Zero-shot Motion Control for Multiple Objects Using Video Diffusion Models

Ryota Kaji, Keiji Yanai
The University of Electro-Communications



Introduction

- Many high-quality video generation models have been proposed.



Generation video with Open AI Sora

Introduction

- While generative quality is high, control over motion and other aspects is often limited.



- There are many situations where we want to control multiple objects.

Introduction

- While generative quality is high, control over motion and other aspects is often limited.
- There are many situations where we want to control multiple objects.



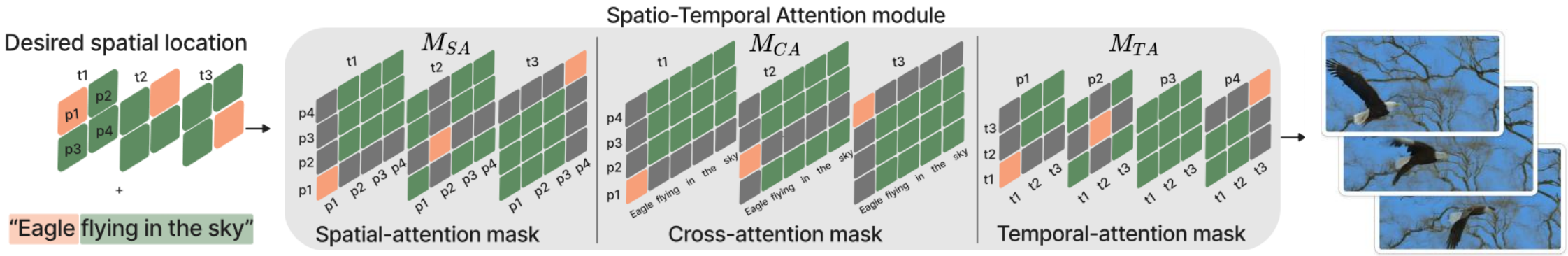
Multiple videos need to be generated and selected to create the desired video output.



Proposing a Zero-Shot Method for Controlling Multiple Objects

Related Works: PEEKABOO [CVPR'24]

- A method to control generated videos in a zero-shot manner by manipulating attention maps of pre-trained video generation models.



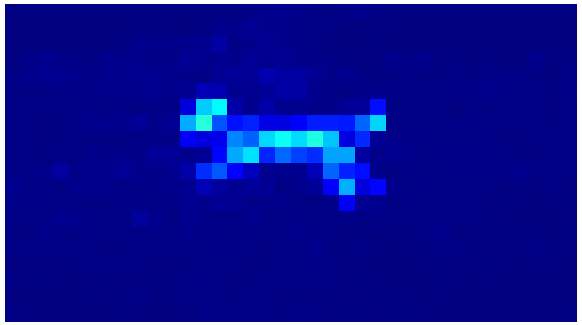
[1] Jain et, al. PEEKABOO: Interactive Video Generation via Masked-Diffusion

Related Works: PEEKABOO [CVPR'24]

- Attention Map



Generation video



Attention Map of "dog"

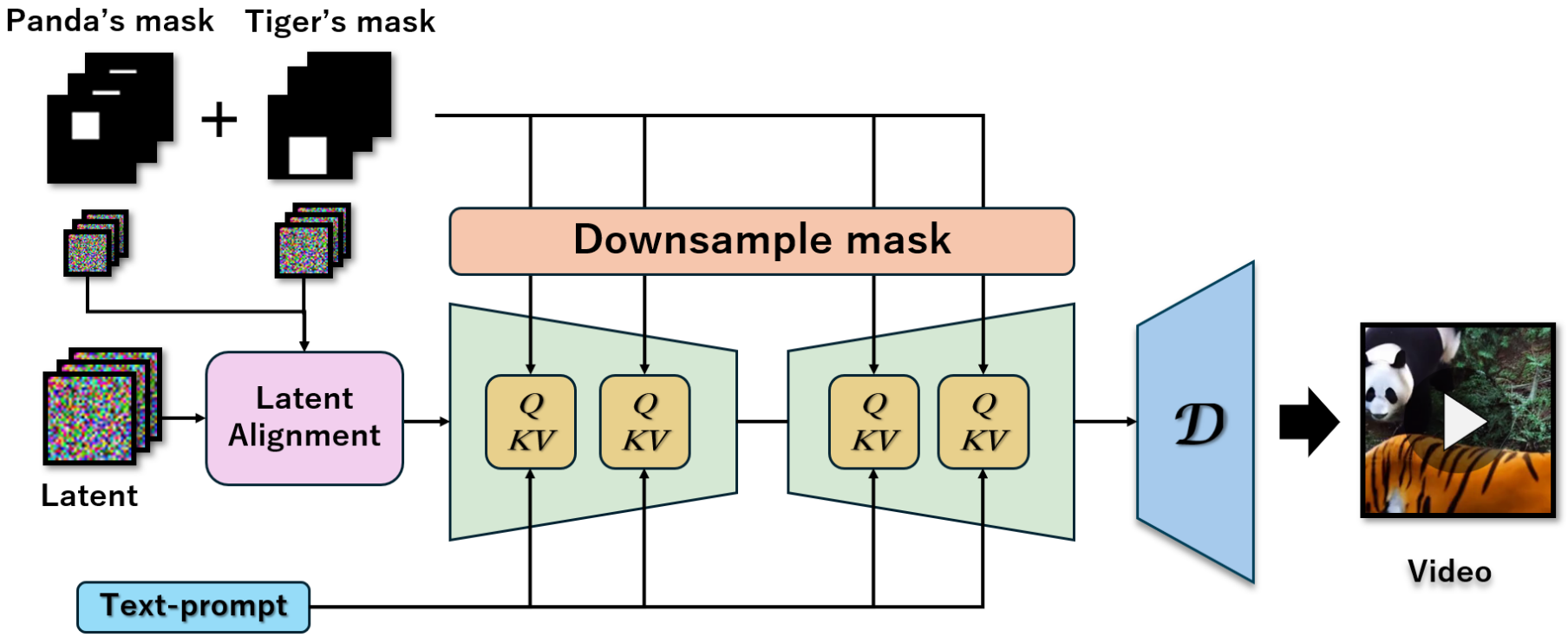
- Computing Masked Attention for controlling areas of focus

$$\text{MaskedAttention}(Q, K, V, M) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + M\right)V$$

$$\text{where} \quad M[i, j] = \begin{cases} -\infty & \text{if } M[i, j] = 0 \\ 0 & \text{if } M[i, j] = 1 \end{cases}$$

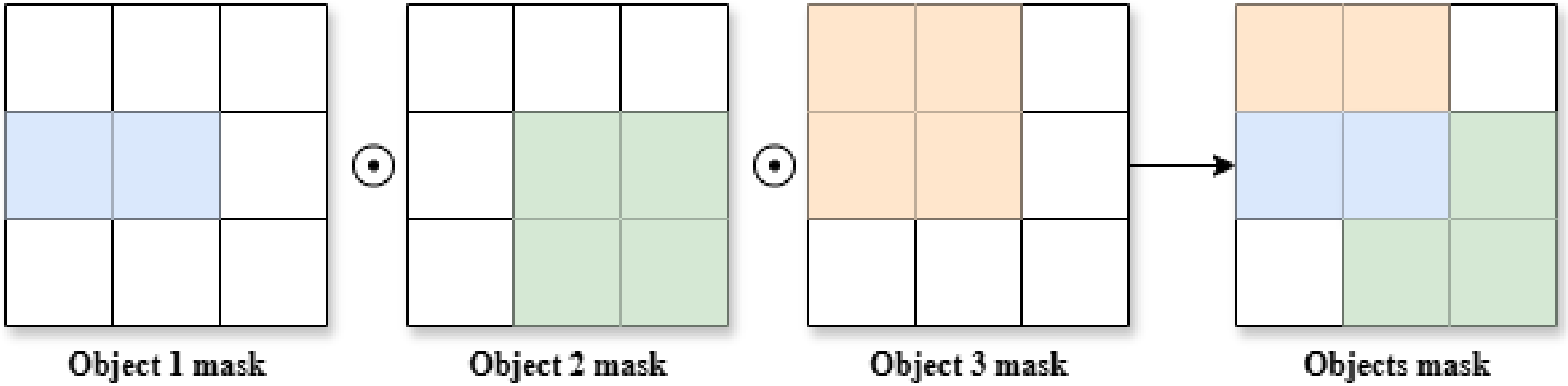
Overview of the proposed method

1. Priority Multi Masked Attention
2. Latent Alignment



Priority Multi Masked Attention

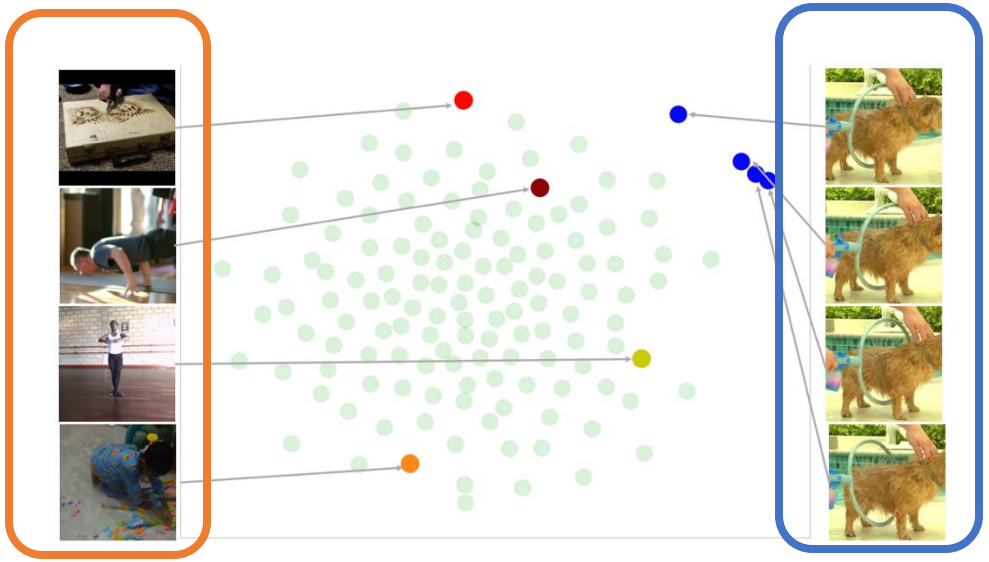
- Our method designs Attention Masks that also separate objects from each other within the foreground.



- Modified so that objects do not reference each other within each attention mechanism.
- Through this design, each object has an independent generation process.

Latent Alignment

- Visually similar between frames of the same video, and their latent variables are also similar.



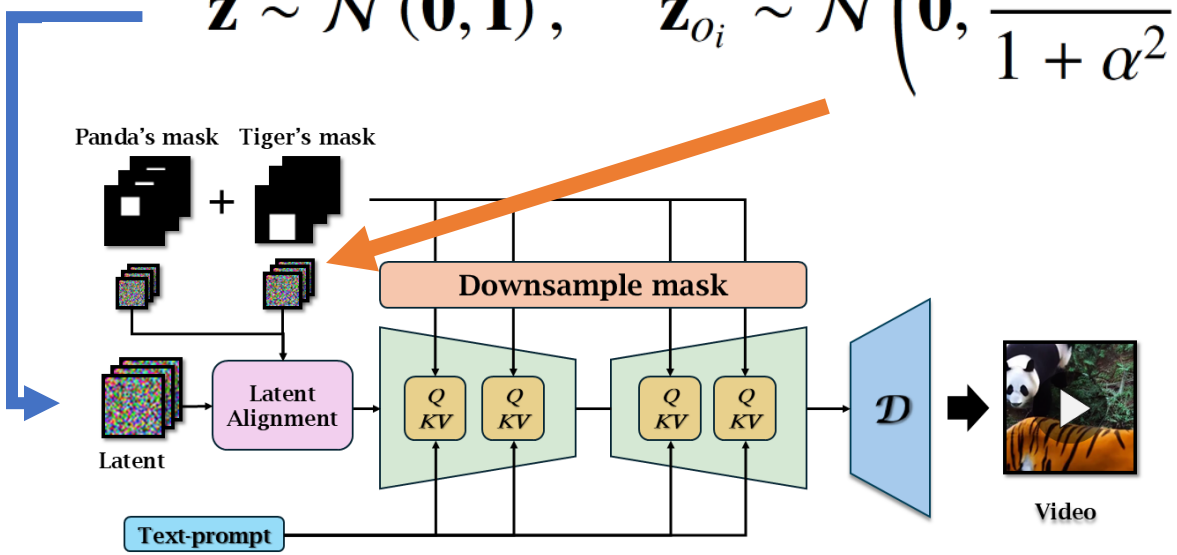
[3] Ge et, al. Preserve your own correlation: A noise prior for video diffusion models

- Use similar latent variables for each object position between frames

Latent Alignment

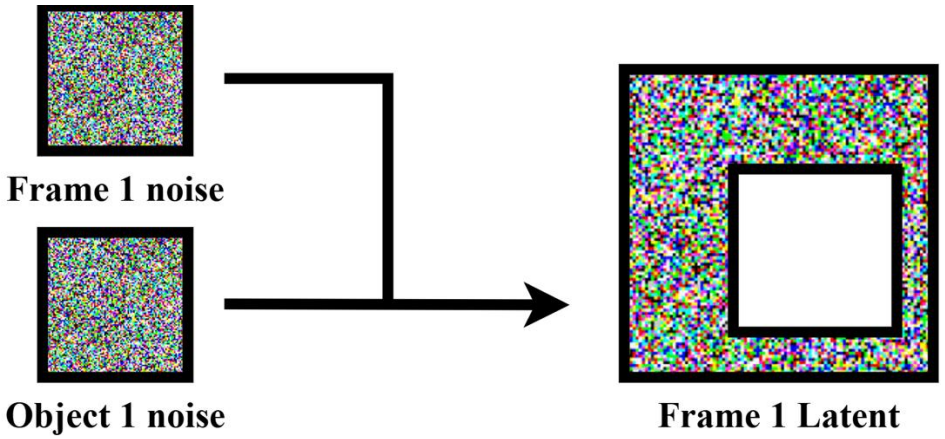
- Sampling each object-specific latent variable separately from the initial latent variable at generation.

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{z}_{o_i} \sim \mathcal{N}\left(\mathbf{0}, \frac{\alpha^2}{1+\alpha^2}\mathbf{I}\right)$$



Latent Alignment

- Perturbation is added to the intrinsic noise in each frame



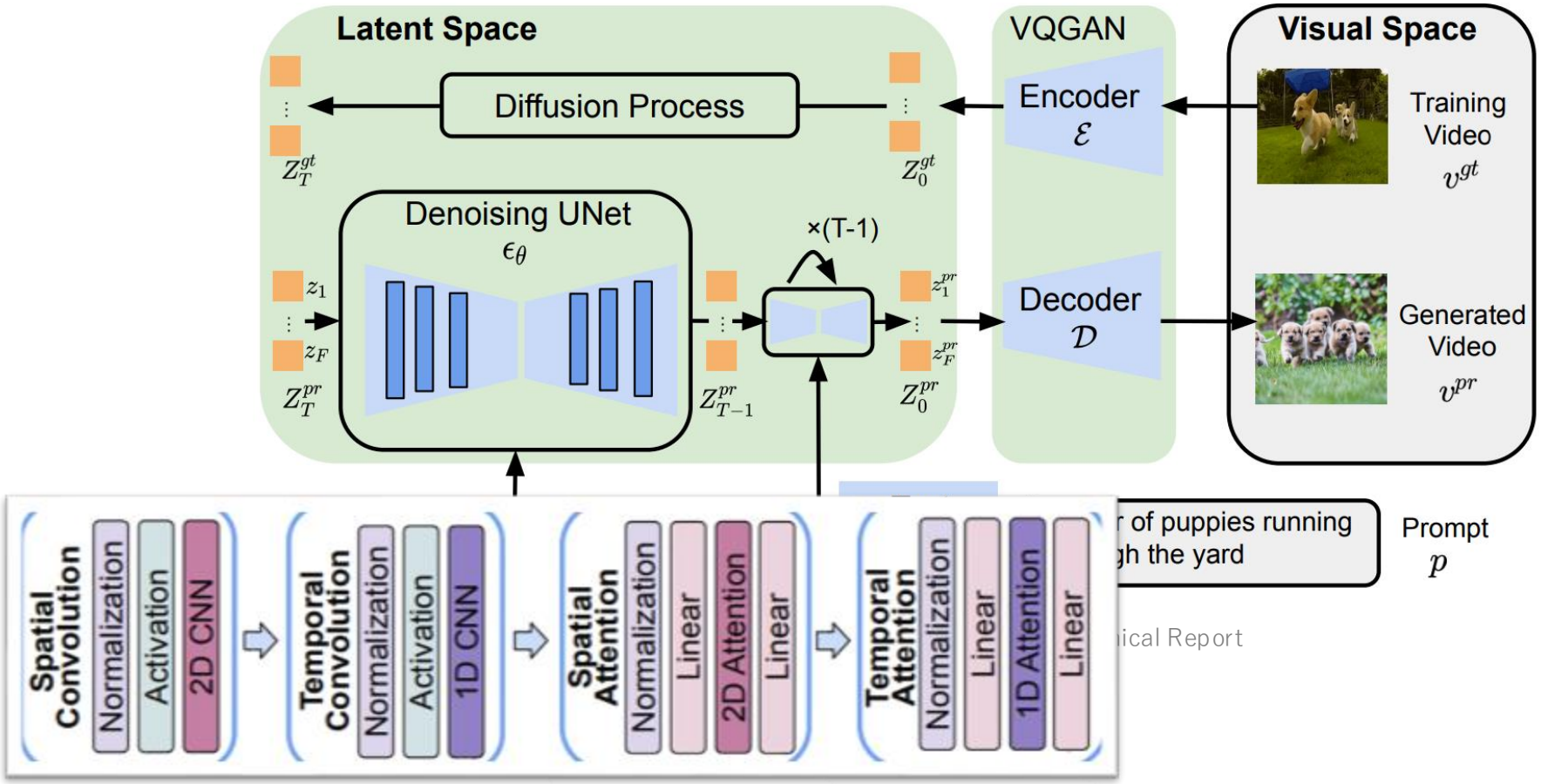
$$\epsilon^f_{o_i} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{1+\alpha^2}\mathbf{I}\right), \quad \mathbf{z}^f_{o_i} = \mathbf{z}_{o_i} + \epsilon^f_{o_i}$$

$$\mathbf{z}[\hat{M}^{o_i}_{input}] = \mathbf{z}_{o_i}$$

- α is a hyperparameter that controls strength of conditioning.
  - $\alpha = 0$: Each noise is independent.
  - $\alpha \to \infty$: Each noise is equal.

Experiments

- Text-to-Video Model (ModelScope/ZeroScope)

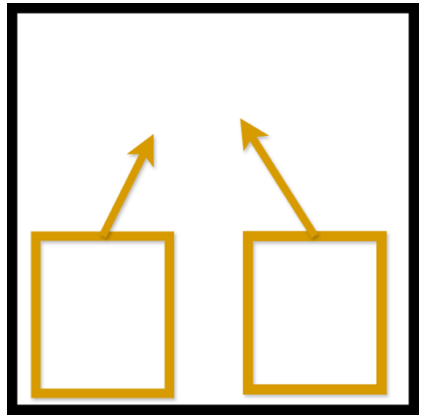


ical Report

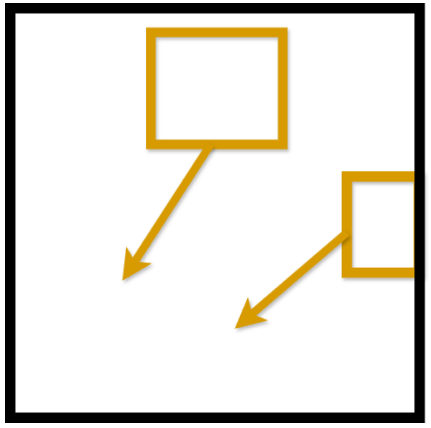
Experiment settings

- Create dataset for evaluation using LLM
- Text-to-Video: ModelScope/ZeroScope
- Masked Attention performs only the first 2 of 40/50 generation steps
- Parameter α is set to 0.2/0.18 from the experiment
- All experiments were performed with NVIDIA A100 80GB x 1

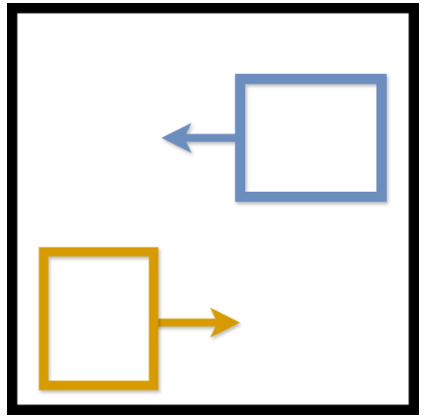
ModelScope Results



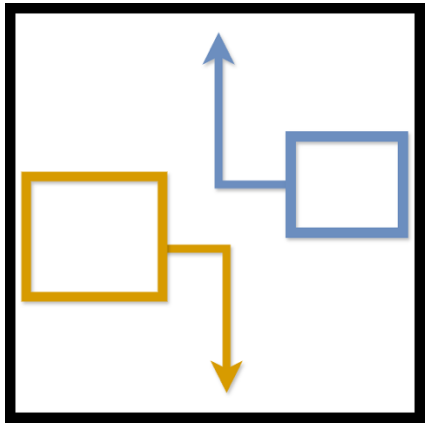

two frogs jumping up to catch a fly




ducks in a pond



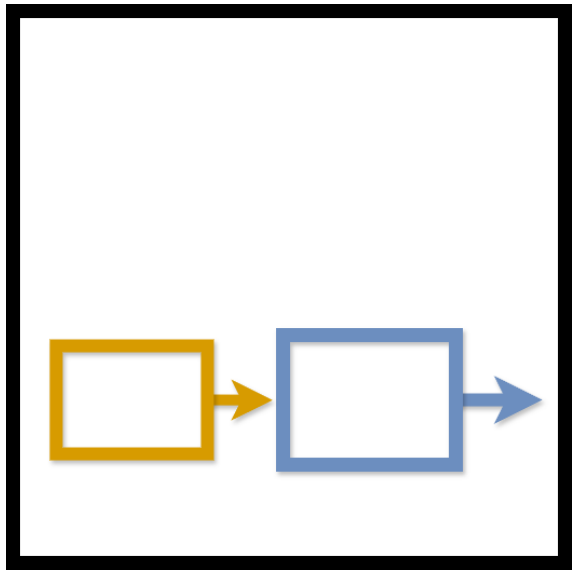

cruiser and fishing boat pass each other at the sea




Panda and tiger walking in the forest

ZeroScope Results

Bounding Box



PEEKABOO



Ours



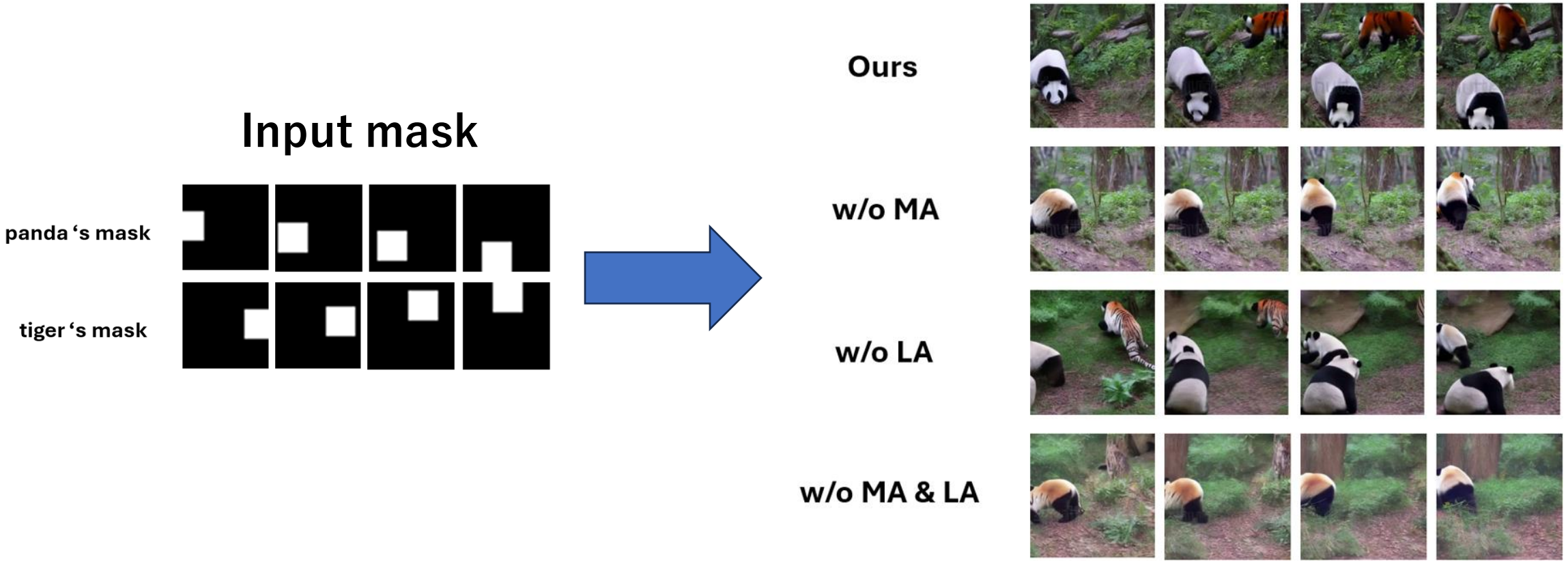
A panda and a tiger walking in the forest

Comparison of Quantitative Evaluation

- Generated by ZeroScope using our evaluation dataset
  - mIoU: IoU between input BBox and generated object's BBox
  - CLIPSim: CLIP similarity between each frame and text prompt
  - T-Cons: CLIP similarity between feature of consecutive frame

Method	mIoU (↑)	CLIPSim (↑)	T-Cons (↑)
ZeroScope	7.19	30.8	0.982
PEEKABOO	11.4	30.4	0.975
Ours	18.1	29.9	0.978

Ablation Study (ModelScope)



MA: Multi Masked Attention
LA: Latent Alignment

Conclusion

Proposed method of controlling multiple objects with zero shots.

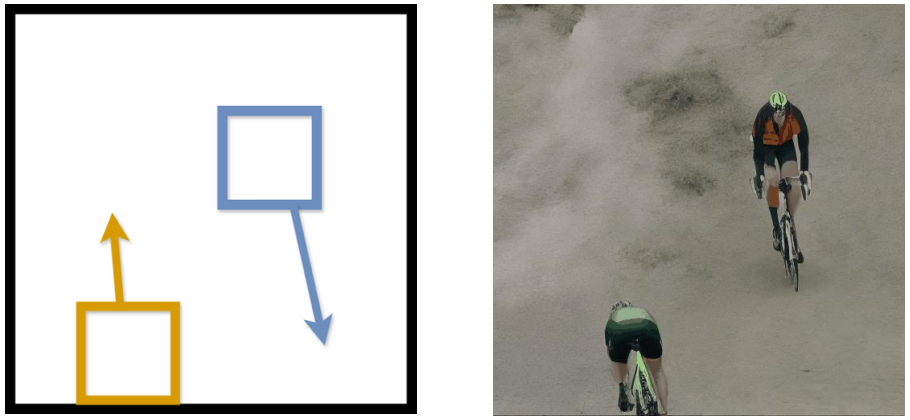
1. Priority Multi Masked Attention
2. Latent Alignment

Quantitative and qualitative evaluation showed the effectiveness of our method.

Future work

- Application to Diffusion Transformer
- Conditioning of the entire sampling

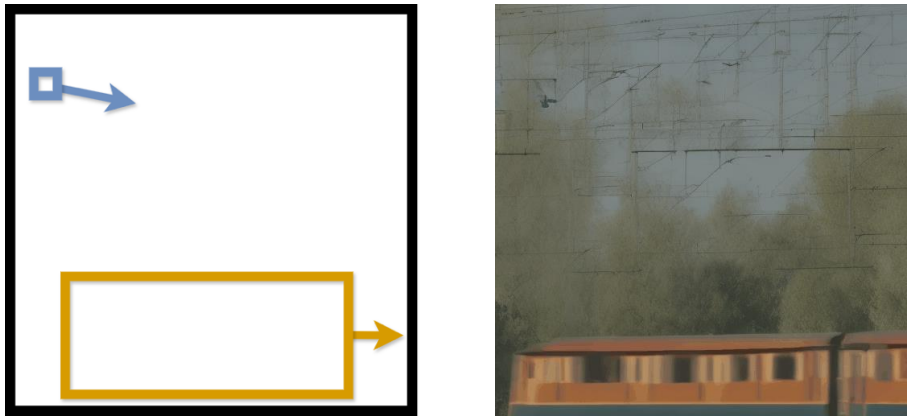
Generation Videos



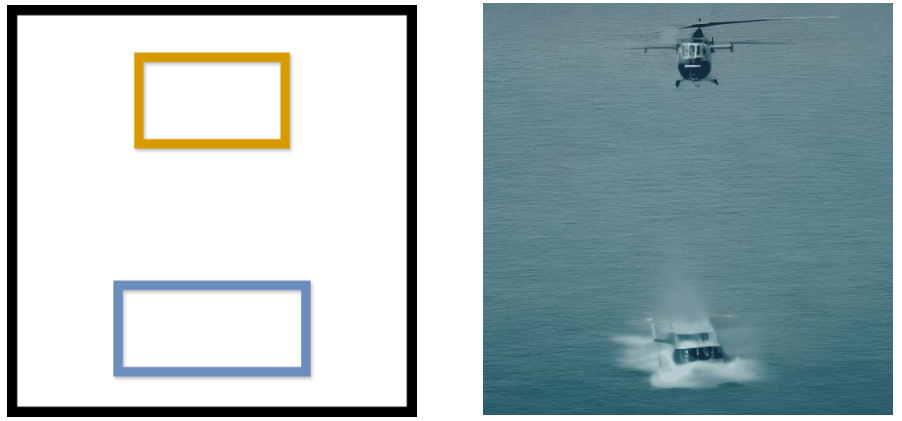
A cyclist pedaling uphill as another speeds downhill



A firefighter spraying water at a burning house



A train speeding past as a bird flies overhead



A helicopter hovering above a boat on the ocean

Evaluation Datasets with LLM

- Model: GPT-4
- In-context Learning: 3-shot learning

```
"0003": {
    "prompt": "A panda and tiger walking in the forest",
    "objects": ["panda", "tiger"],
    "resolution": [512, 512],
    "bbox": {
        "panda": [
            [50, 300, 100, 100],
            [80, 300, 100, 100],
            [110, 300, 100, 100],
            [140, 300, 100, 100],
            [170, 300, 100, 100],
            [200, 300, 100, 100]
        ],
        "tiger": [
            [200, 320, 120, 80],
            [236, 320, 120, 80],
            [272, 320, 120, 80],
            [308, 320, 120, 80],
            [344, 320, 120, 80],
            [380, 320, 120, 80]]}},
```