Tune-A-Video: One-Shot Tuning of Image Diffusion Models For Text-to-Video Generation

Jay Zhangjie Wu[1] Yixiao Ge[2] Xintao Wang[2] Stan Weixian Lei[1] Yuchao Gu[1]
Wynne Hsu[4] Ying Shan[2] Xiaohu Qie[3] Mike Zheng Shou[1*]

[1]Show Lab, National University of Singapore [2]ARC Lab,[3]Tencent PCG
[4]School of Computing, National University of Singapore

arXiv 2022.12

STRUCT Group Seminar
Presenter: Wenjing Wang
2023.01.15

OUTLINE

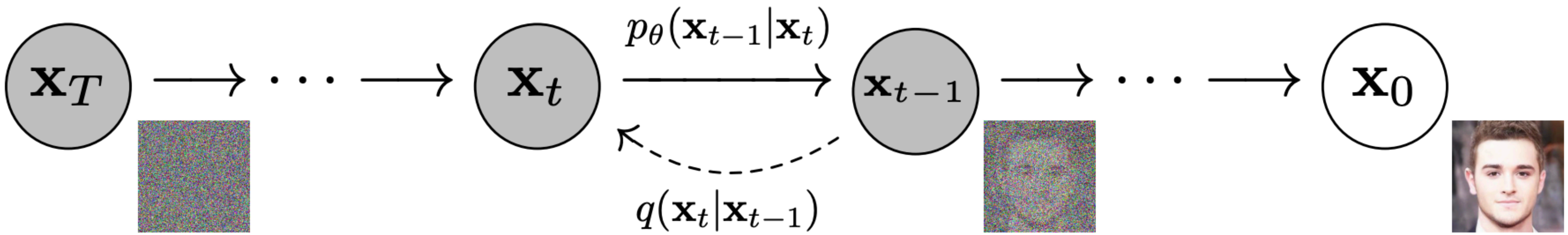
- Authorship
- Background
- Method
- Experiments
- Conclusion

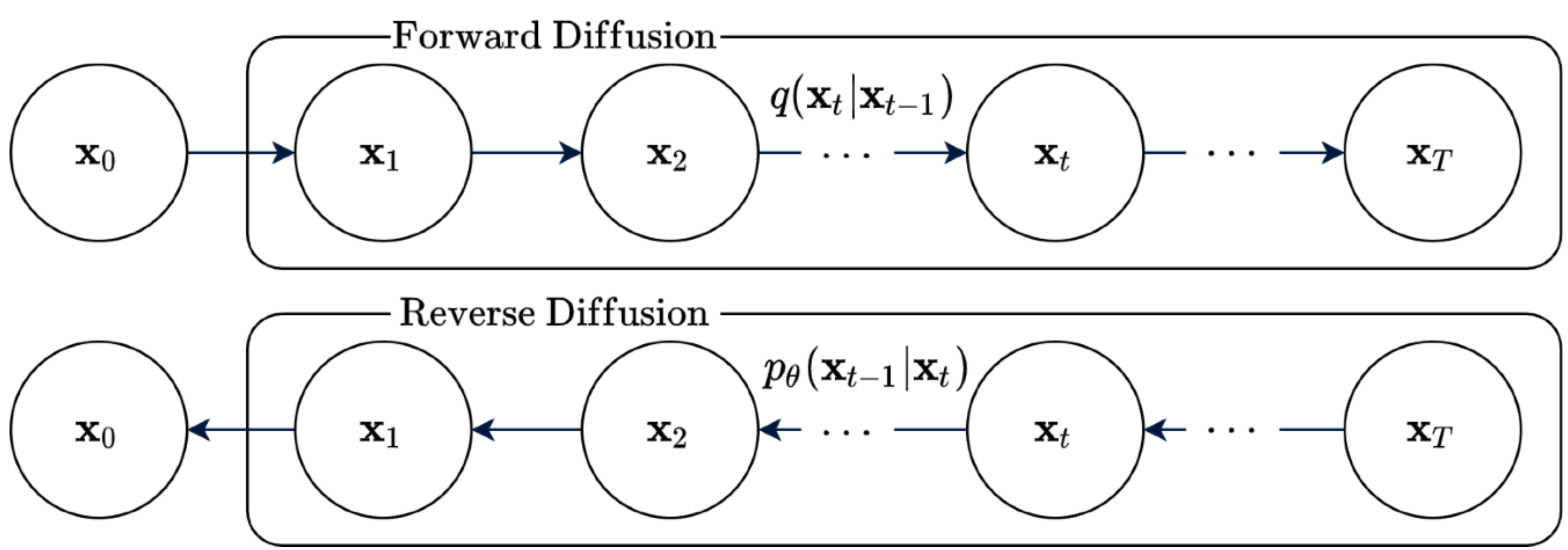
OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

BACKGROUND

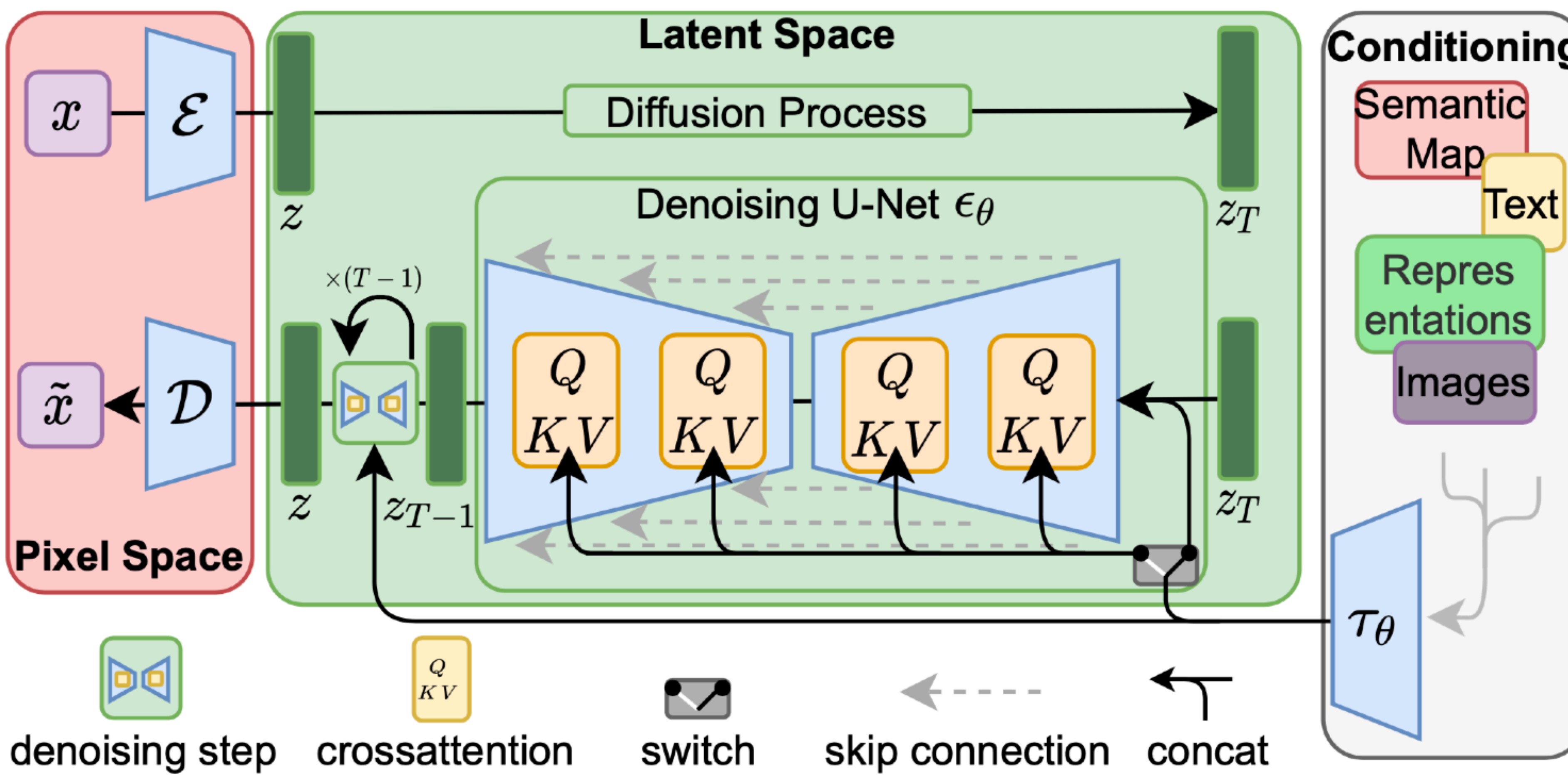
- Denoising diffusion probabilistic model (DDPM)





BACKGROUND

- Latent Diffusion Models (LDMs)
  - DDPM on latent (4×32×32) rather than on images (3×512×512)



BACKGROUND

- Stable Diffusion

Stable Diffusion

Stable Diffusion was made possible thanks to a collaboration with Stability AI and Runway and builds upon our previous work:

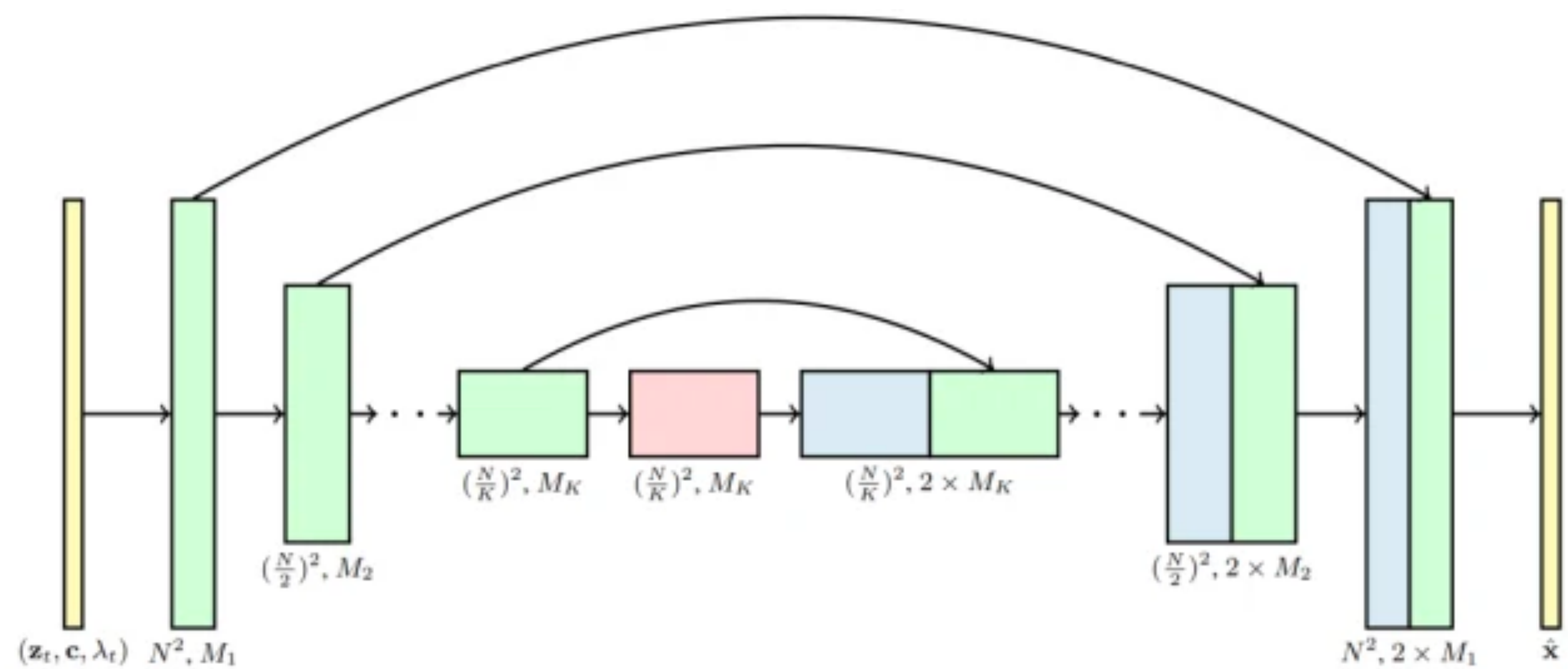
High-Resolution Image Synthesis with Latent Diffusion Models
Robin Rombach*, Andreas Blattmann*, Dominik Lorenz, Patrick Esser, Björn Ommer
CVPR '22 Oral | *GitHub* | *arXiv* | *Project page*

Stable Diffusion is a latent text-to-image diffusion model. Thanks to a generous compute donation from Stability AI and support from LAION, we were able to train a Latent Diffusion Model on 512x512 images from a subset of the LAION-5B database. Similar to Google's Imagen, this model uses a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts. With its 860M UNet and 123M text encoder, the model is relatively lightweight and runs on a GPU with at least 10GB VRAM. See this section below and the model card.

DIFFUSION-BASED VIDEO GENERATION

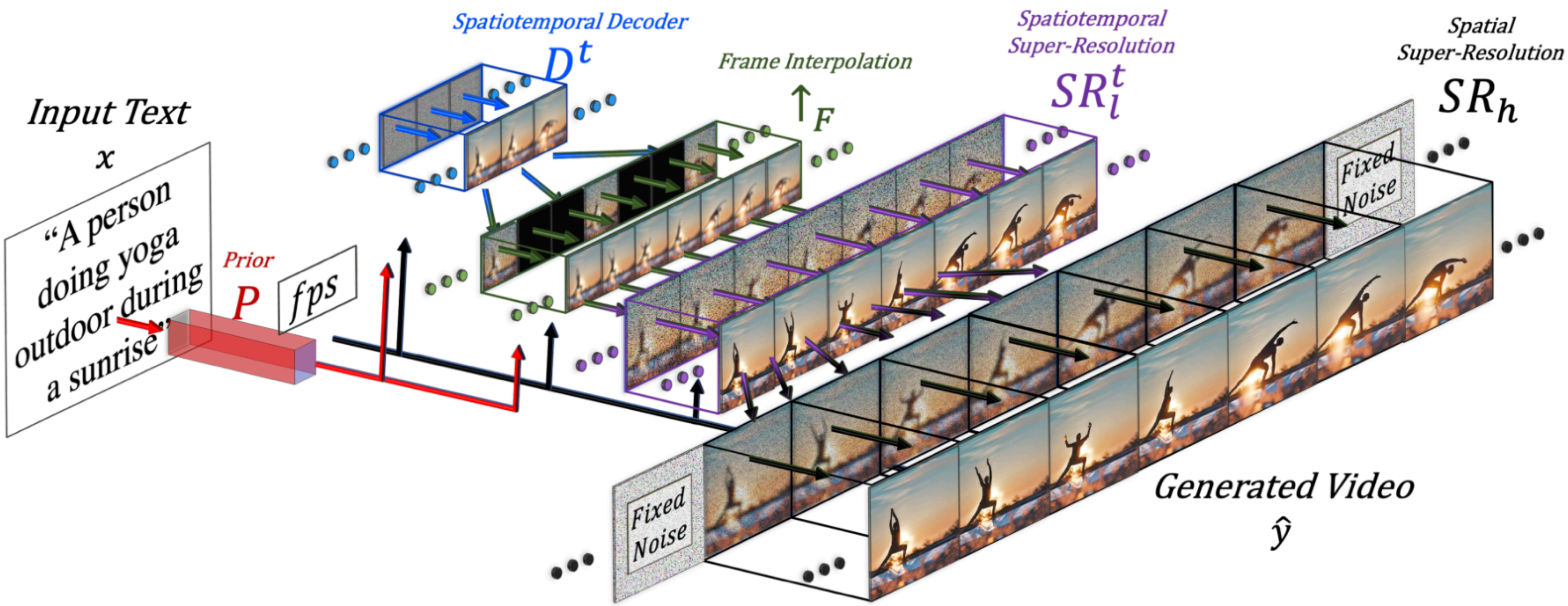
➤ Video Diffusion Models by Google Brain (ICLR 2022 Workshop)

- A space-time factorized U-Net
- Joint image and video data training



DIFFUSION-BASED TEXT-TO-VIDEO

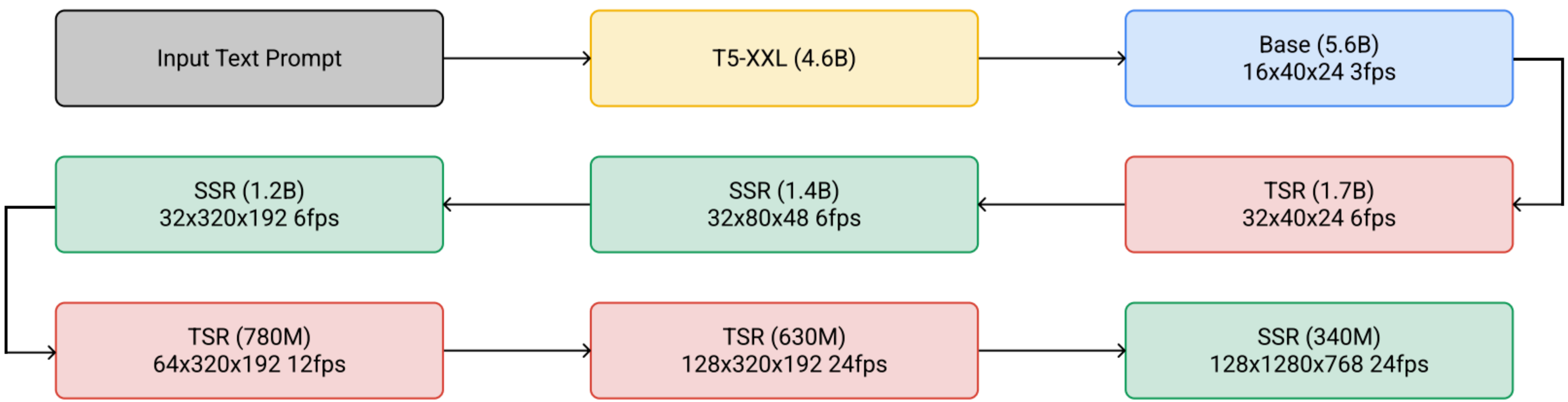
- ➤ Make-A-Video by Meta-AI (arXiv 2022.09)
- Text-to-image pairs and unlabeled videos



DIFFUSION-BASED TEXT-TO-VIDEO

➤ Imagen Video by Google Brain (arXiv 2022.10)

- No text-to-image pretraining
- Train on text-video pairs

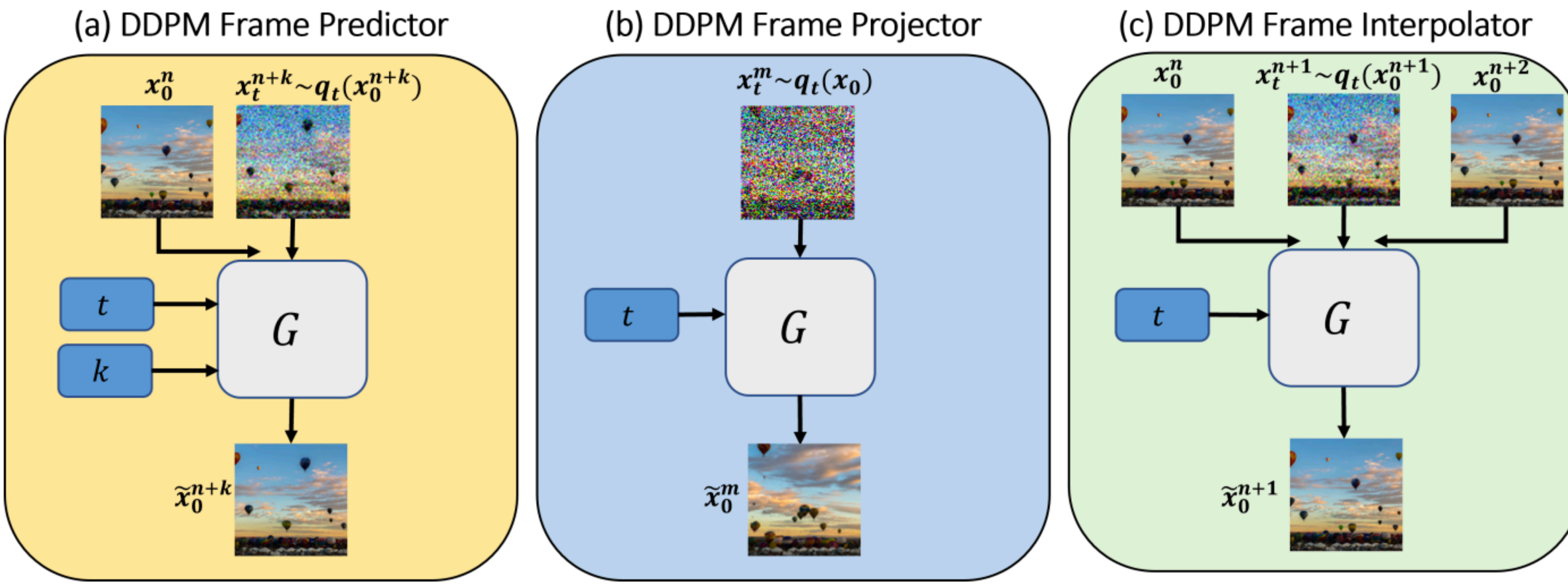


DIFFUSION-BASED TEXT-TO-VIDEO

- SinFusion (arXiv 2022.11)

DIFFUSION-BASED TEXT-TO-VIDEO

- SinFusion (arXiv 2022.11)



OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

METHOD

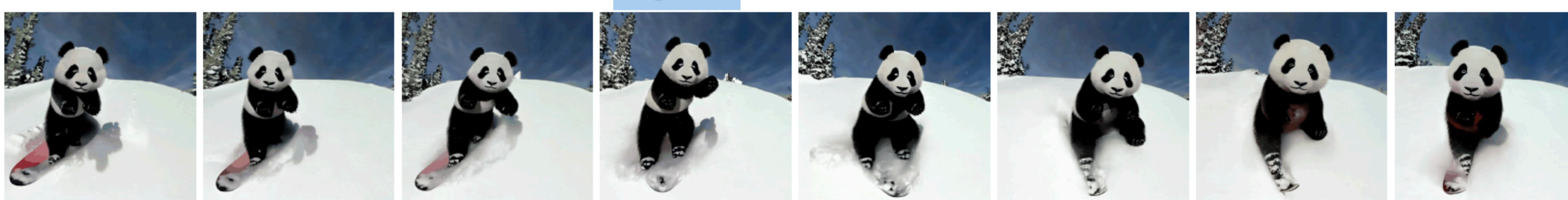
- Task: open-domain one-shot text-to-video (T2V) generation
- Model is trained with:
    - An open-domain pre-trained text-to-image (T2I) model
        - Stable diffusion v1.4
        - 1.4 billion parameter
    - A single text-video pair

METHOD - OVERVIEW

[Training video] A man is skiing on snow.

A man wearing red clothes, is skiing on snow.

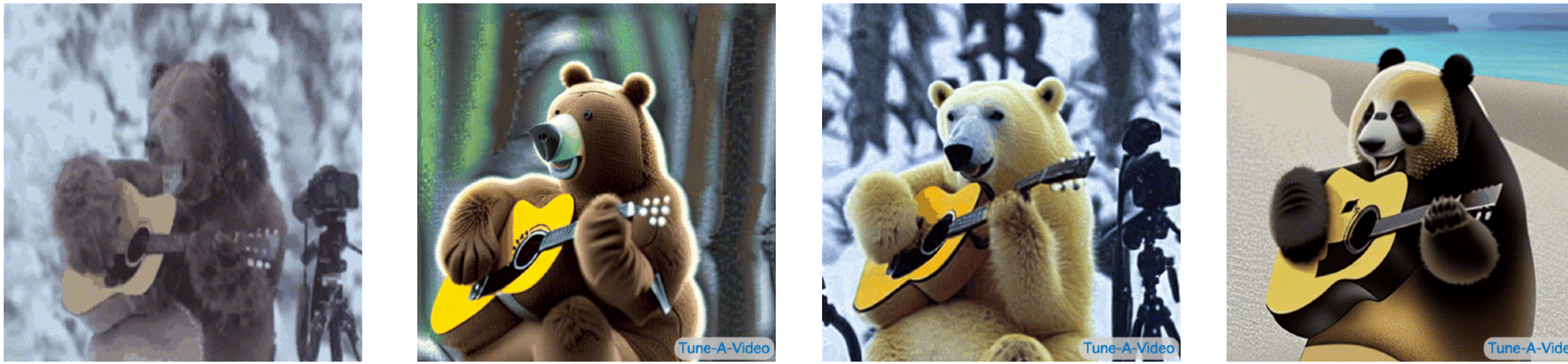
A panda is skiing on snow.



An astronaut is skiing on the moon.

METHOD - OVERVIEW

A bear is playing guitar. | "in the forest" | "polar bear" | "panda", "on the beach"



https://tuneavideo.github.io/

METHOD - OVERVIEW

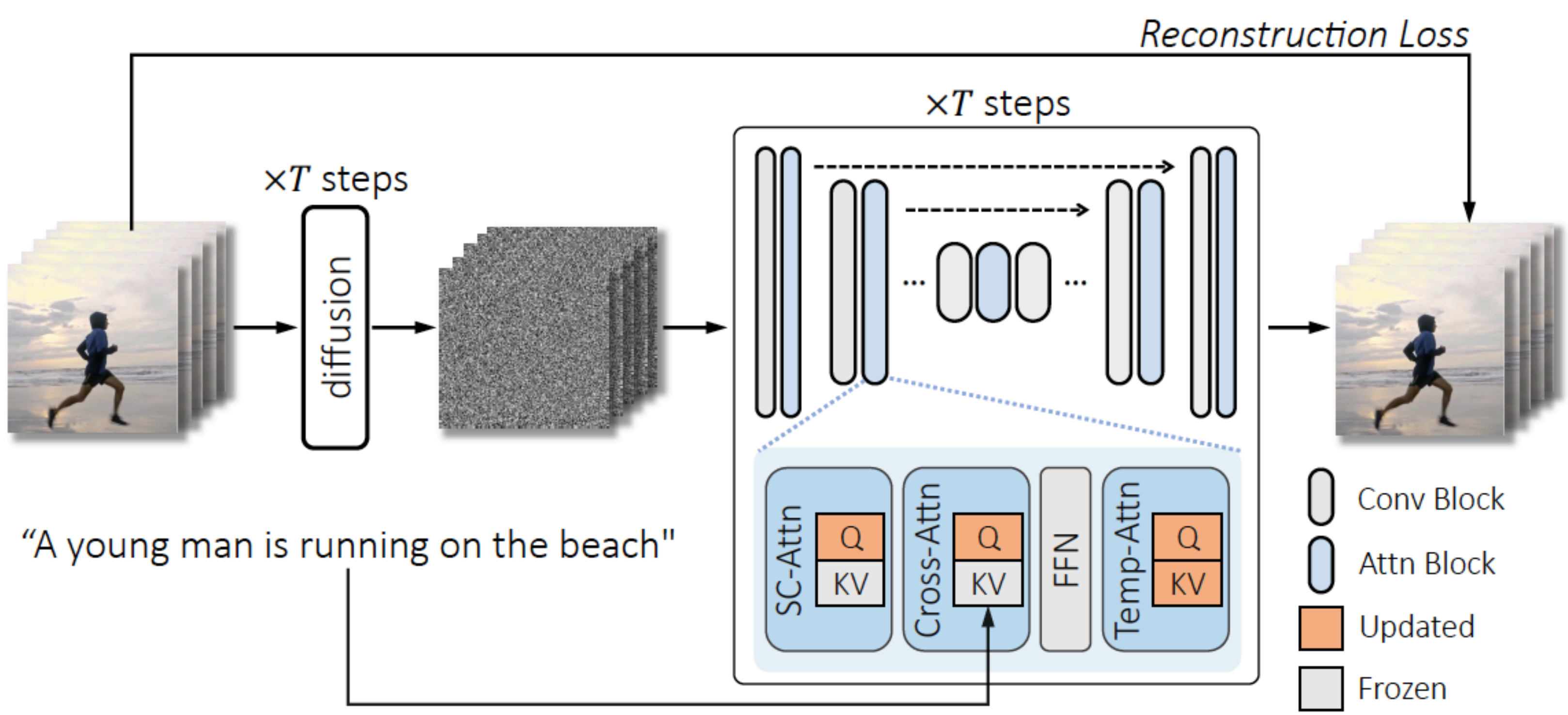
- Two key observations:
  - T2I models are able to generate images that align well with the verb terms
  - A simple cross-frame attention that attends the first video frame enables generating a sequence of frames that are consistent in content

Figure 2. Observations on T2I models. 1) T2I models are able to generate images that are well-aligned with the verb terms. 2) Extending self-attention over images maintains content consistency.

METHOD

- ➤ T2I diffusion model
- U-Net, 2D residual blocks followed by attention blocks
- Attention block: <u>spatial self-attention</u> + <u>cross-attention</u> + feed-forward

image only   image - text



T2I to T2V:

- Conv 3×3 → Conv 1×3×3
- Attention

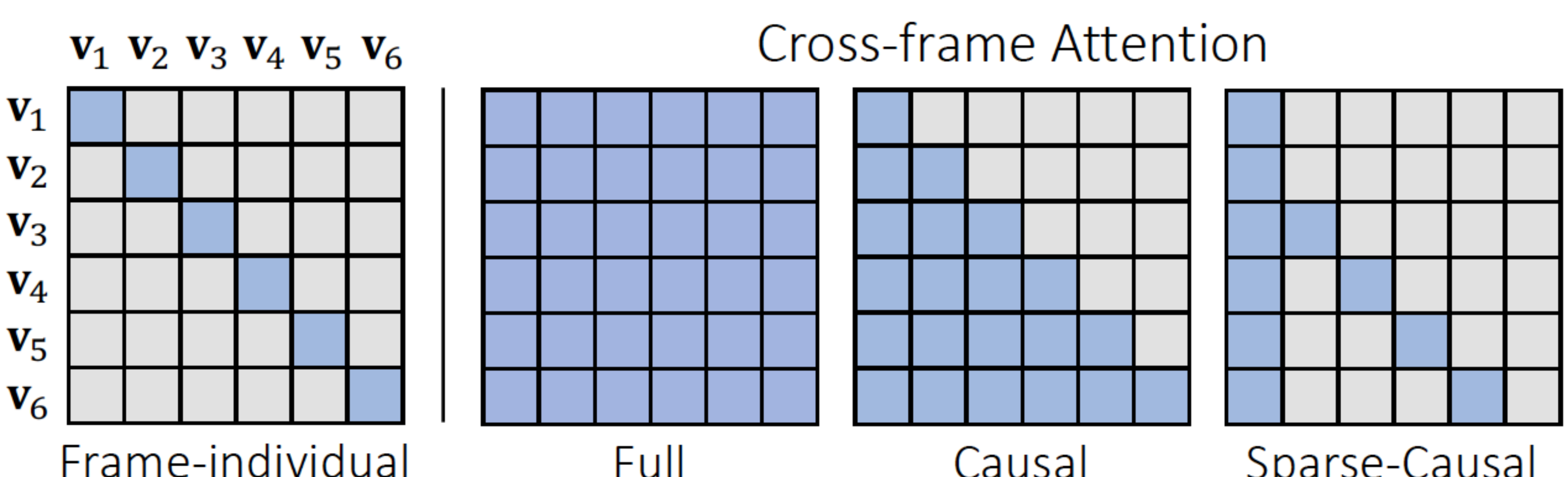
METHOD

➤ Sparse-Causal Attention

- Spatial self-attention → spatio-temporal
- Spatial self-attention: $\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$, with

$$Q = W^Q \mathbf{v}_i, K = W^K \mathbf{v}_i, V = W^V \mathbf{v}_i,$$

- Spatio-temporal cross-frame attention:



$$Q = W^Q \mathbf{v}_i$$

$$K = W^K [\mathbf{v}_1; \mathbf{v}_{i-1}]$$

$$V = W^V [\mathbf{v}_1; \mathbf{v}_{i-1}]$$

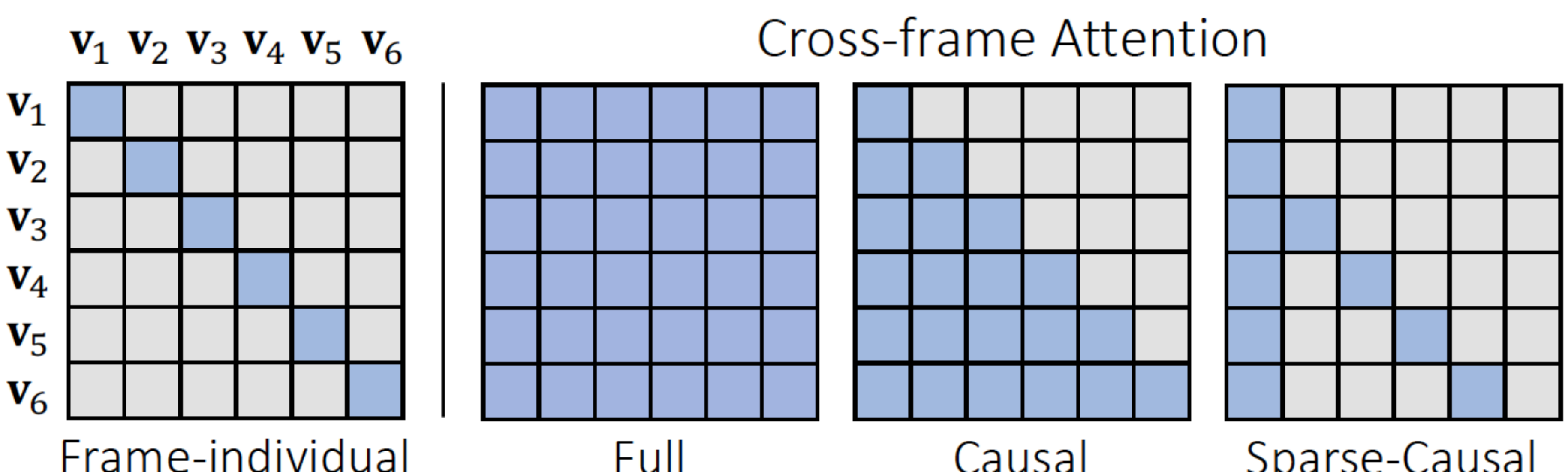
METHOD

➤ Sparse-Causal Attention

- Spatial self-attention → spatio-temporal
- Spatial self-attention: $\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$, with

$$Q = W^Q \mathbf{v}_i, K = W^K \mathbf{v}_i, V = W^V \mathbf{v}_i,$$

- Spatio-temporal cross-frame attention:



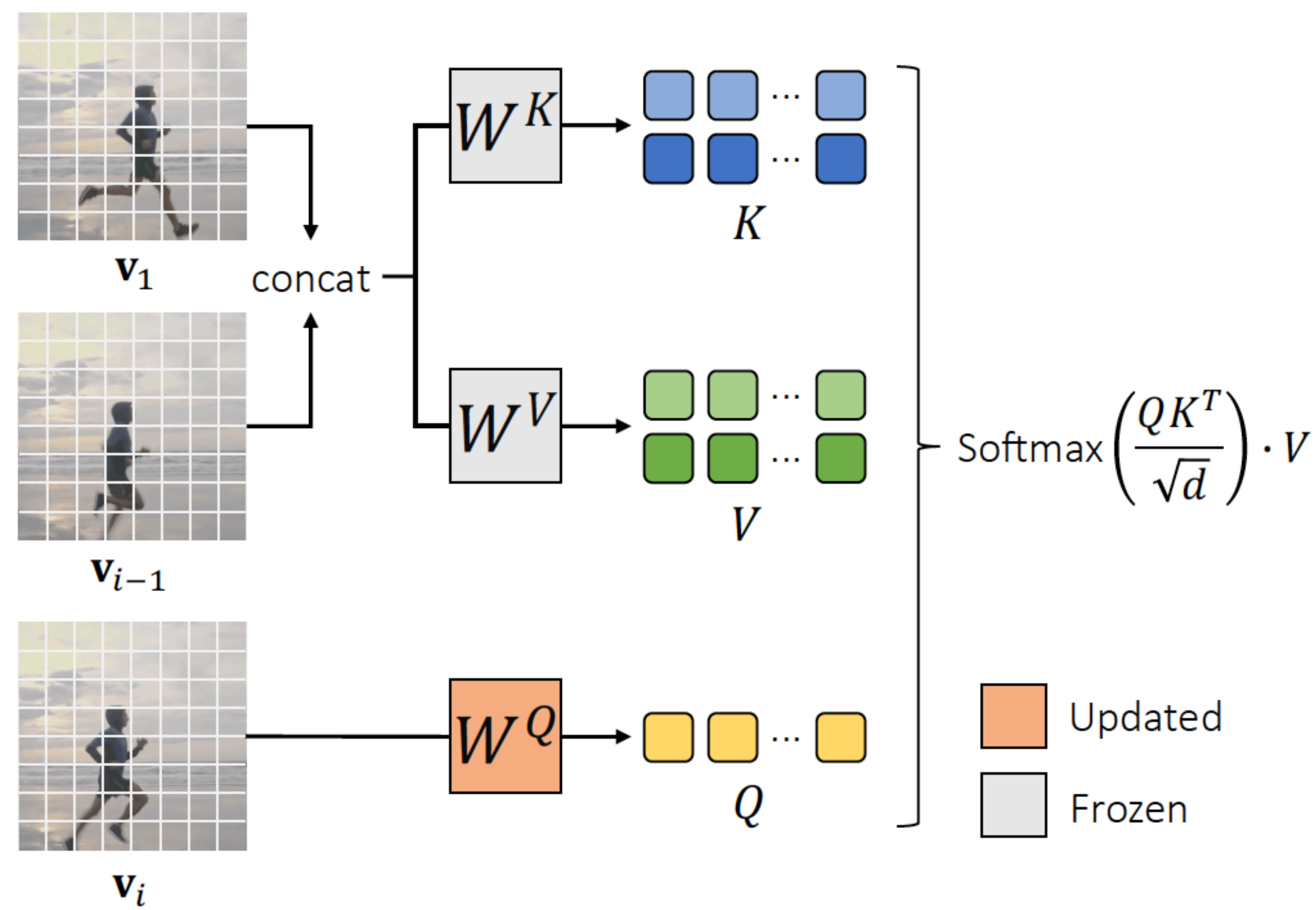
V_1: global coherence in terms of generated content

V_{i-1}: motion between consecutive frames

METHOD

➤ Sparse-Causal Attention

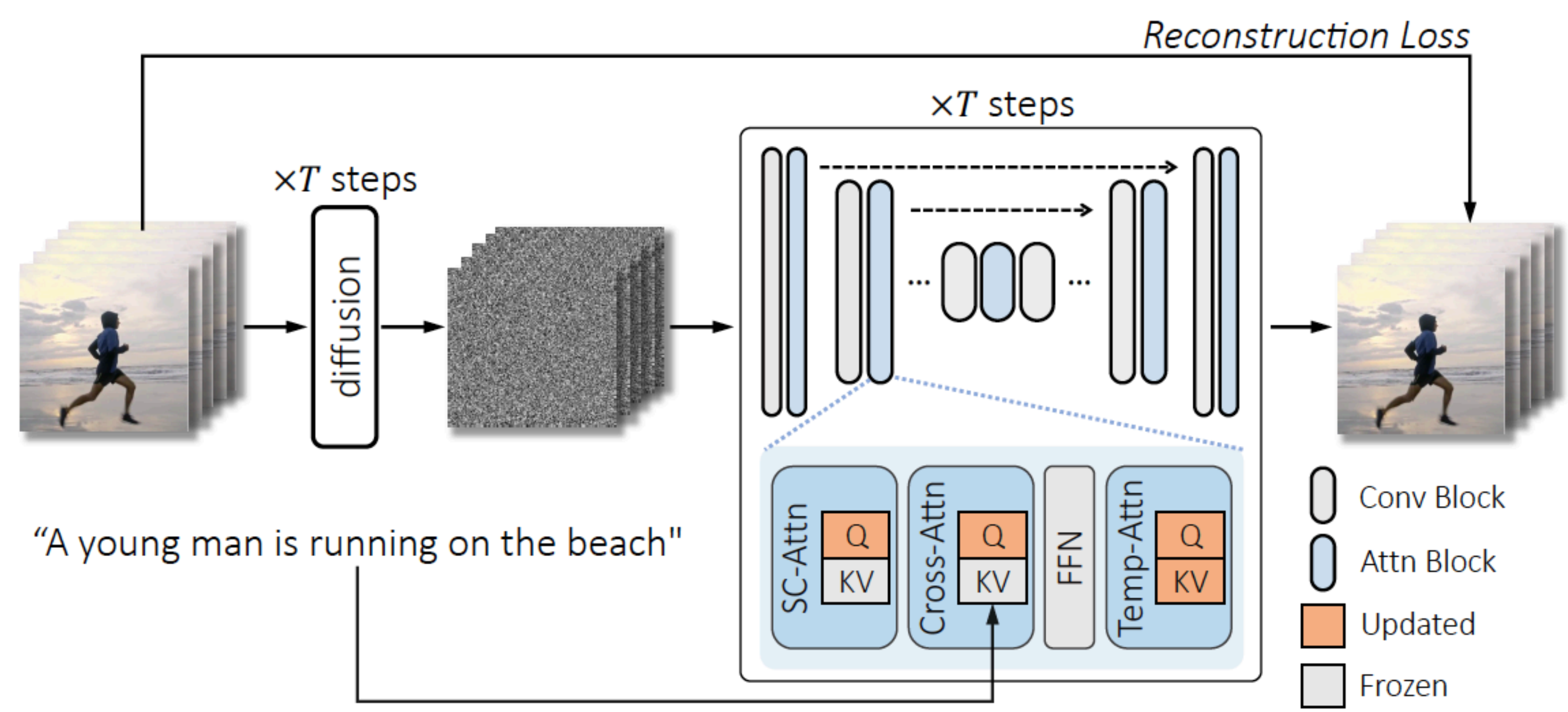
- One-way mapping from frame $\mathbf{v}_i$ to its previous frames ($\mathbf{v}_1$ and $\mathbf{v}_{i-1}$)
  - Key and value features derived from previous frames are independent to the output of $\mathbf{v}_i$
  - Therefore, we can fix W^K and W^V, only update W^Q



METHOD

➤ One-shot T2V diffusion model

- Train Q in SC-Attn
- Train Q in cross-Attn for better video-text alignment
- Tran QKV in Temp-Attn



OUTLINE

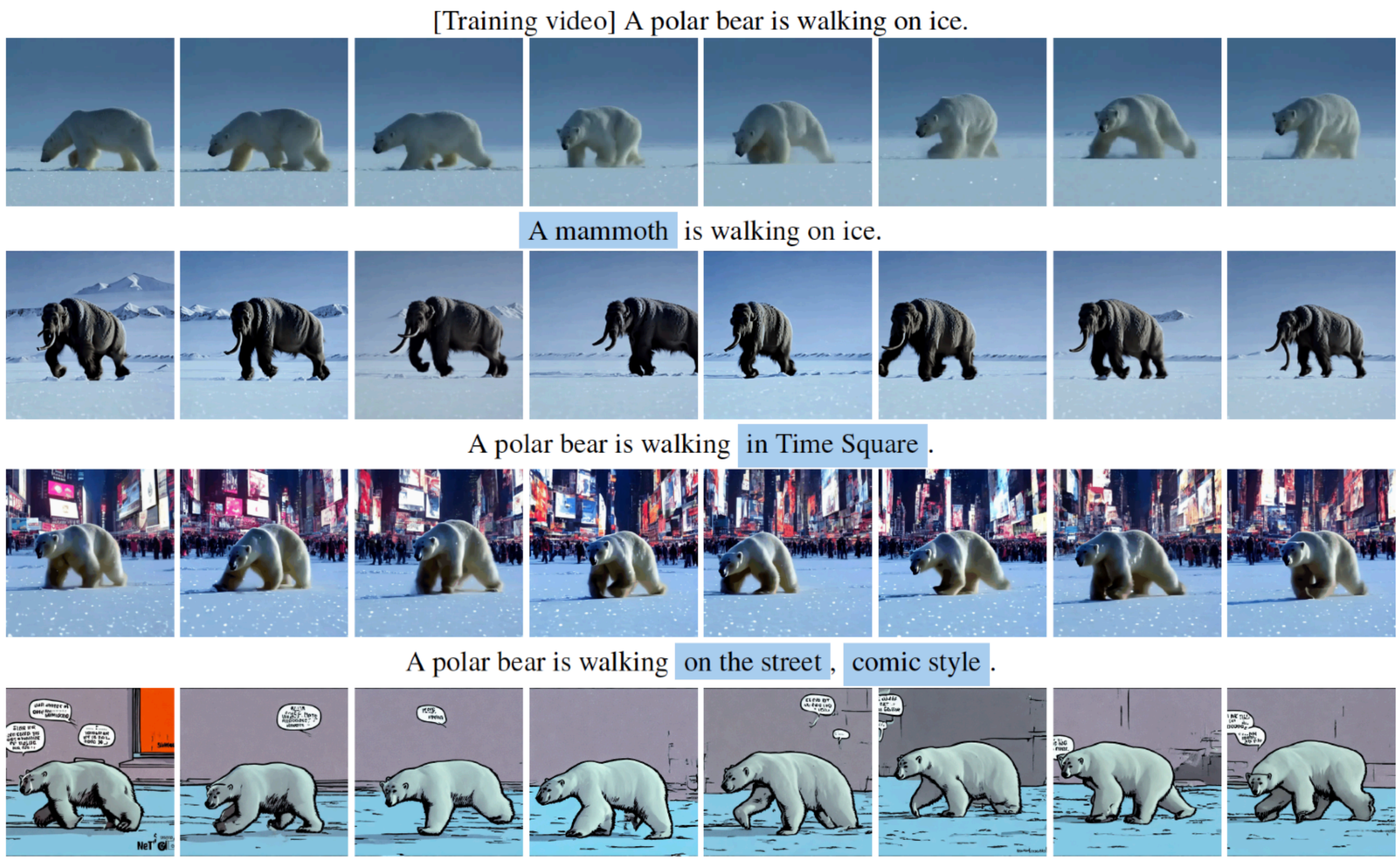
- Authorship
- Background
- Method
- Experiments
- Conclusion

EXPERIMENTS

- Implementation Details
    - Sample 8 uniform frames, resolution of 512×512 from input video
    - Train for 300 steps
    - Inference: DDIM sampler with classifier-free guidance

EXPERIMENTS

- Subject, background, attribute and style modification



EXPERIMENTS

- Subject, background, attribute and style modification

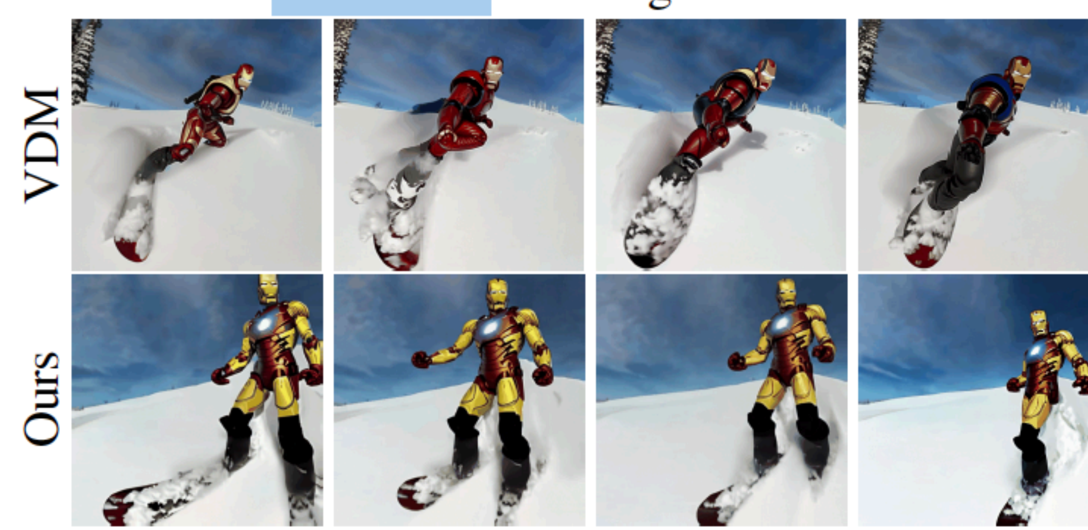
EXPERIMENTS

➤ Comparison with VDM Baselines

- VDM baselines:
  - Factorize space and time by appending a temporal attention after each spatial attention block in T2I diffusion models
  - The original 2D spatial blocks are kept in space only

A man is skiing on snow at night.

Iron Man is skiing on snow.



EXPERIMENTS

➤ Quantitative results

- CogVideo:
  - Based on a pre-trained T2I model CogView2
  - 9.4 billion parameters (~6 larger than Tune-A-Video)
  - Trained on a large-scale dataset of 5.4 million captioned videos.

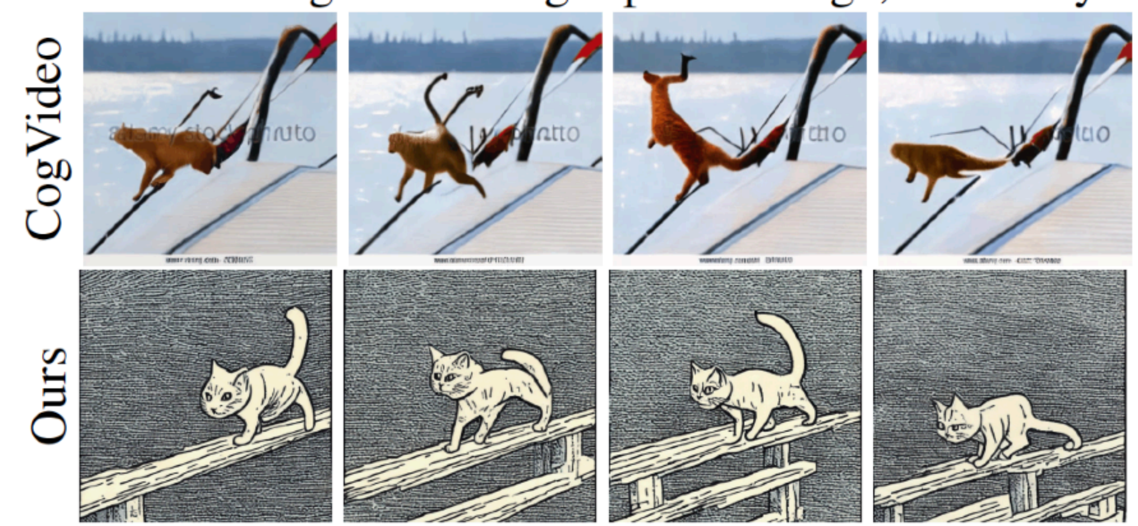
Method	CLIP Score	User study	
		Quality (%)	Faithfulness (%)
CogVideo [19]	23.66	13.76	9.38
Tune-A-Video	**26.57**	**86.24**	**90.62**

CLIP score : on 1024 video samples

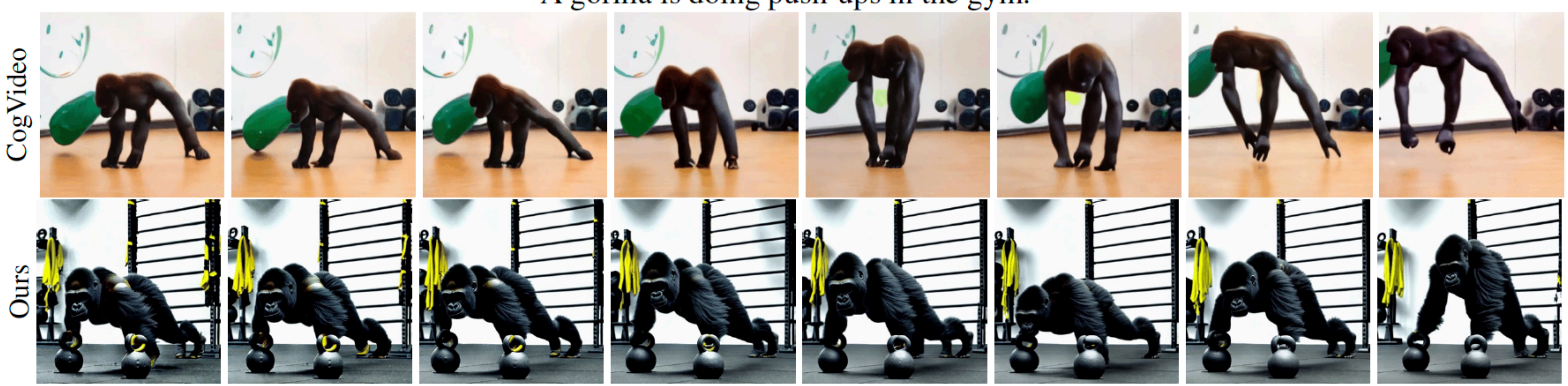
EXPERIMENTS

A woman is running on the lawn.

A cat is running on the single-plank bridge, comic style.

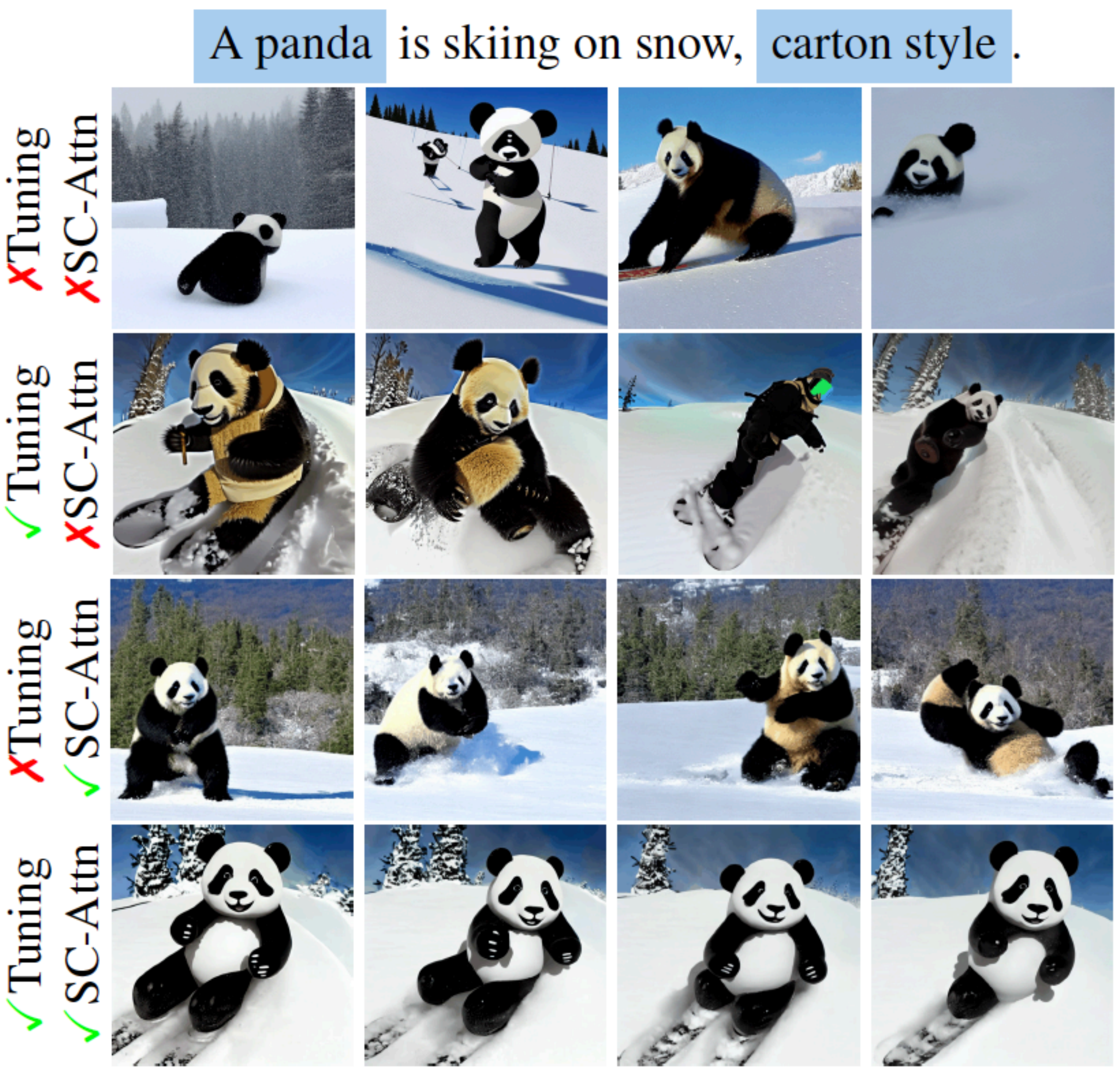


A gorilla is doing push-ups in the gym.



EXPERIMENTS

➤ Ablation Study: The effect of SC-Attn and One-Shot Tuning



OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

CONCLUSION

- ➤ Task: One-Shot Video Generation
- ➤ Sparse-Causal Attention