diffusion inversion vs Null-text Inversion on the left and Prompt-to-Prompt based Image editing on the right.[MHA+22]

# Our method overview



# Dataset


Figure: EPIC-KITCHENS-100 dataset samples [DDF+22]

# Dataset

| | | Hours | Videos | Action Seg. | Unique Narr. | Verb Cls. | Noun Cls. | Action Cls. | Object Masks | Hand BB | Int. Obj |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Source** | Videos from [1] | 54.6 | 432 | 39,432 | 11,423 | 93 | 272 | 2,747 | 35,682,398 | 18,234,678 | 22,156,746 |
| | Extension | 45.4 | 268 | 50,547 | 11,236 | 91 | 266 | 2,900 | 29,987,598 | 12,999,913 | 16,043,057 |
| | **Overall** | **100.0** | **700** | **89,977** | **20,580*** | **97** | **300** | **4,053** | **65,669,996** | **31,234,591** | **38,199,803** |
| **Splits** | Train | 74.7 | 495 | 67,217 | 15,968 | 97 | 289 | 3,568 | 48,896,723 | 23,186,294 | 28,190,446 |
| | Val | 13.2 | 138 | 9,668 | 3,835 | 78 | 211 | 1,352 | 8,714,871 | 4,462,472 | 5,513,884 |
| | Test | 12.1 | 67 | 13,092 | 4,324 | 84 | 207 | 1,487 | 8,058,402 | 3,585,825 | 4,495,473 |

Figure: EPIC-KITCHENS-100 dataset stats [DDF+22]

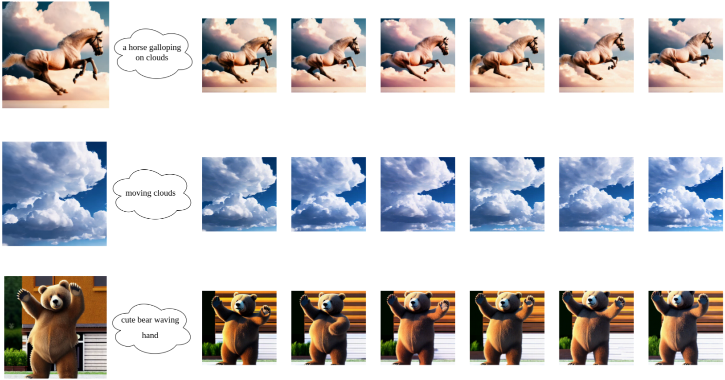
# Quantitative evaluation

- Median 3D-SSIM (3D-Structural Similarity Index)
- Median PSNR (Peak Signal to Noise Ration)
- FVD (Frechet Video Distance)

| | | **Kitchen** | | |
|---|---|---|---|---|
| **Scenario** | **Resolution** | 3D-SSIM ↑ | PSNR ↑ | FVD ↓ |
| MMVG | 128 × 128 | 0.3495 ± 0.1353 | **15.6479 ± 3.438** | **561.68** |
| **Ours** | 128 × 128 | 0.2479 ± 0.081 | 13.9159 ± 2.6998 | 593.04 |
| **Ours** | 512 × 512 | **0.4542 ± 0.078** | 13.6417 ± 2.47 | 600.8 |

Table: Quantitative evaluation comparison between the state-of-the-art MMVG architecture and our proposed method.

# Qualitative evaluation



# Limitations and challenges

Limitations:

- computational resources
- pre-trained backbone

Challenges:

- long video generation
- temporal consistency
- computational complexity
- ethical concerns

# References I


- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray, *Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100*, International Journal of Computer Vision (IJCV) **130** (2022), 33–55.
- Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell, *Tell me what happened: Unifying text-guided video completion via multimodal masked video generation*, CoRR **abs/2211.12824** (2022).
- Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or, *Null-text inversion for editing real images using guided diffusion models*, CoRR **abs/2211.09794** (2022).

# References II


- Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav-Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen, *Dreamix: Video diffusion models are general video editors*, CoRR **abs/2302.01329** (2023).
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, *High-resolution image synthesis with latent diffusion models*, IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 10674–10685.
- Xue Song, Jingjing Chen, Bin Zhu, and Yu-Gang Jiang, *Text-driven video prediction*, CoRR **abs/2210.02872** (2022).