Spójność obiektów w czasie w zadaniu generowania wideo za pomocą sieci neuronowych

Mikołaj Małkiński

24.01.2024

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). **Generating videos with scene dynamics**. Advances in neural information processing systems, 29.

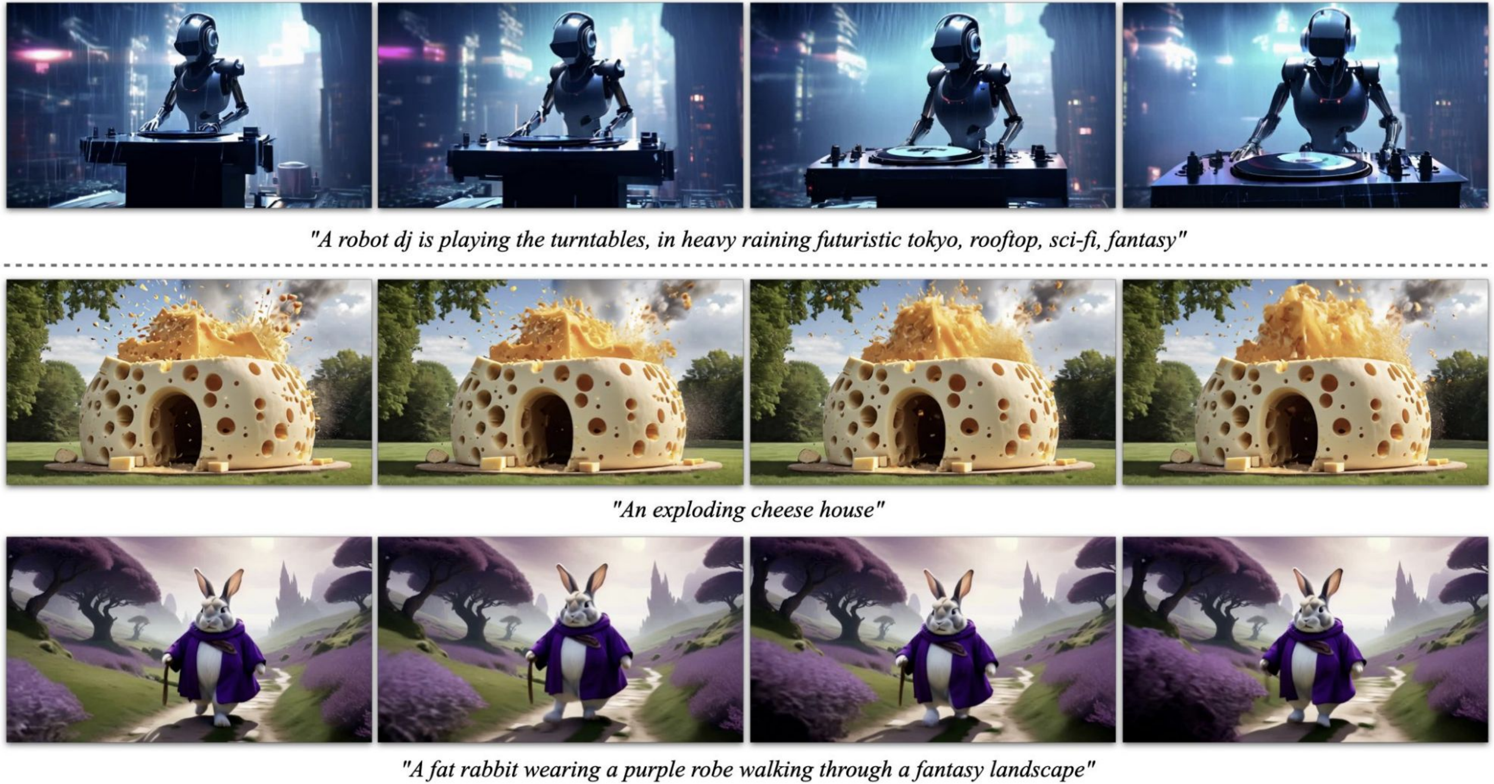
Beach Generated Videos

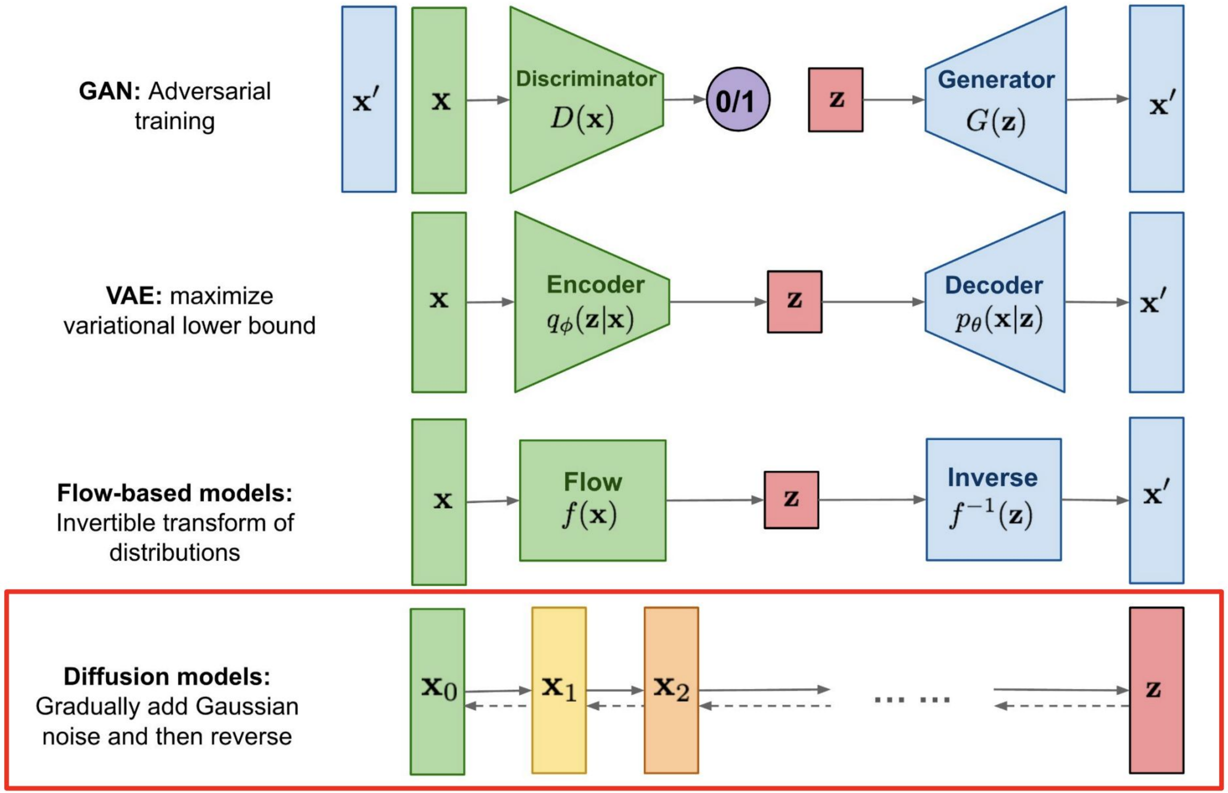
Hospital / Baby Generated Videos

Tulyakov, S., Liu, M. Y., Yang, X., & Kautz, J. (2018). **MoCoGAN: Decomposing motion and content for video generation**. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1526-1535).

(a) Generated by MoCoGAN (b) Generated by VGAN [41]

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., ... & Rombach, R. (2023). **Stable video diffusion: Scaling latent video diffusion models to large datasets**. arXiv preprint arXiv:2311.15127.





Text-to-Image Synthesis with Conditional Diffusion Models, Aman Shrivastava, University of Virginia
https://www.cs.rice.edu/~vo9/cv-seminar/slides/aman-diffusion.pdf

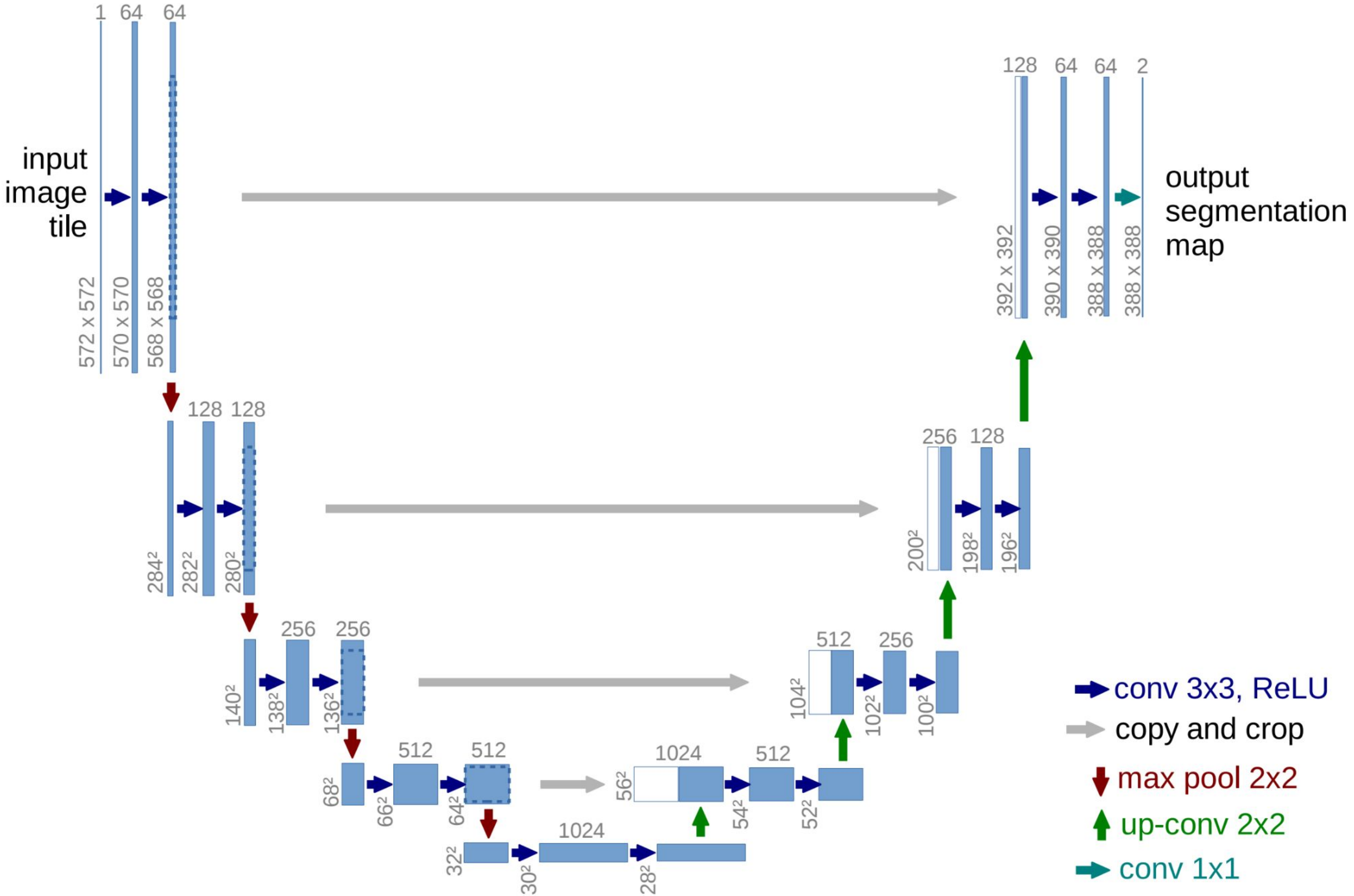
Preliminaries: text-to-image (T2I) models

1. Large models trained on web-scale image-text pairs
2. Diffusion models learn a data distribution by gradually denoising a normally distributed variable, i.e. “noise”, to generate the output
3. Pixel diffusion models denoise in the pixel space
4. Latent diffusion models denoise in the latent space
5. Conditional diffusion models denoise conditioned on the input c

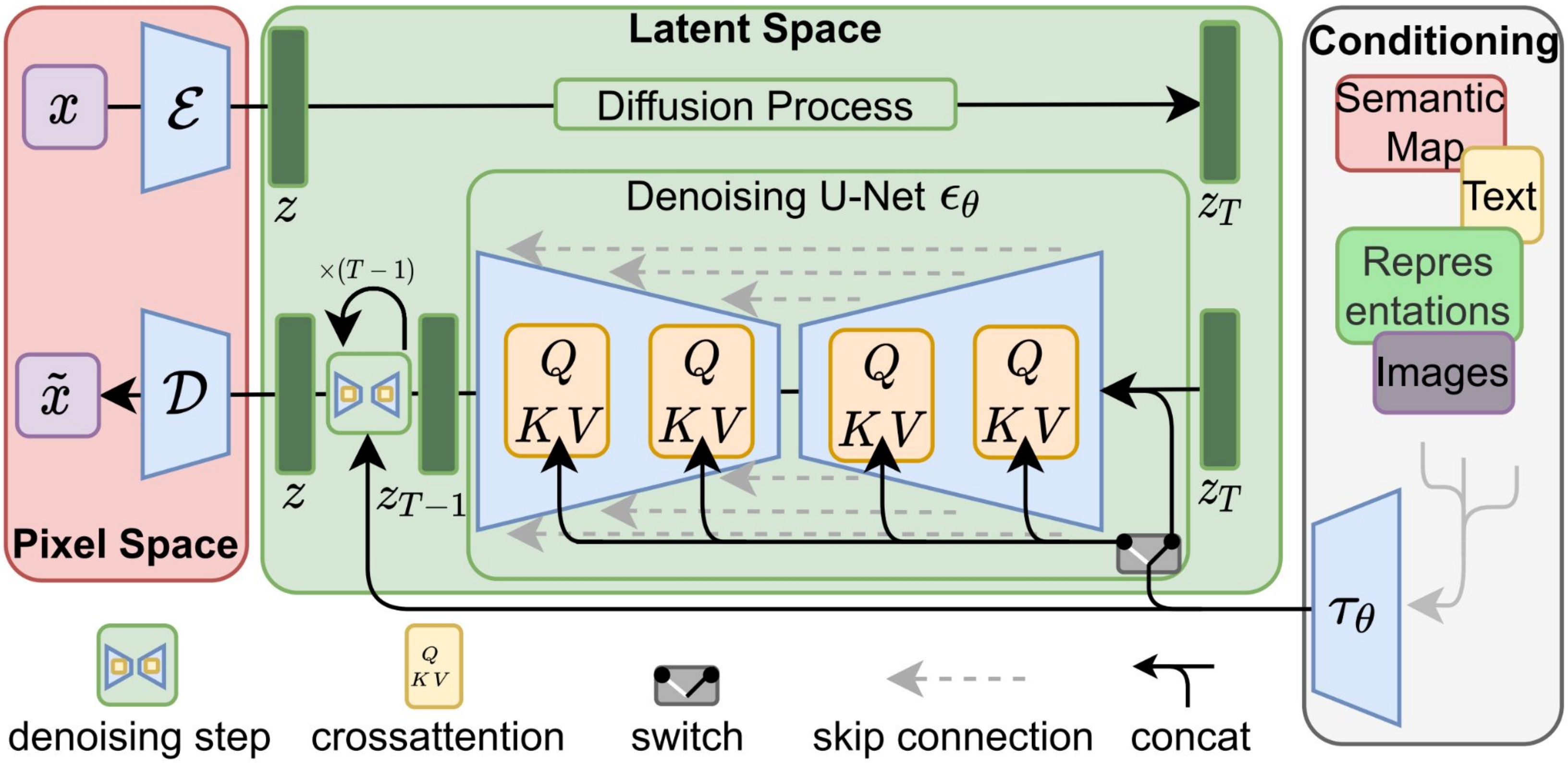
Diffusion models: Denoising Diffusion Probabilistic Models
https://pages.mini.pw.edu.pl/~mandziukj/2022-11-30.pdf

Preliminaries: text-to-image (T2I) models

Nichol, A. Q., & Dhariwal, P. (2021, July). **Improved denoising diffusion probabilistic models**. In International Conference on Machine Learning (pp. 8162-8171). PMLR.



Ronneberger, O., Fischer, P., & Brox, T. (2015). **U-net: Convolutional networks for biomedical image segmentation**. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.



Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). **High-resolution image synthesis with latent diffusion models**. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

Text-to-video (T2V) models

1. Adapt T2I models by using video-text pairs
2. Use diffusion models to generate all frames at once
3. Autoregressive generation

Challenges in T2V generation

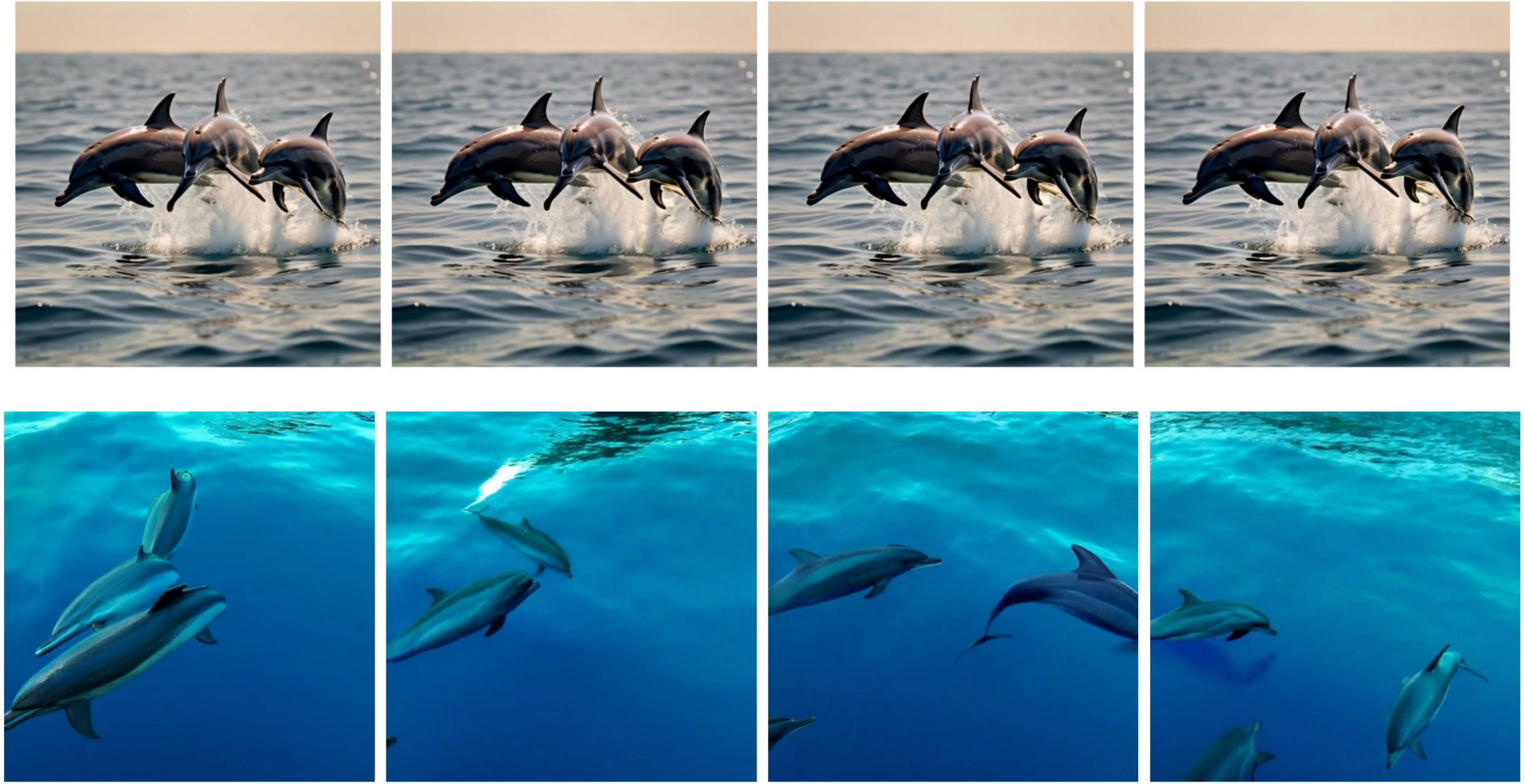
1. The scarcity and weak relevance of text-video datasets
2. Big computational cost of training from scratch
3. Autoregressive video generation is prohibitively expensive
4. Higher spatio-temporal output space as compared to T2I
5. Weak conditioning, text-only
6. Fine-tuning T2I models decreases quality due to lower diversity of T2V data

Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S. S., ... & Misra, I. (2023). **Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning**. arXiv preprint arXiv:2311.10709.

Motivation

1. Utilize a pre-trained T2I model (latent diffusion model with frozen weights)
2. Explicitly generate the starting frame
3. Condition on the text and the initial generated image

Dolphins jumping in the ocean – w/o image conditioning



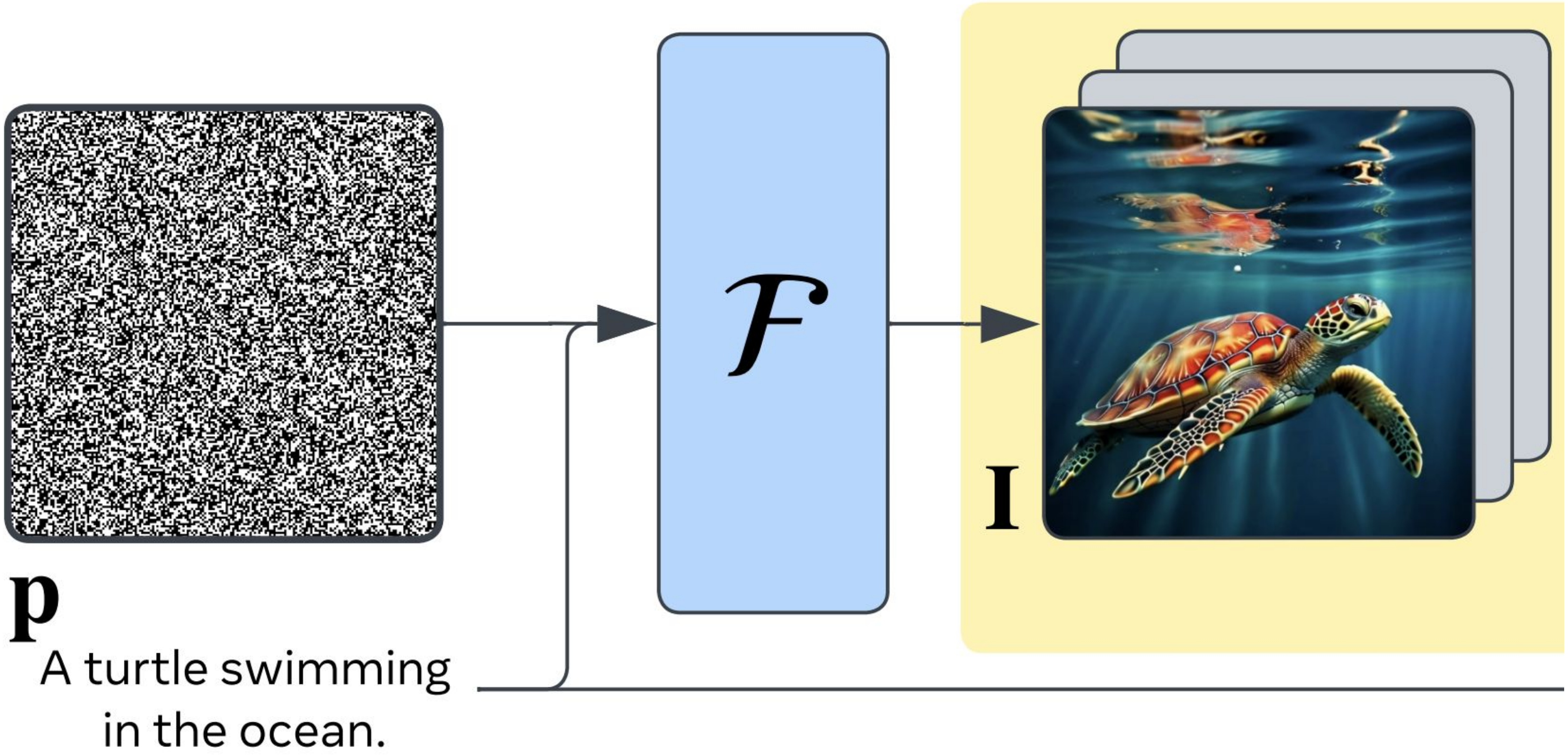
Unicorns running along a beach – w/o image conditioning

T2I model

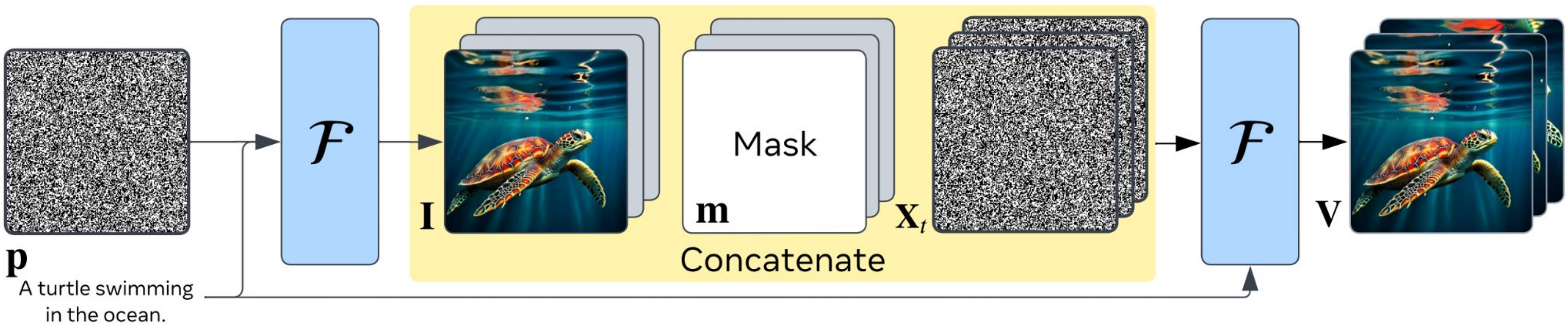
Dai, X., Hou, J., Ma, C. Y., Tsai, S., Wang, J., Wang, R., ... & Parikh, D. (2023). **Emu: Enhancing image generation models using photogenic needles in a haystack**. arXiv preprint arXiv:2309.15807.

1. Pre-train a on 1.1B image-text pairs
2. Latent diffusion model
3. U-Net backbone with 2.7B parameters
4. Condition on text embedded with CLIP and T5-XL
5. Fine-tune with a few thousand high-quality images
6. Emu achieves win rate of 82.9% compared to the pre-trained only counterpart

Step 1: Generate the starting frame with the T2I model



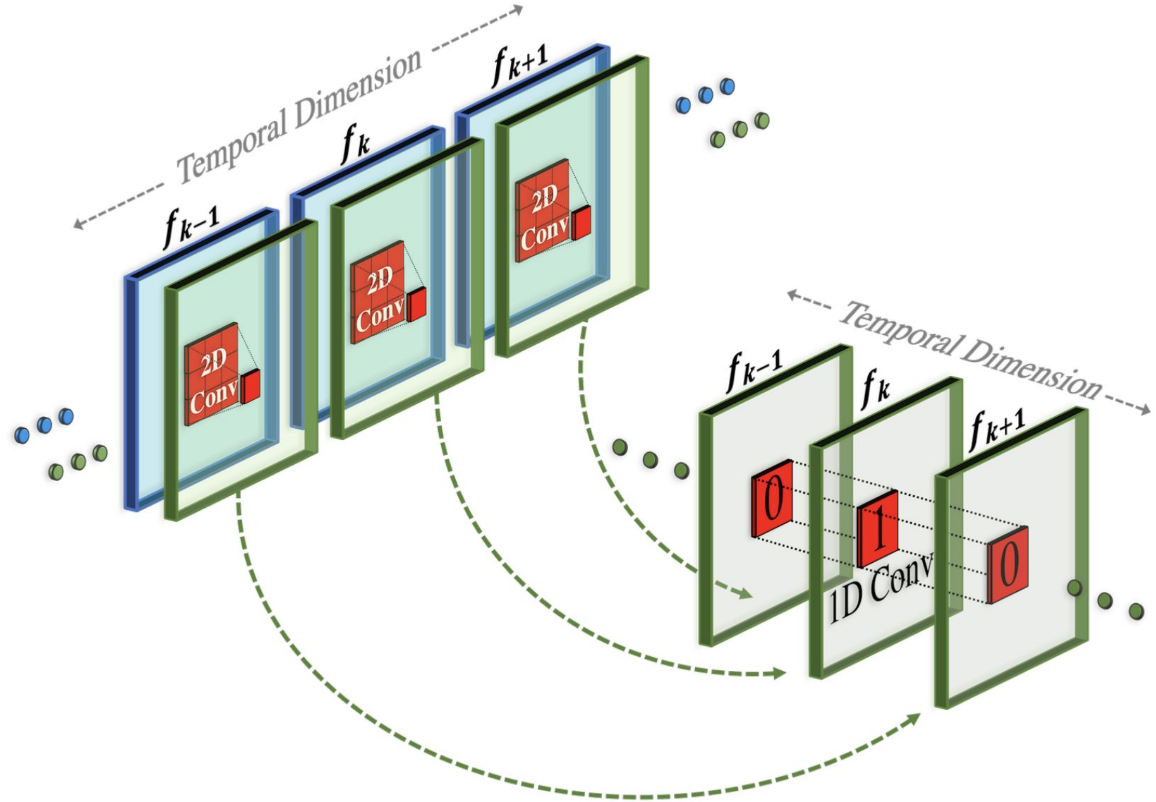
Step 2: Predict T subsequent frames



T2V model

1. Add new learnable parameters
    a. 1D temporal convolution after every spatial convolution
    b. 1D temporal attention layer after every spatial attention layer
    c. 1.7B of new learnable parameters
    d. 4.3B parameters including the T2I model
2. T2I layers are kept frozen and applied to each frame independently
3. New learnable zero-initialised channels are added to the UNet’s input layer
4. Identity initialisation for temporal parameters (improves convergence by 2X)
5. The model produces videos with T = 8 or 16 frames of 512px resolution

Pseudo-3D Convolutional Layer



$$Conv_{P3D}(h) := Conv_{1D}(Conv_{2D}(h) \circ T) \circ T$$

Depth-wise separable convolution

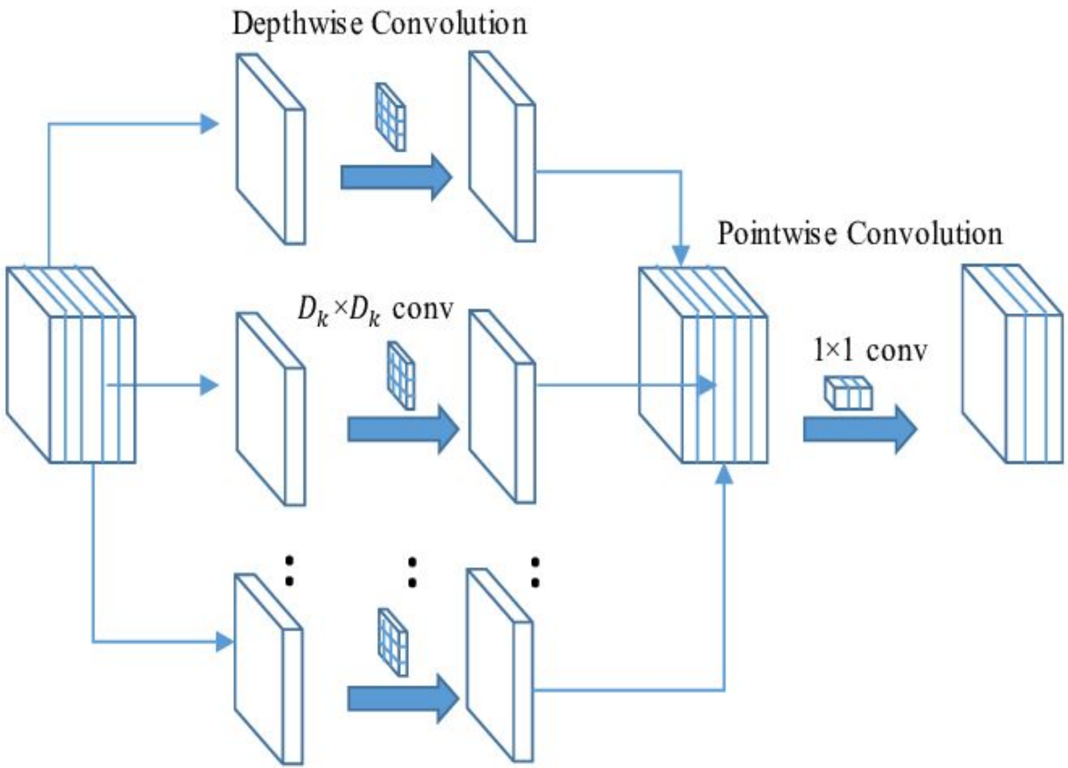
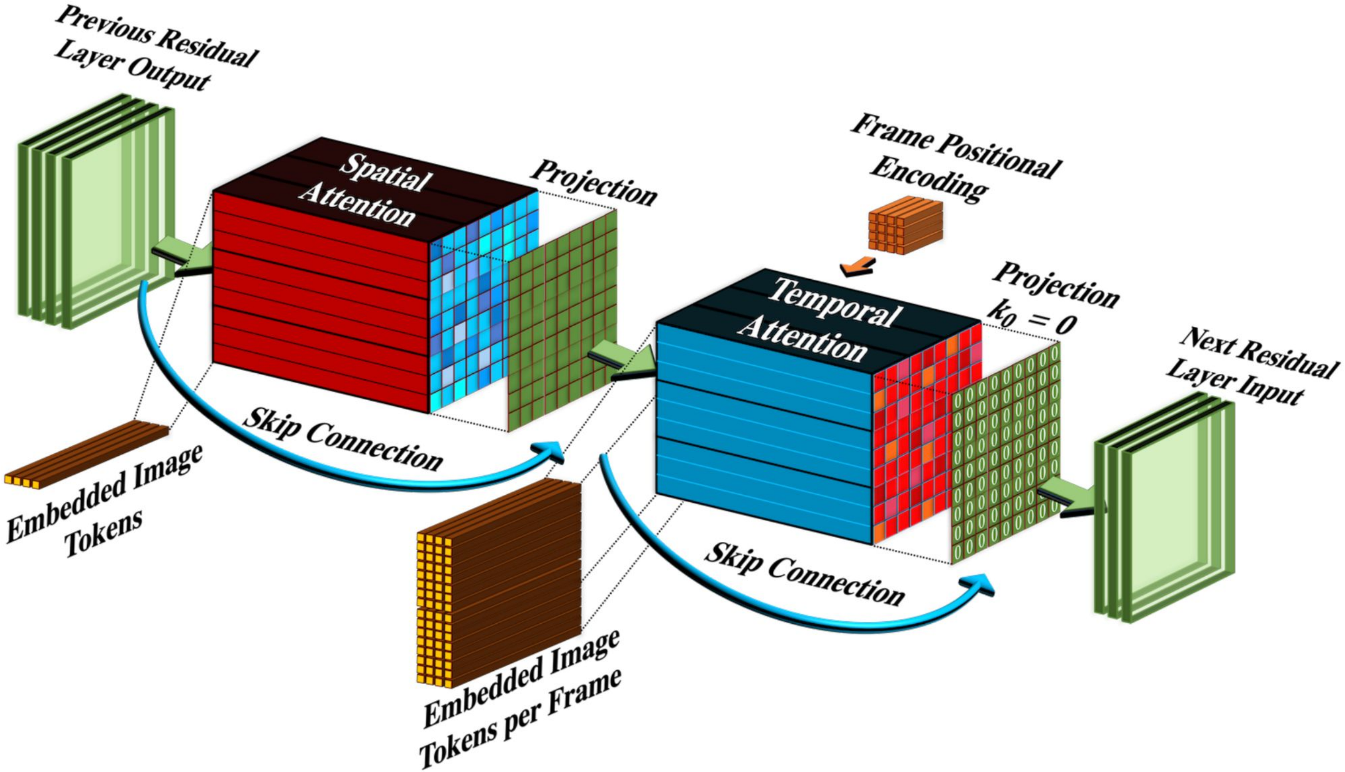


Image from: Hossain, D., Imtiaz, M. H., Ghosh, T., Bhaskar, V., & Sazonov, E. (2020, July). Real-time food intake monitoring using wearable egocnetric camera. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 4191-4195). IEEE.

Pseudo-3D Attention Layer



$$ATTN_{P3D}(h) = unflatten(ATTN_{1D}(ATTN_{2D}(flatten(h)) \circ T) \circ T)$$

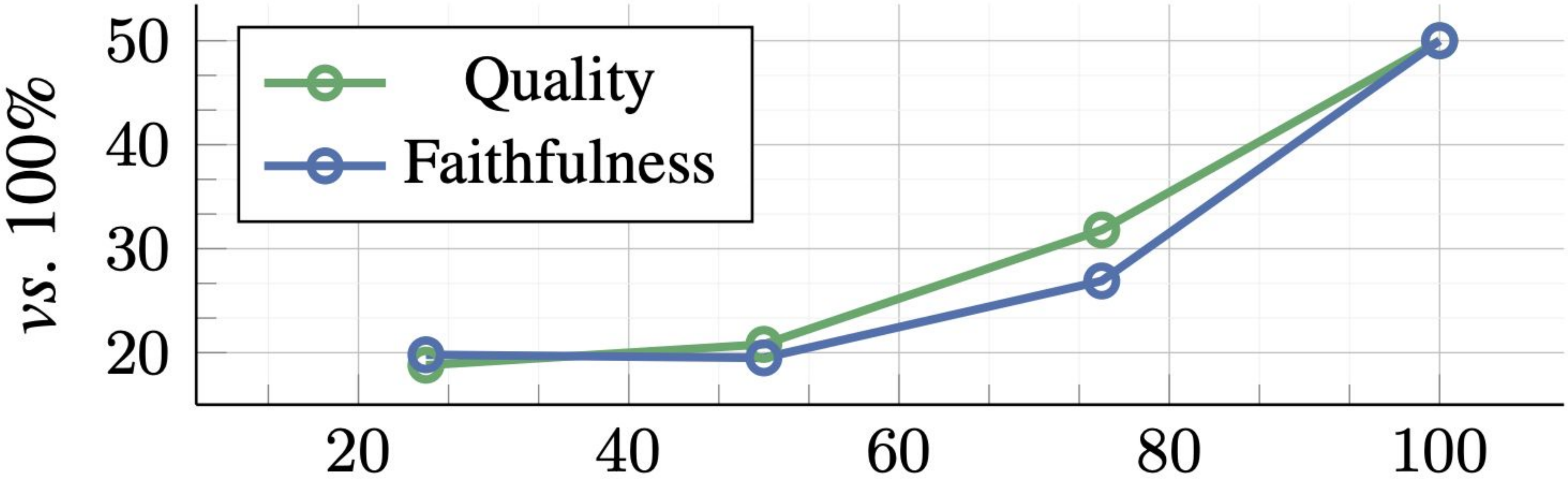
Dataset

1. 34M licensed video-text pairs
2. Videos range from 5 to 60 seconds
3. Covers a variety of natural world concepts
4. Unfiltered

Training: Multi-stage multi-resolution

1. Uses video clips of 1, 2 or 4 seconds sampled at 8fps or 4fps
2. First stage:
    a. 70K iterations
    b. 256px 8fps 1s videos
    c. Classical noise schedule (from LDM, Rombach et al. 2021)
    d. Smaller spatial resolution reduces per-iteration time by 3.5x
3. Second stage
    a. 15K iterations
    b. 512px 4fps 2s videos
    c. Zero terminal-SNR
4. Optional third stage
    a. 25K iterations
    b. 512px 4fps 4s videos
    c. Increases video duration

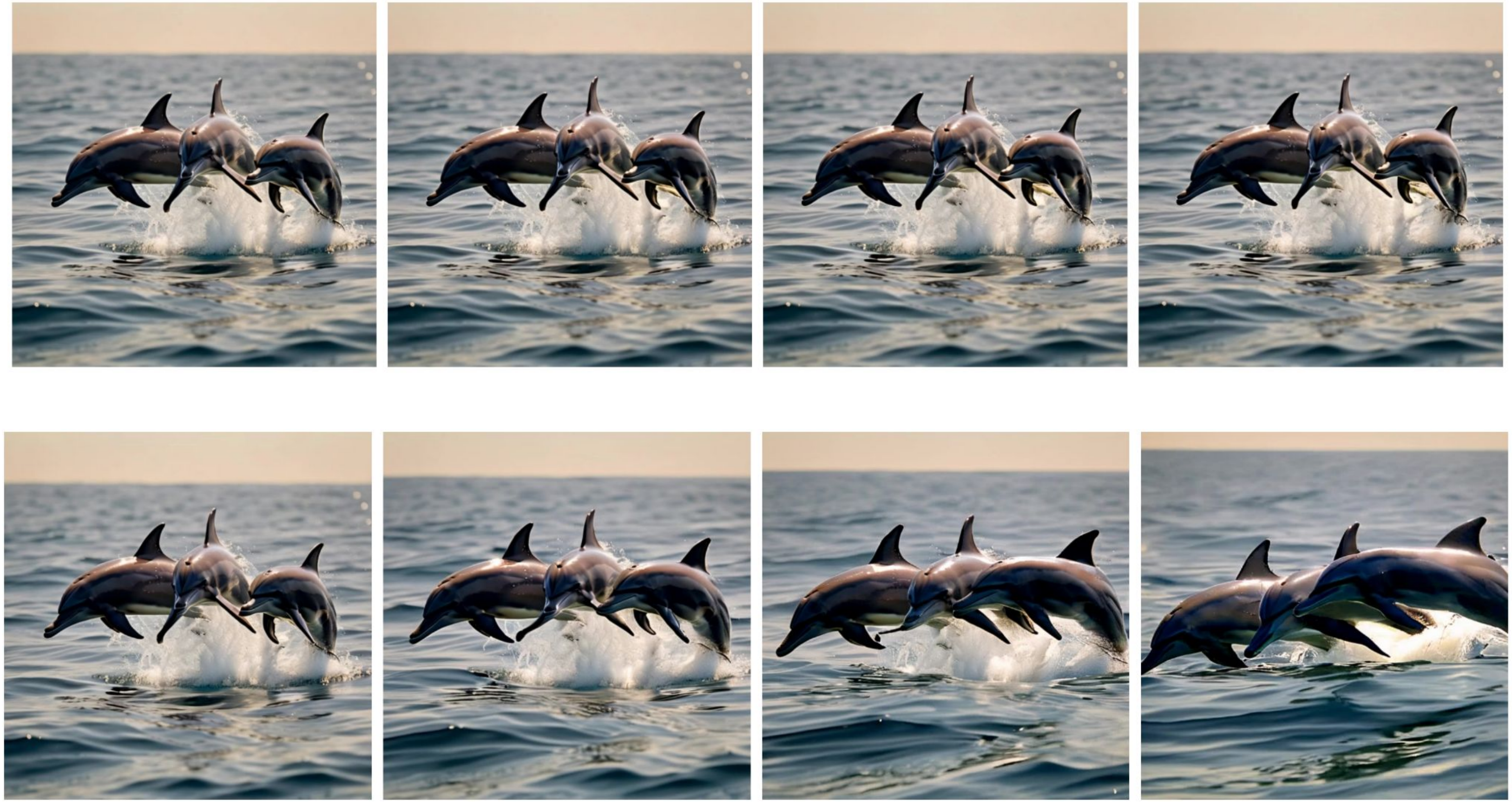
Performance vs. training iterations in the low-resolution stage



Training: Fine-tuning for better quality

1. Fine-tuning for better quality
2. Small subset of high motion high quality videos
3. 1.6K videos from the training set
4. Filtering based on automatic metrics (e.g. CLIP similarity between the video’s text and the first frame)

Dolphins jumping in the ocean – w/o HQ Finetune

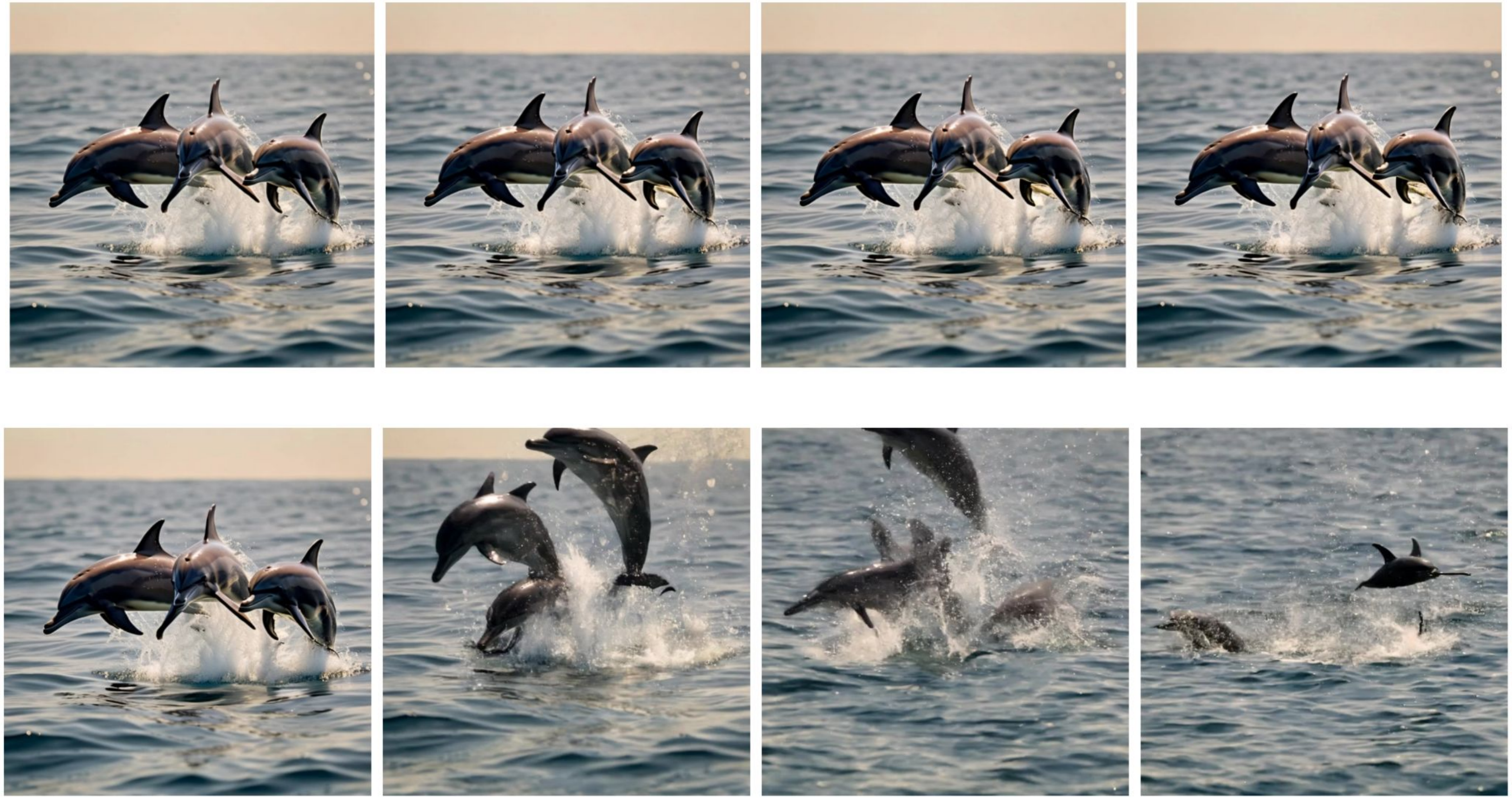


Unicorns running along a beach – w/o HQ Finetune

Training: Zero terminal-SNR noise schedule

1. At training the noise schedule has non-zero signal-to-noise (SNR) ratio even at the terminal diffusion time step N
2. At test-time, however, the initial noise has 0 SNR
3. This issue is exacerbated in the video domain, as videos have spurious pixels across space and time
4. To mitigate, the noise schedule is scaled so that SNR in the final noised input is 0

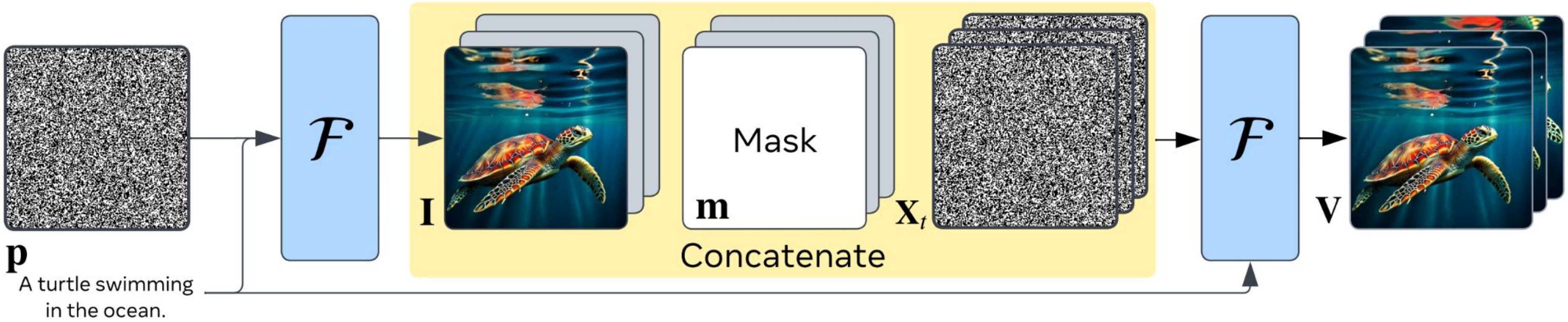
Dolphins jumping in the ocean – w/o Zero SNR



Unicorns running along a beach – w/o Zero SNR

Interpolation model – analogous to the T2V model

1. Initialised from the video model F and only the temporal parameters are fine-tuned
2. Takes 8 frames as input
3. Outputs 37 frames at 16fps as output



Inference

1. The T2I model is run without the temporal layers to generate the initial image
2. The T2V model generates the video frames
3. Interpolation model increases the frame rate
4. All models are implicitly conditioned on the text due to the underlying T2I model

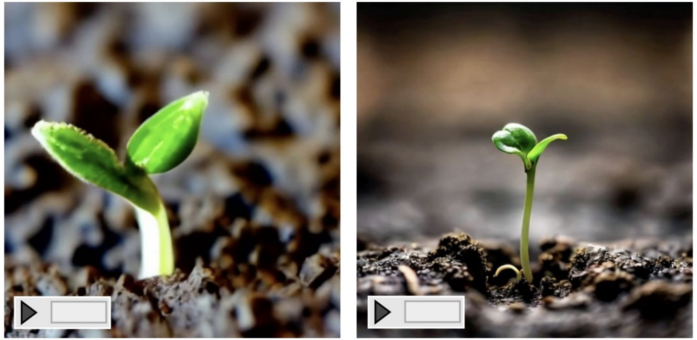
Evaluation: Human preference tests

1. Robust human evaluation scheme, where evaluators are asked to JUstify their choICE (JUICE) in the pairwise comparisons
2. Pre-defined justification reasons
   a. Quality: pixel sharpness, motion smoothness, recognisable objects/scenes, frame consistency, amount of motion
   b. Faithfulness: spatial text alignment, temporal text alignment
3. Win-rate in terms of quality and faithfulness (alignment of the generated video to the text prompt)
4. Majority vote from 5 evaluators for each comparison

Human evaluations

Which video do you prefer?

Video A Video B



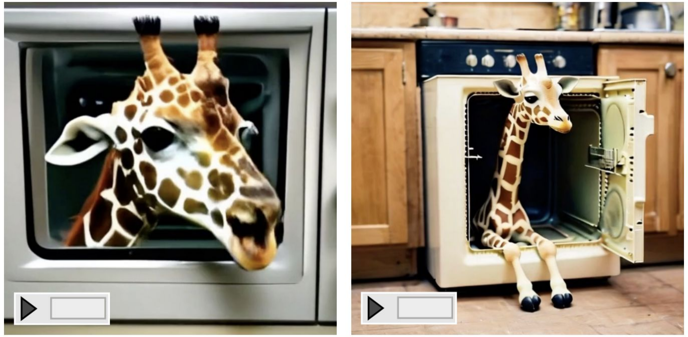
Which factors contributed towards making this choice?
(Select all that apply)

- ❑ Motion smoothness
- ❑ Object/scene consistency
- ❑ Pixel sharpness
- ❑ Recognizable objects/scenes
- ❑ Amount of motion

(a) Video Quality

Which video aligns better with the text prompt?

Video A Video B



A giraffe underneath a microwave.

Which factors contributed towards making this choice?
(Select all that apply)

- ❑ Spatial text alignment
- ❑ Temporal text alignment

(b) Video-Text Faithfulness

Ablation study – preference on adopting a design decision

Method	Q	F
Factorized	70.5	63.3

(a)

Method	Q	F
Zero SNR	96.8	88.3

(b)

Method	Q	F
Multi-stage	81.8	84.1

(c)

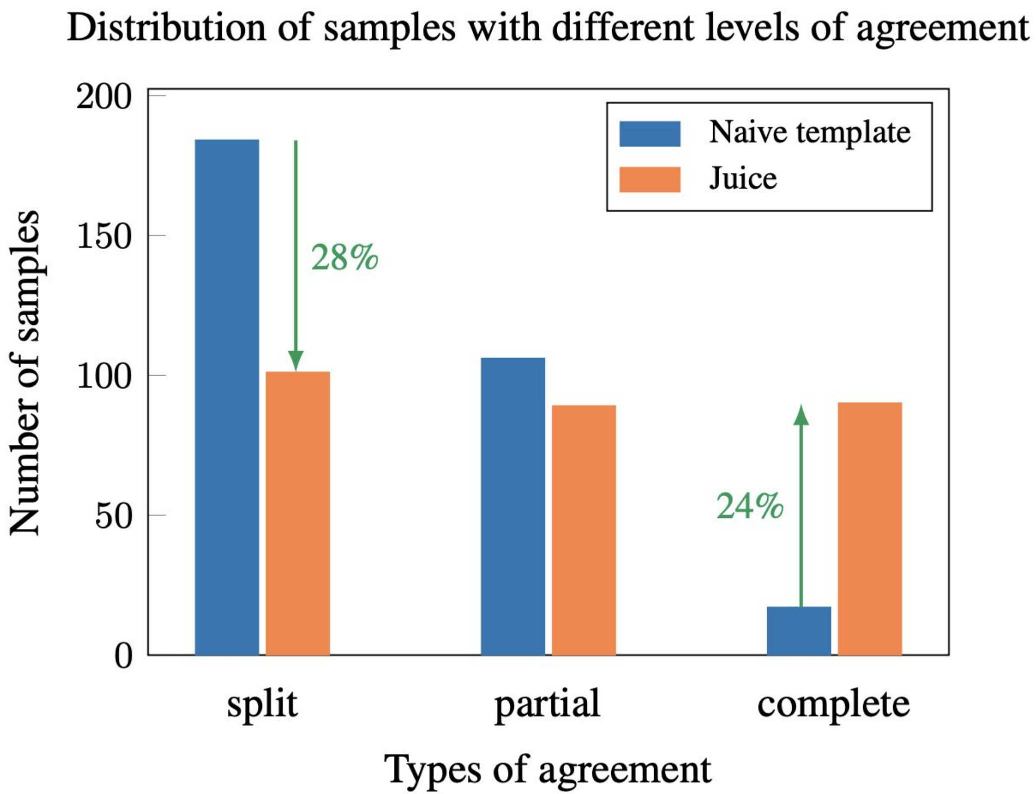
Method	Q	F
HQ finetuned	65.1	79.6

(d)

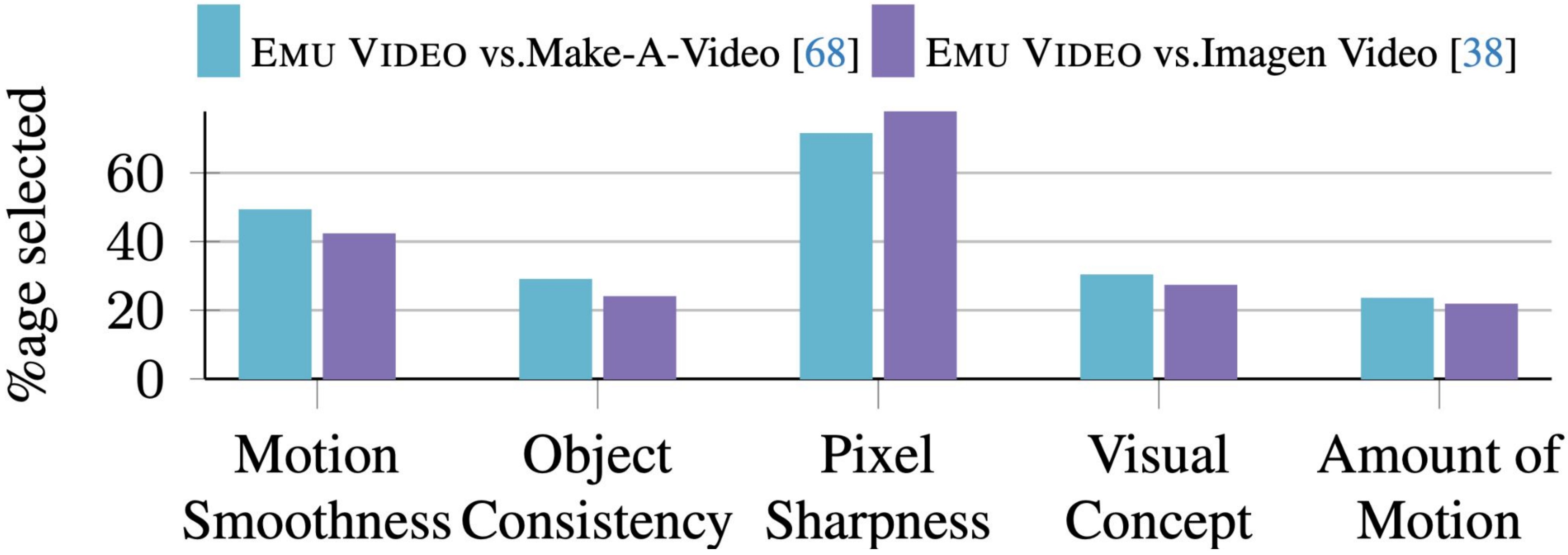
Method	Q	F
Frozen spatial	55.0	58.1

(e)

Human agreement in Emu Video vs. Make-A-Video



Why human evaluators prefer EmuVideo?



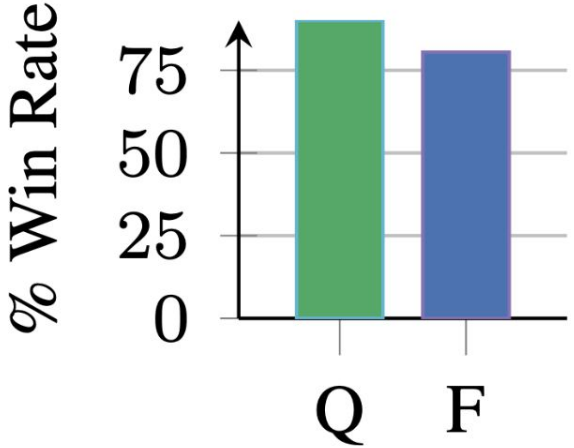
Evaluation: Automated metrics

1. Faithfulness (CLIP-Text)
2. Temporal coherency (CLIP-Image)
3. Temporal coherency (Pixel-MSE)

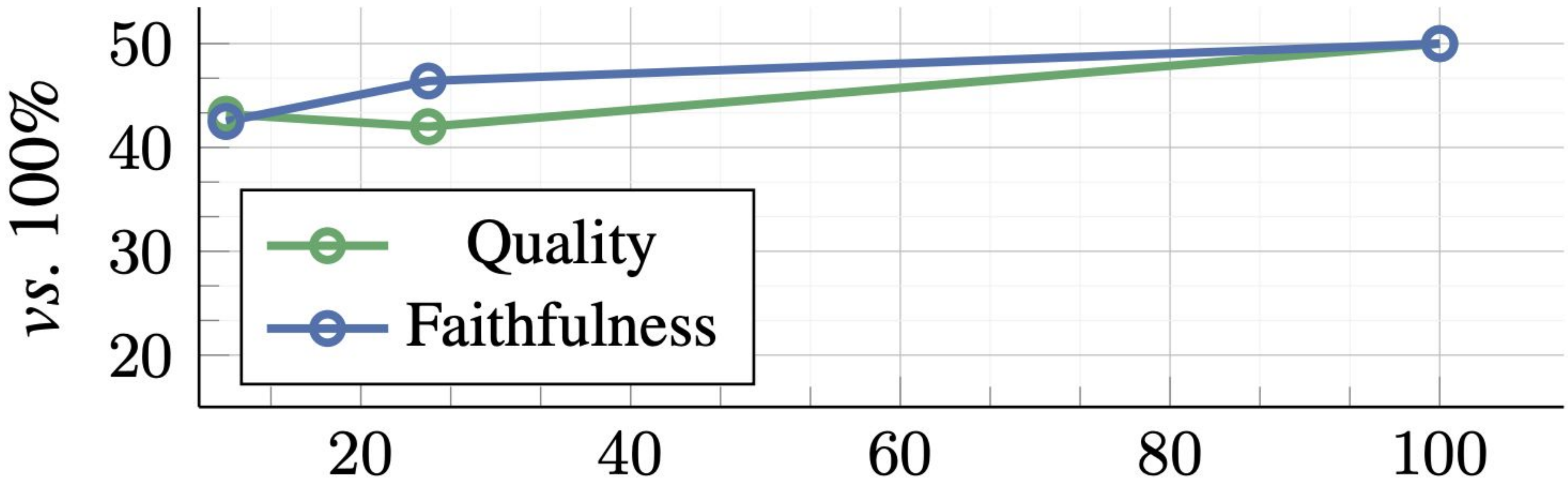
Automated metrics vs. Human evaluation

Method	Automated FVD ↓	IS ↑
MagicVideo [88]	655.0	-
Align Your Latents [7]	550.6	33.5
Make-A-Video [68]	367.2	33.0
PYOCO [30]	355.2	47.8
EMU VIDEO	606.2	42.7

Human Evaluation
vs. Make-A-Video



Performance vs. training data

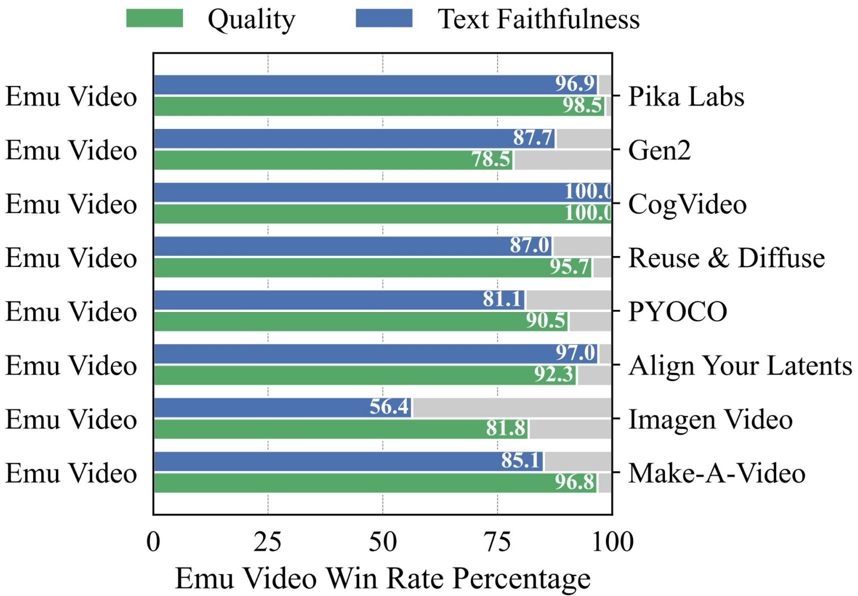


Evaluation: Strong retrieval baseline

1. A nearest neighbor baseline retrieves videos from the training set (34M videos)
2. Relies on the text's CLIP similarity to the training prompts
3. Human evaluators prefer EmuVideo over real videos (81.1% in Faithfulness)

Evaluation: Commercial solutions

1. The models behind commercial solutions are often kept private and only examples (probably the best ones) of their generations are shared
2. Reuse the text prompt and the input image to generate a video



Conclusions

1. Stronger conditioning of image and text shifts the task towards predicting how an image evolves into the future
2. Key design decisions:
   a. Multi-stage multi-resolution training
   b. High-quality fine-tuning
   c. Adjusted noise schedules for diffusion
   d. No need for a deep cascade of models

Future work

1. Stronger text conditioning
    a. During training, they use a video frame sampled from real videos
    b. During inference, the initial frame is generated with a T2I model
    c. The generated image, however, may not be representative of the text prompt
2. Autoregressive generation
    a. The generated videos are rather short (16 frames)
    b. Longer videos require interpolation between frames