SPAD: Spatially Aware Multiview Diffusers

Yash Kant[1], Ziyi Wu[1], Michael Vasilkovsky[2], Guocheng Qian[2,3], Jian Ren[2], Riza Alp Guler[2], Bernard Ghanem[3], *Sergey Tulyakov[2], *Igor Gilitschenski[1], *Aliaksandr Siarohin[2]

University of Toronto[1] Snap Research[2] KAUST[3]

Task: Given **text prompt** of an object, we want **to** generate **consistent novel views**.

Task: Given **text prompt** of an object, we want **to** generate **consistent novel views**.

A knight's armored metal helmet with gold trim and holes

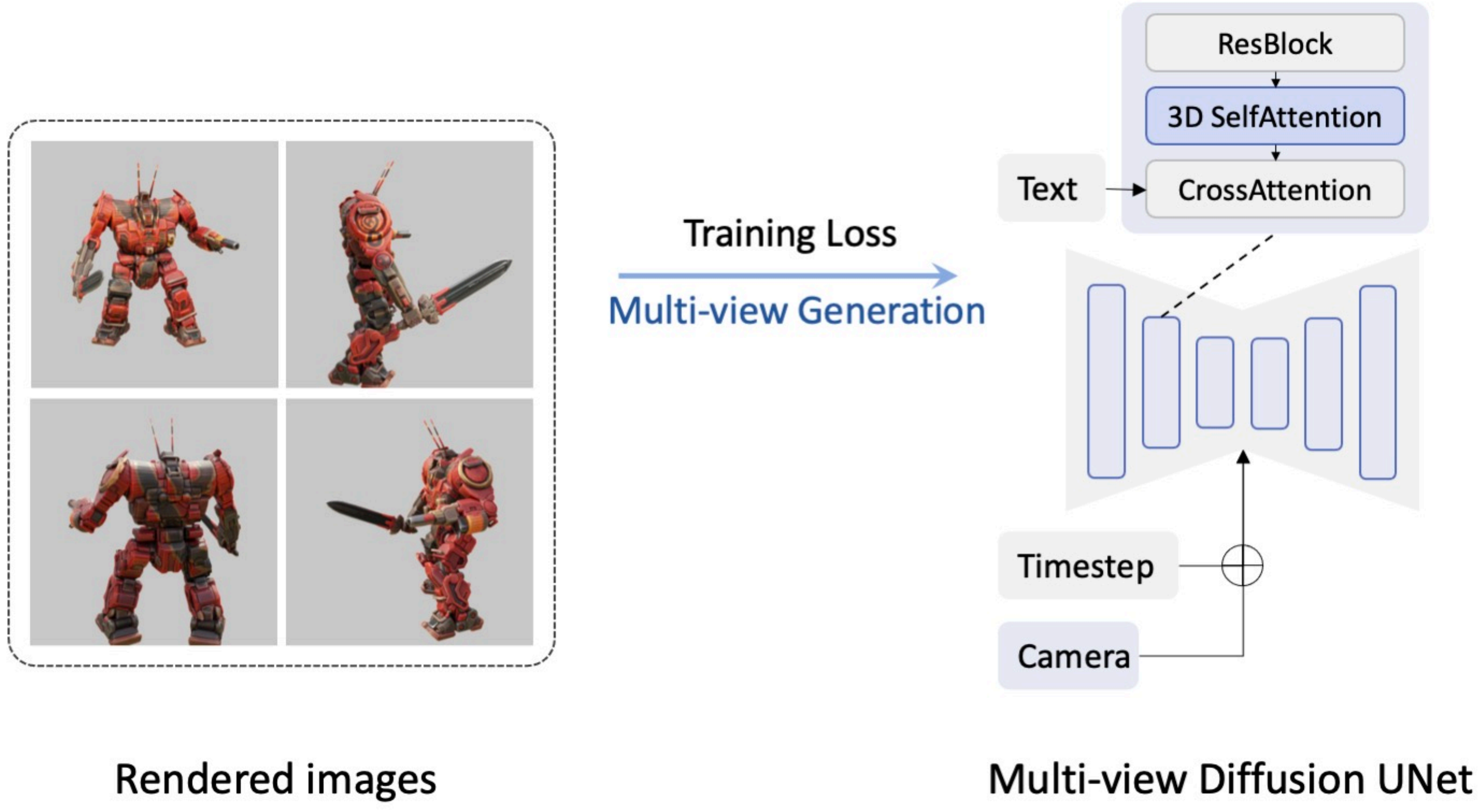
Task: Given **text prompt** of an object, we want **to** generate **consistent novel views**.

A knight's armored metal helmet with gold trim and holes

A small robot with a glass container on its head, metal legs, and a glass top

Prior Works: MVDream

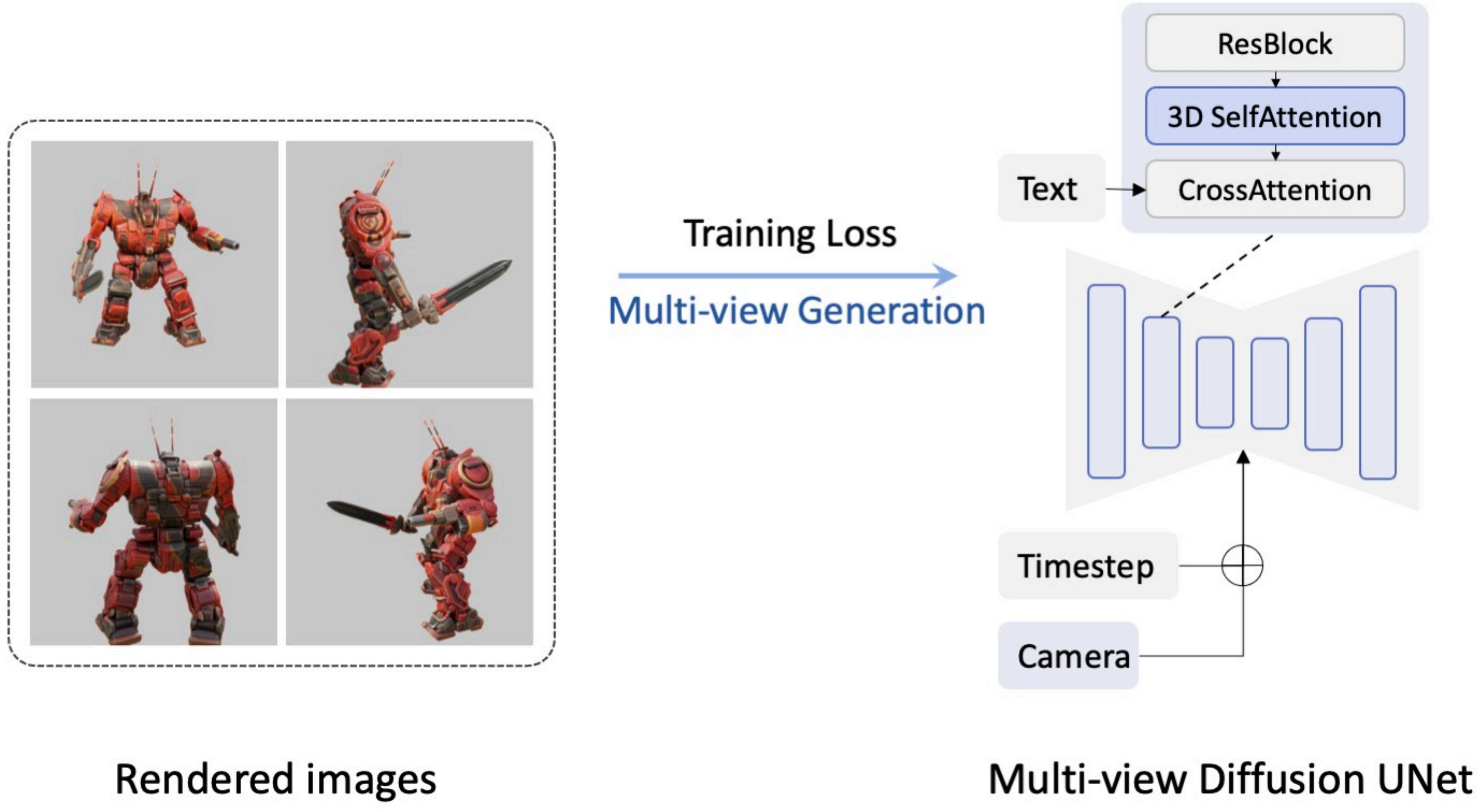
- Trained a Stable Diffusion (SD) to generate four orthogonal views of one object.



Prior Works: MVDream

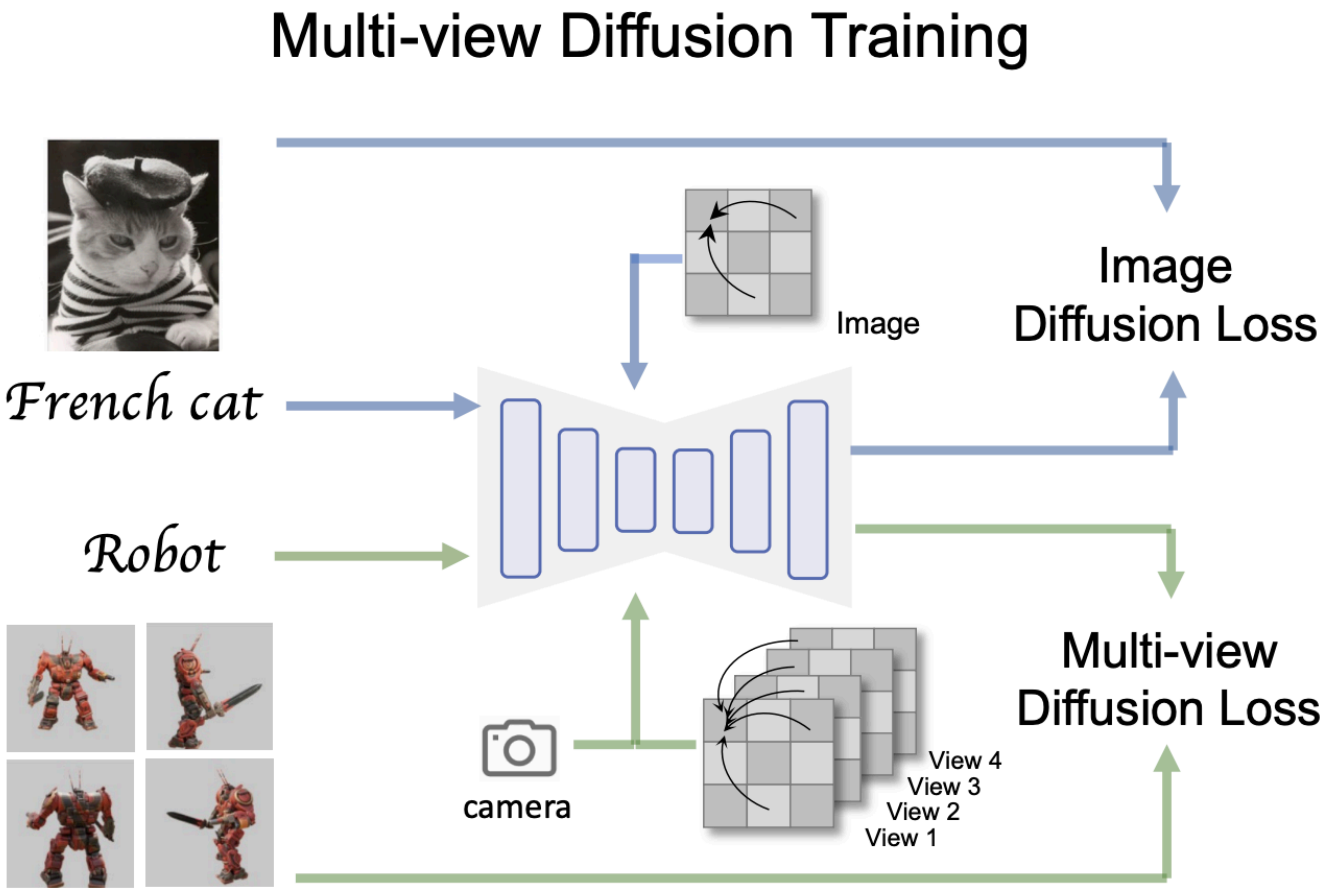
- Trained a Stable Diffusion (SD) to generate four orthogonal views of one object.

- Camera conditioning is injected into Diffusion Model (DM) — similar to timestep conditioning.



Prior Works: MVDream

- Self-attention inside SD is extended to attend 4 views jointly.



Prior Works: MVDream

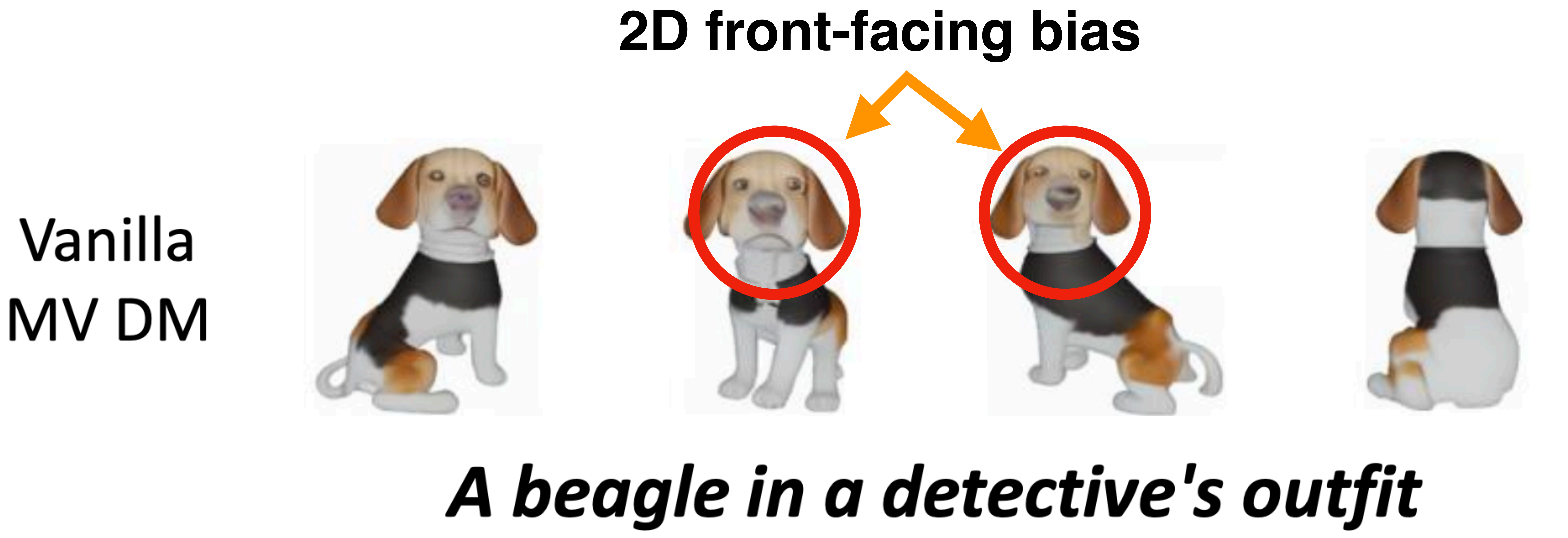
- Limitations.
  - fixed and few viewpoints.

Vanilla
MV DM

A beagle in a detective's outfit

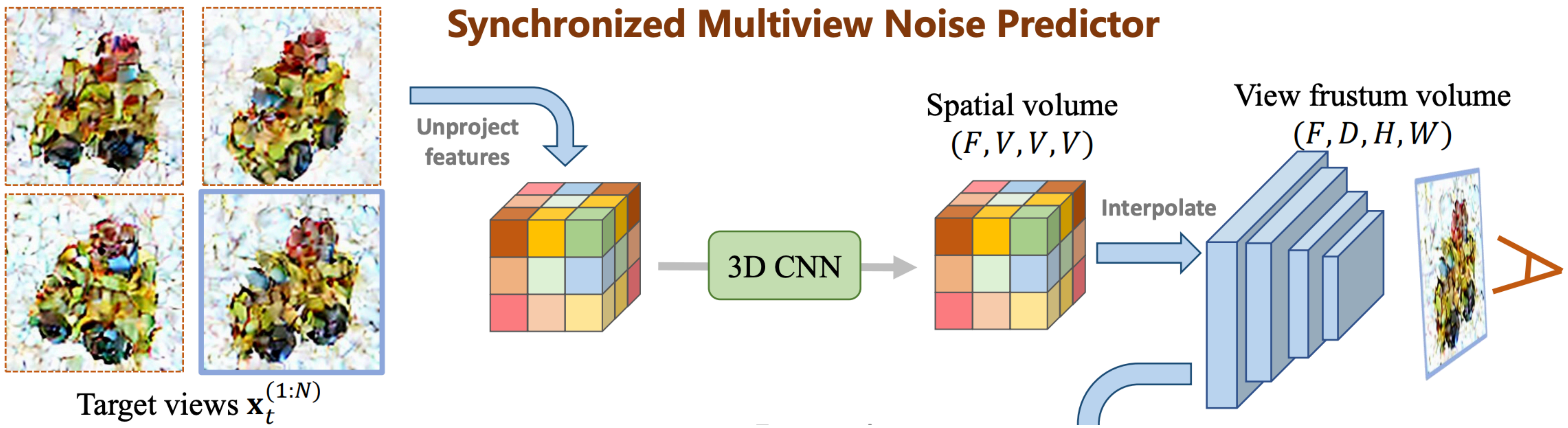
Prior Works: MVDream

- Limitations.
  - fixed and few viewpoints.
  - no 3D bias to maintain consistency — copy-paste views.



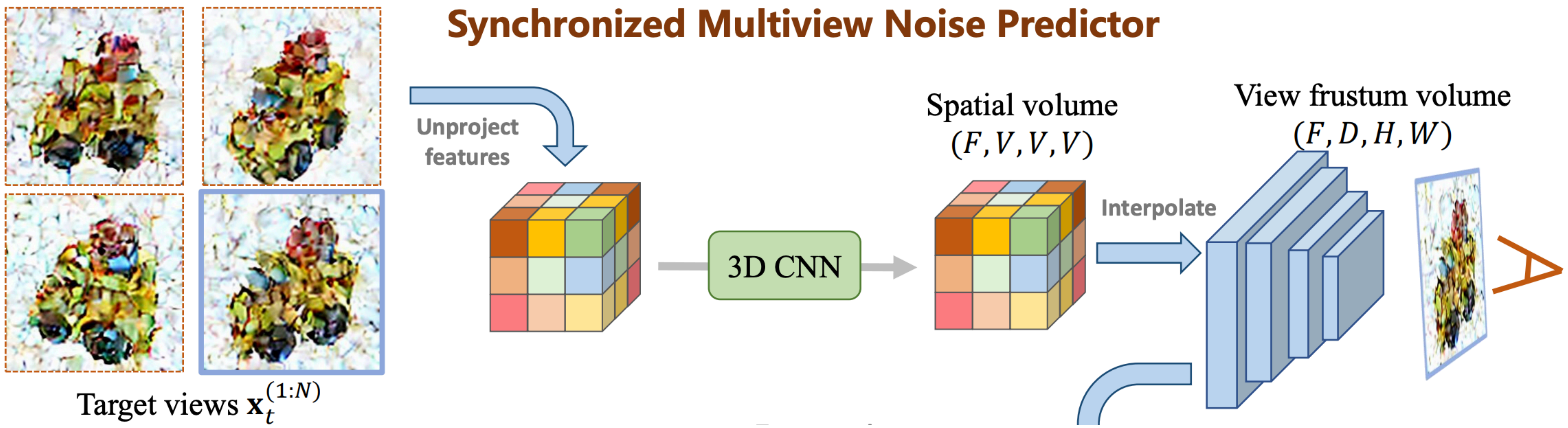
Prior Works: SyncDreamer

- Given an image — generate many views at fixed elevation.



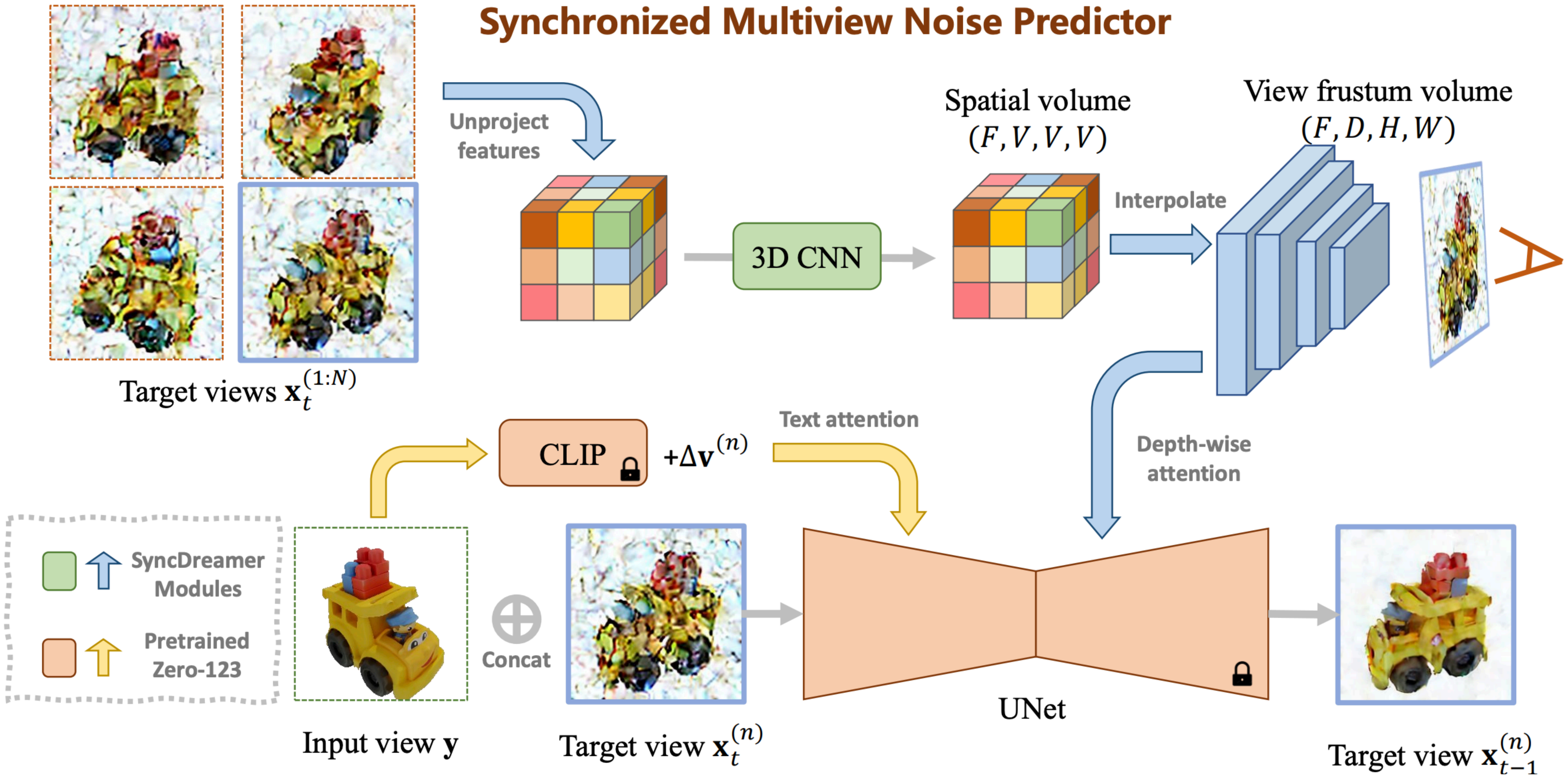
Prior Works: SyncDreamer

- Given an image — generate many views at fixed elevation.
- Extends Zero123 using a 3D latent space — views are aggregated, processed (3D-CNN) and extracted.



Prior Works: SyncDreamer

- Extracted view is then injected within the self-attention layer to generate final target.



Prior Works: SyncDreamer

- Limitations.
    - 3D volume and CNN has huge memory footprint.

Prior Works: SyncDreamer

- Limitations.
  - 3D volume and CNN has huge memory footprint.
  - Tiny volume leads to artefacts

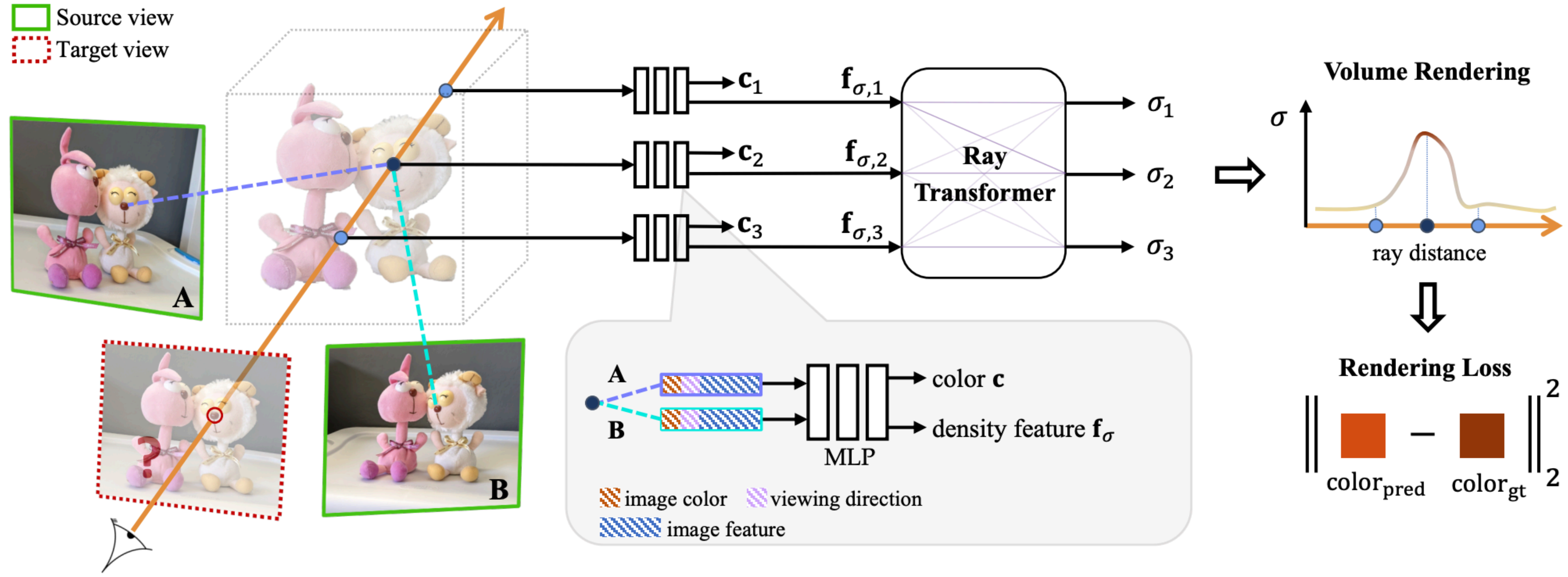
Prior Works: SyncDreamer

- Limitations.
  - 3D volume and CNN has huge memory footprint.
  - Tiny volume leads to artefacts (dragon legs / ballbag)

SPAD: How to inject 3D bias in a scalable way?

Inspired by Image-Based rendering methods [IBRNet] —

"Can we encode the 3D knowledge via correspondences within multi-view self-attention layers?"

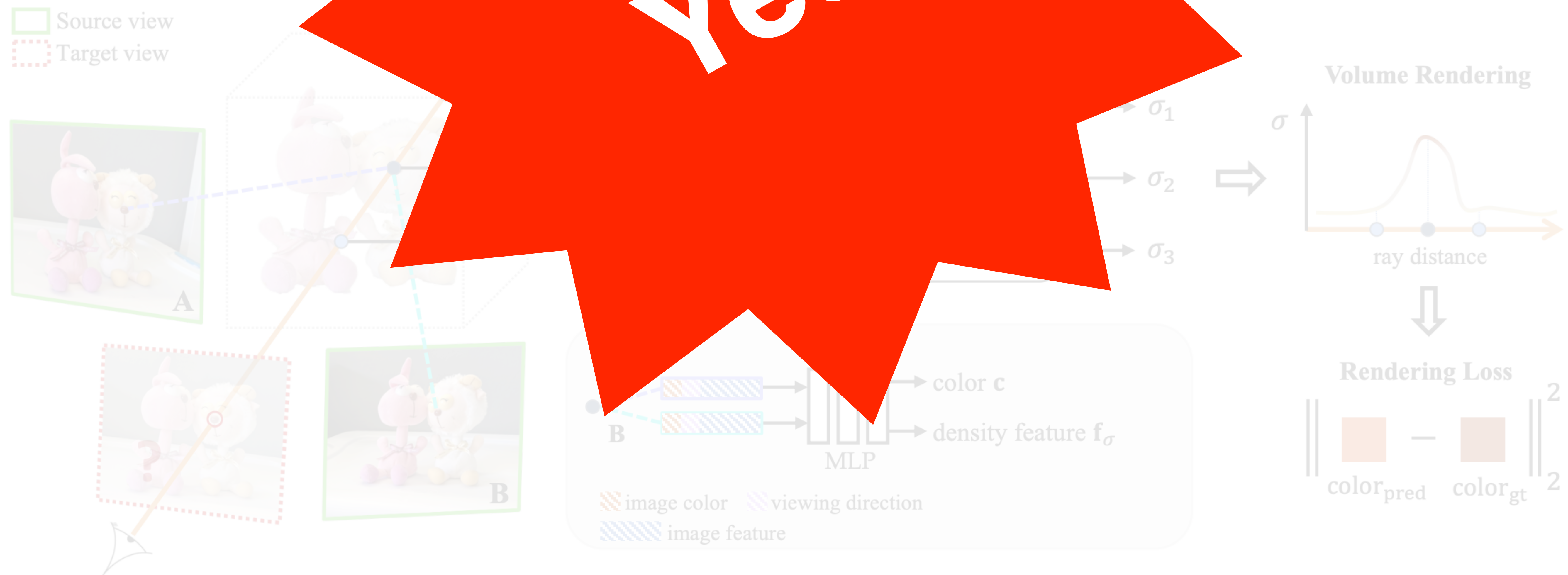


IBRNet: Learning Multi-View Image-Based Rendering [Qianqian Wang, et al. CVPR, 2021]

SPAD: How to inject 3D bias in a scalable way?

Inspired by Image-Based rendering methods [IBRNet] —

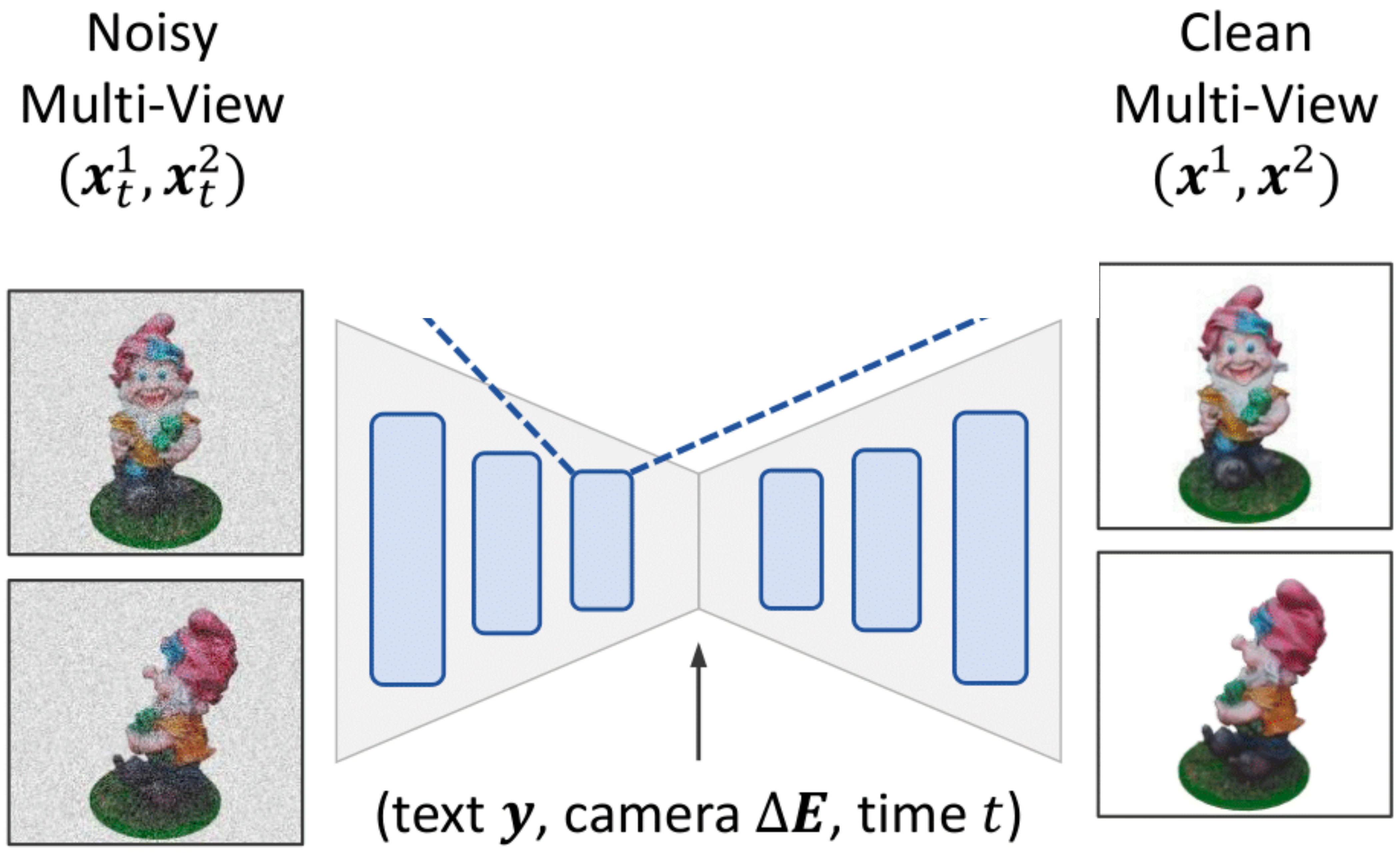
“Can we encode the 3D … e via correspondences within mult… on layers?”



IBRNet: Learning Multi-View Image-Based Rendering [Qianqian Wang, et al. CVPR, 2021]

SPAD: Multi-view Diffusion Model (MV-DM)

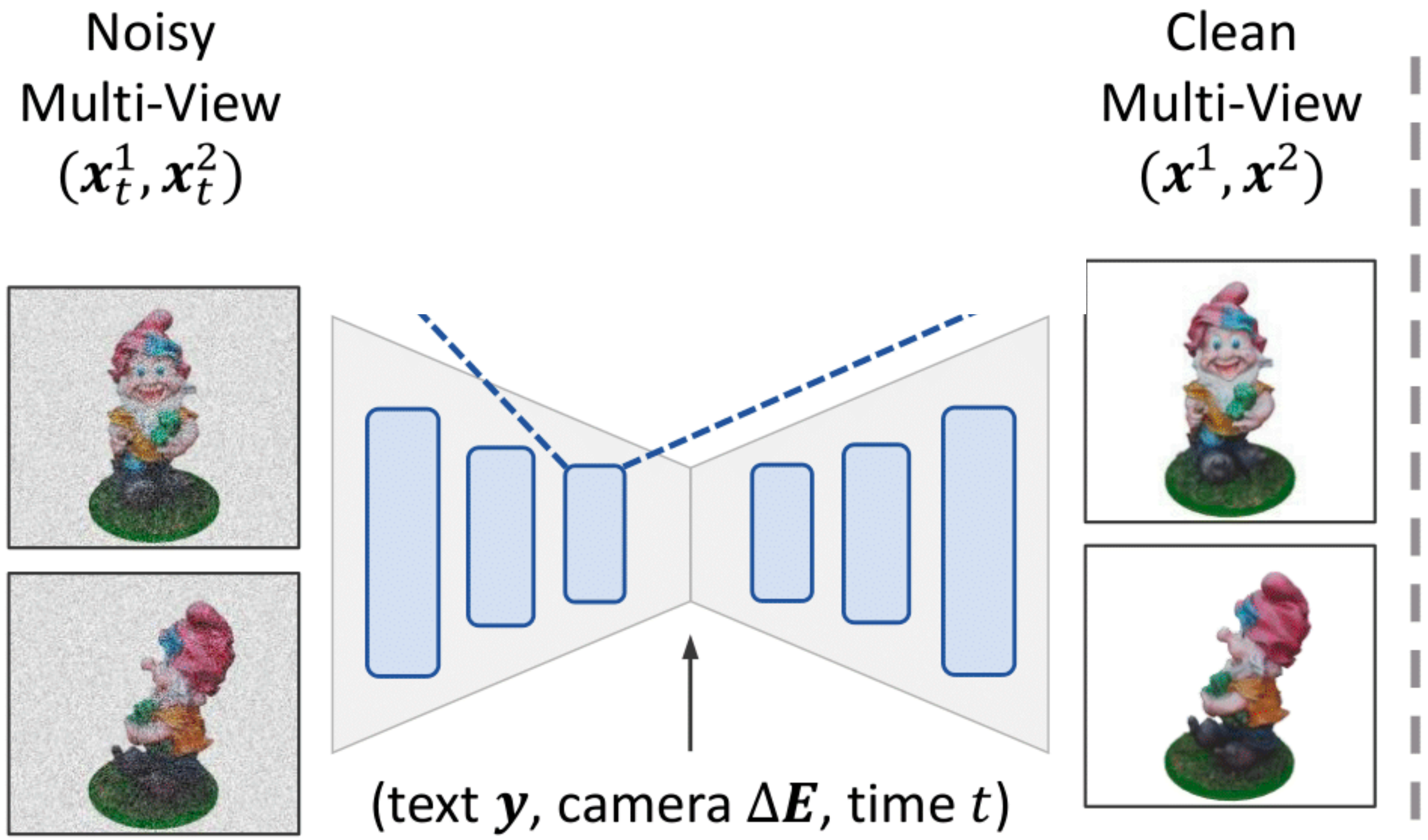
We start with simple MVDream-style model — with dense self-attention layers.



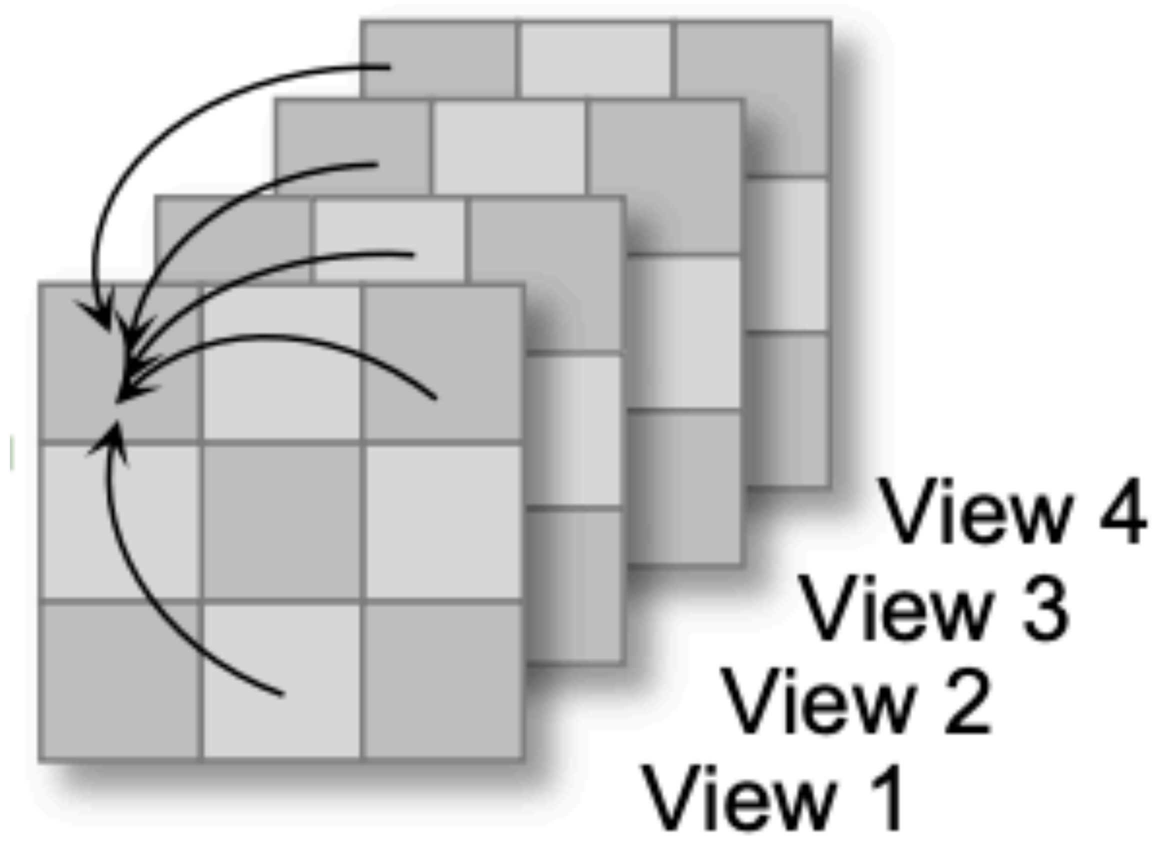
Multi-view Diffusion Model

SPAD: Multi-view Diffusion Model (MV-DM)

We start with simple MVDream-style model — with dense self-attention layers.



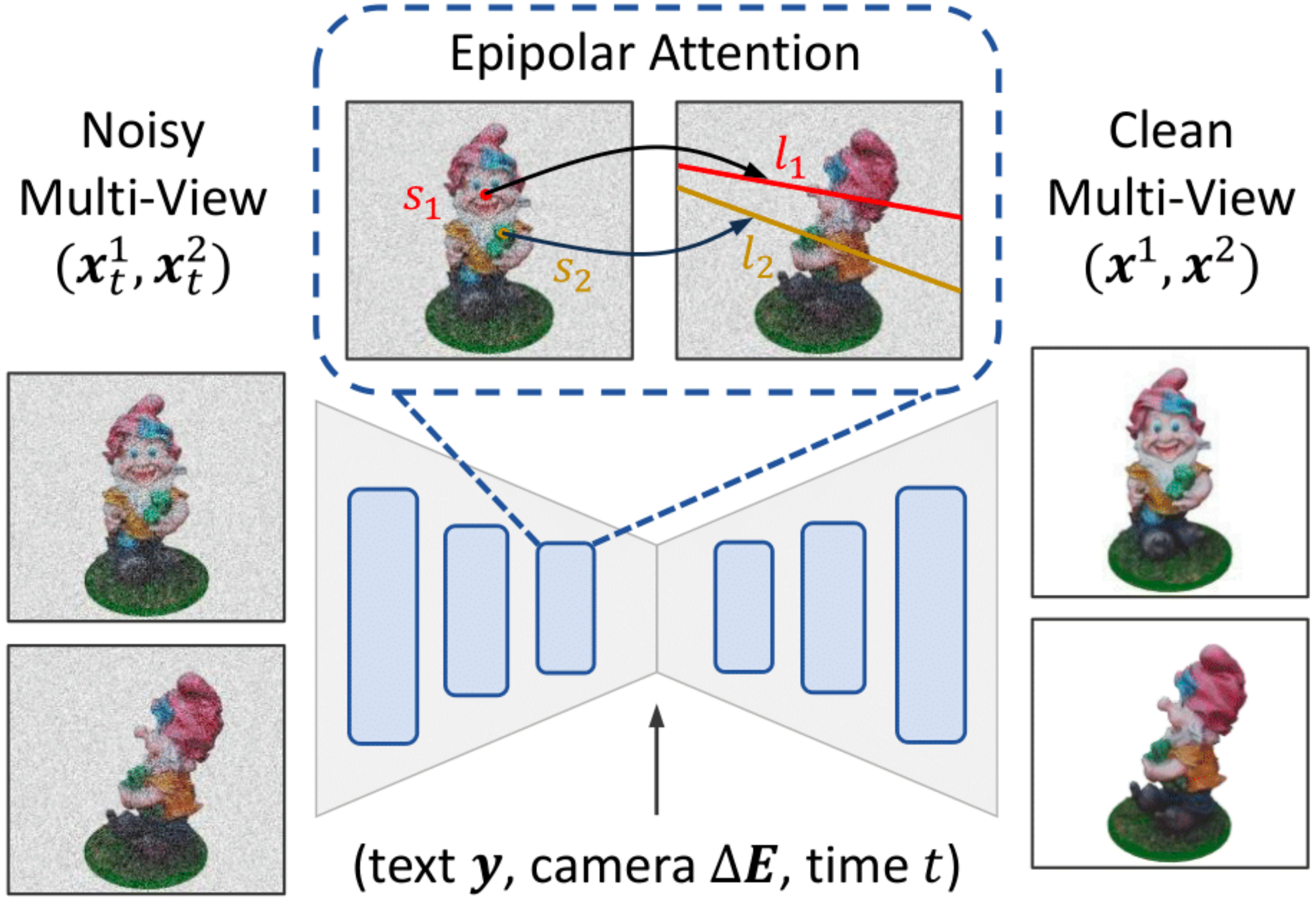
Multi-view Diffusion Model



Dense Attention

SPAD: Epipolar Attention

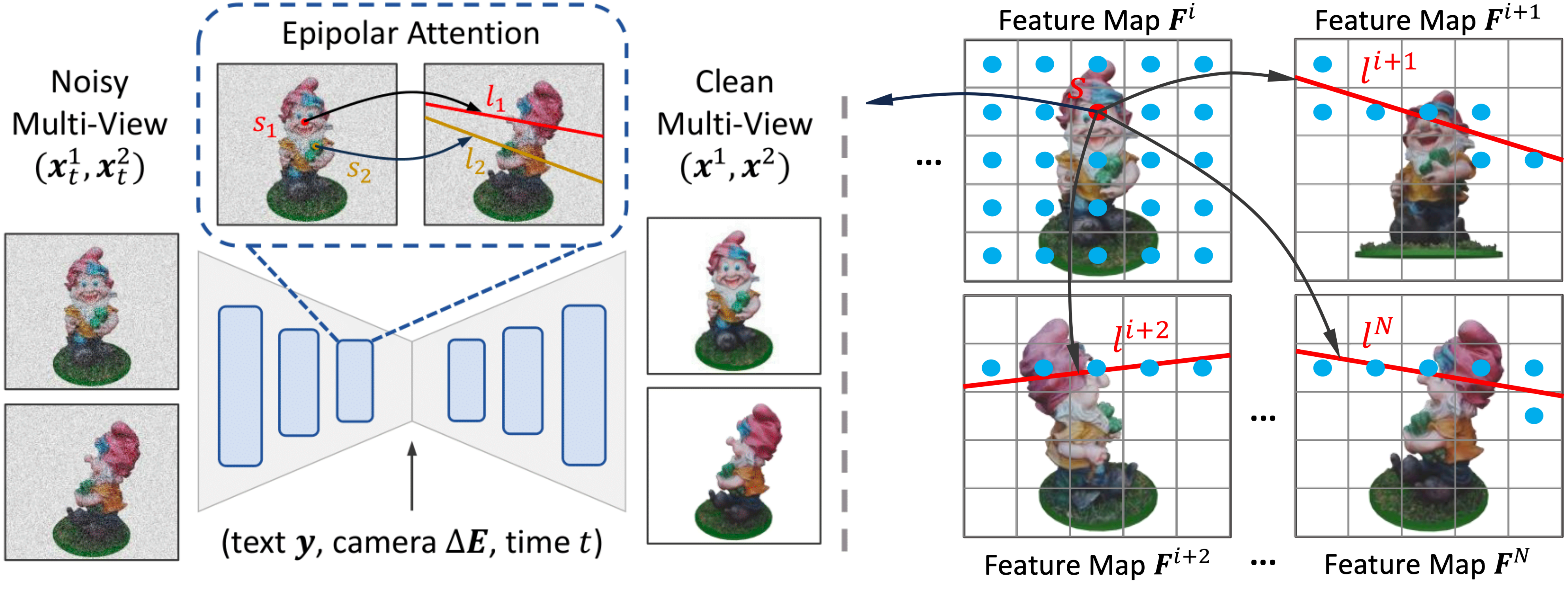
We mask the self-attention to attend along epipolar lines between views.



Multi-view Diffusion Model

SPAD: Epipolar Attention

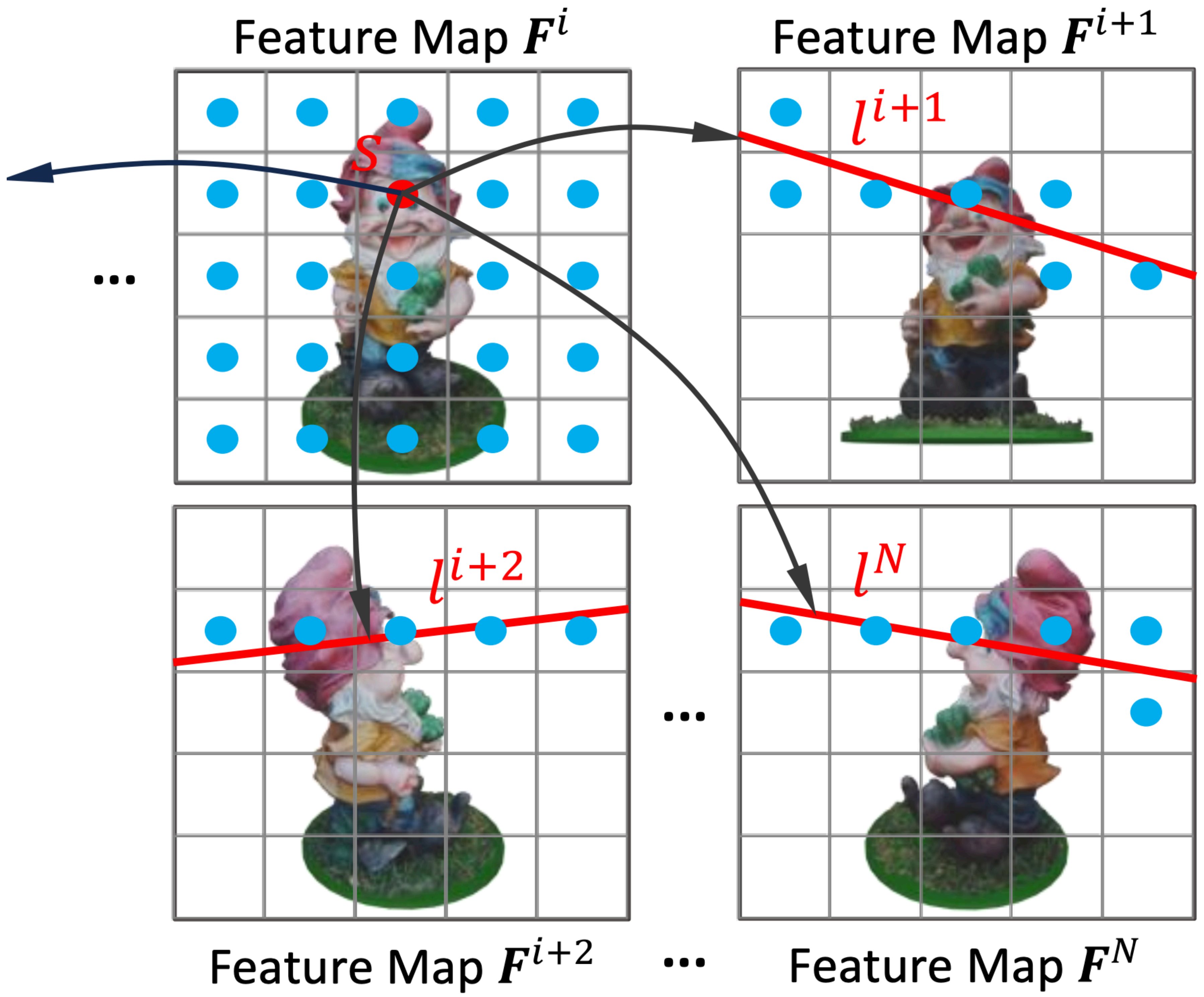
We mask the self-attention to attend along epipolar lines between views.



Multi-view Diffusion Model **Epipolar Attention**

SPAD: Epipolar Attention

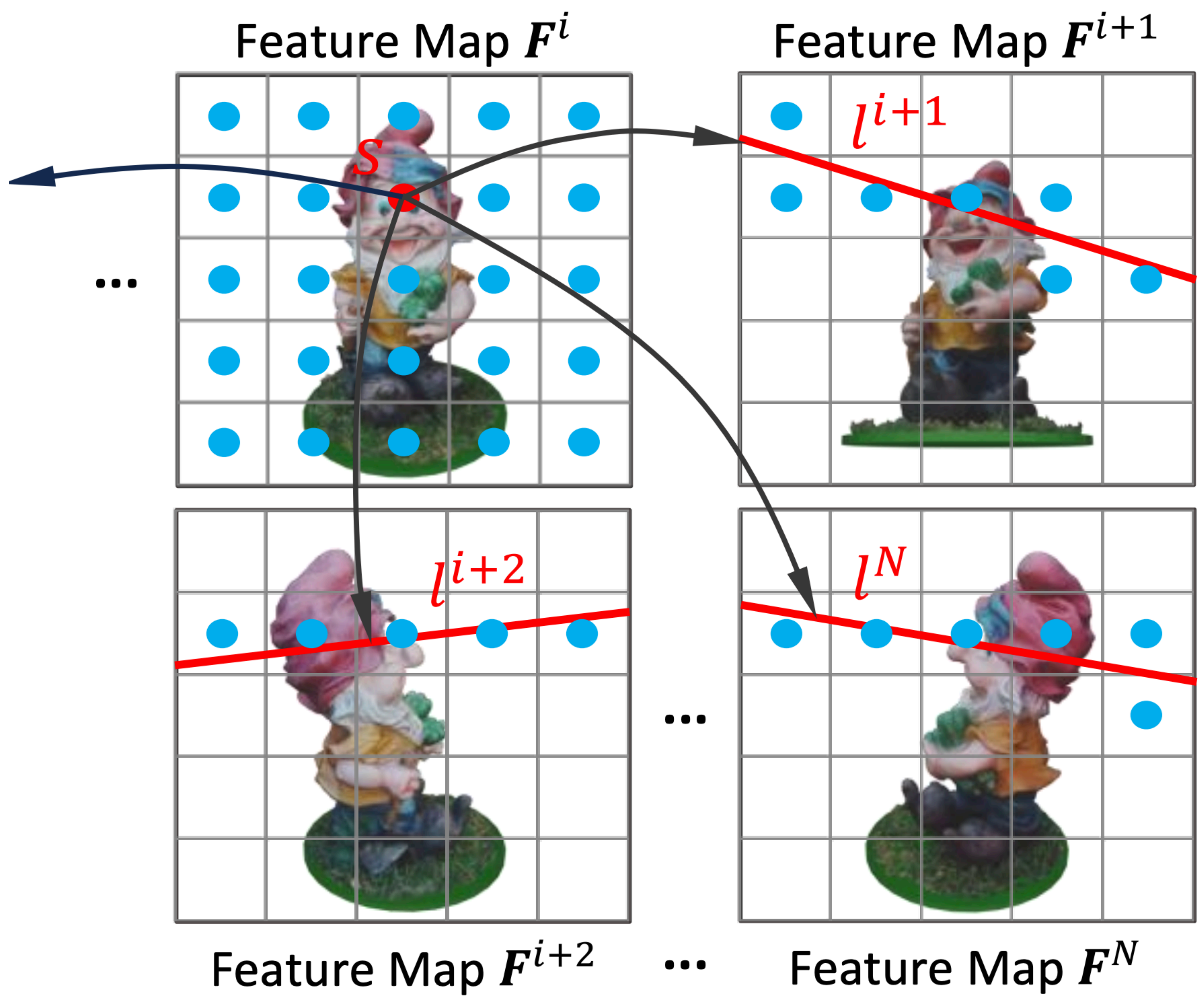
Epipolar lines cannot reason about ray direction — this leads to flipped view generation.



Epipolar Attention

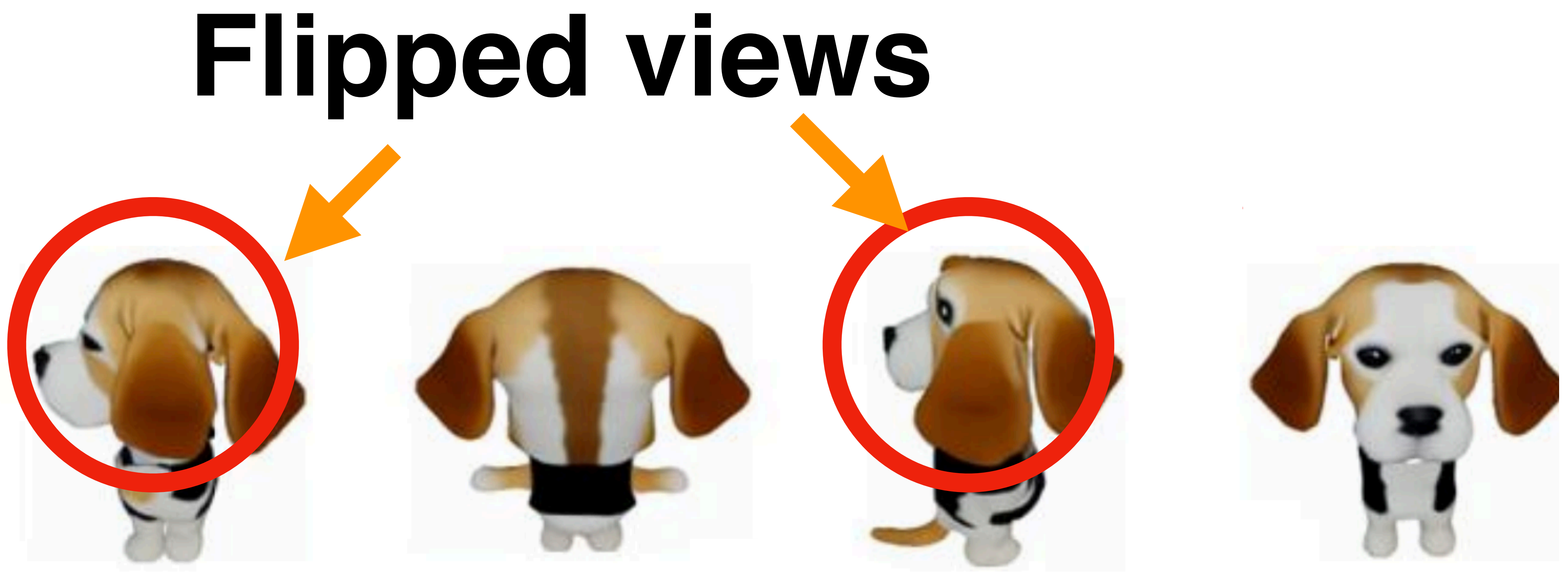
SPAD: Epipolar Attention

Epipolar lines cannot reason about ray direction — this leads to flipped view generation.



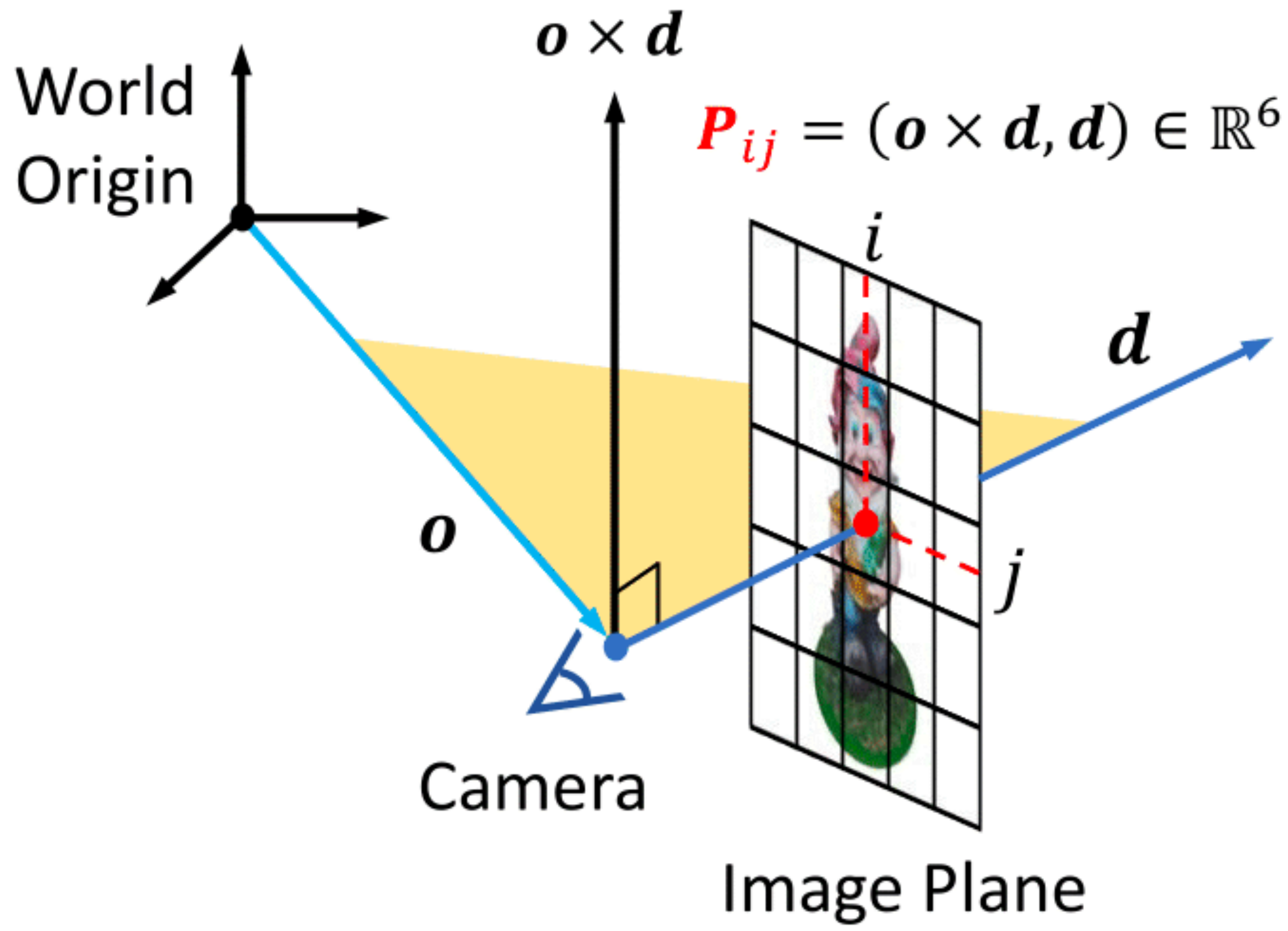
Epipolar Attention

MV DM (Epipolar)



SPAD: Plucker Coordinates

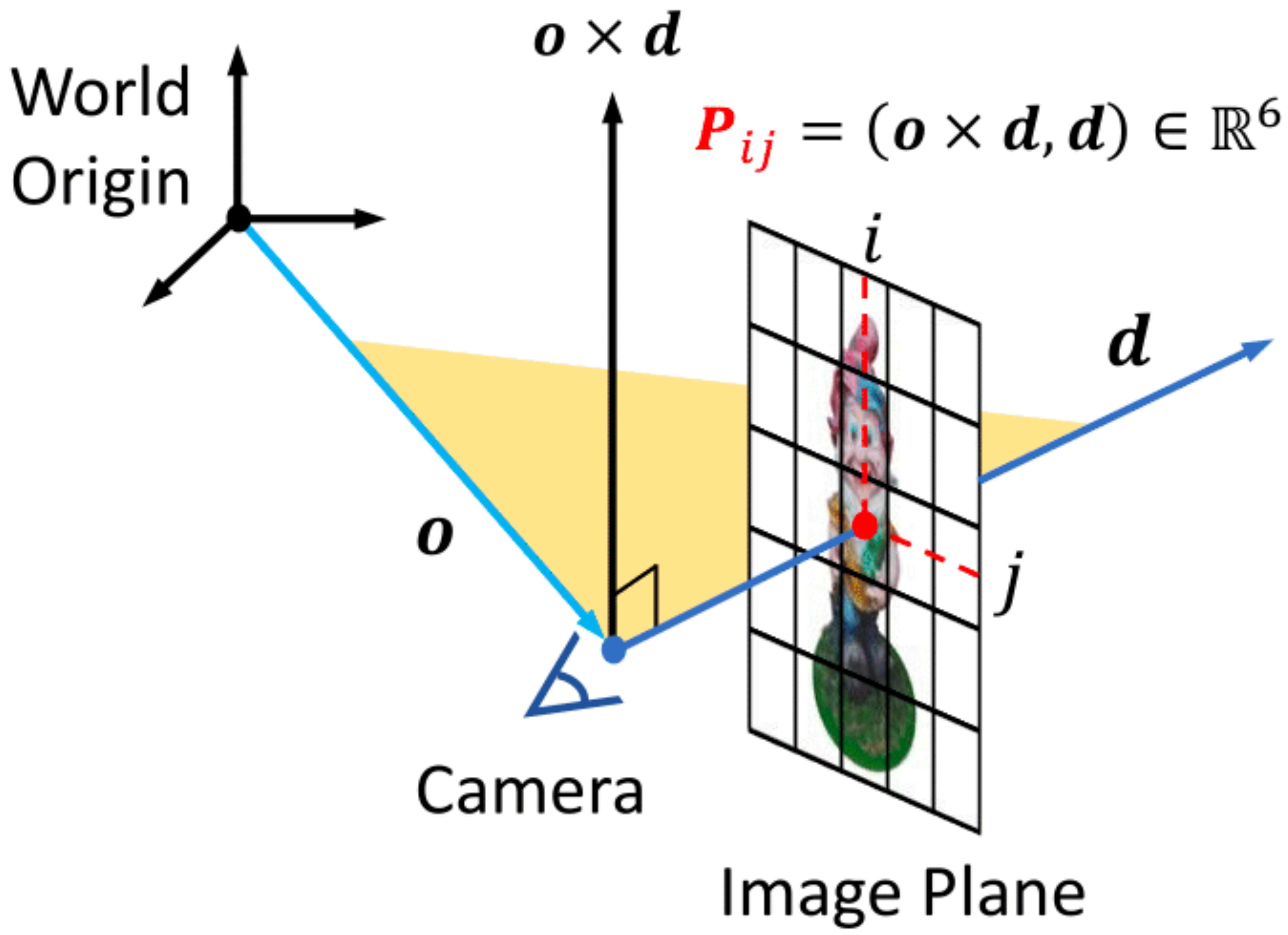
To reason about direction of the rays, we can use plucker coordinates



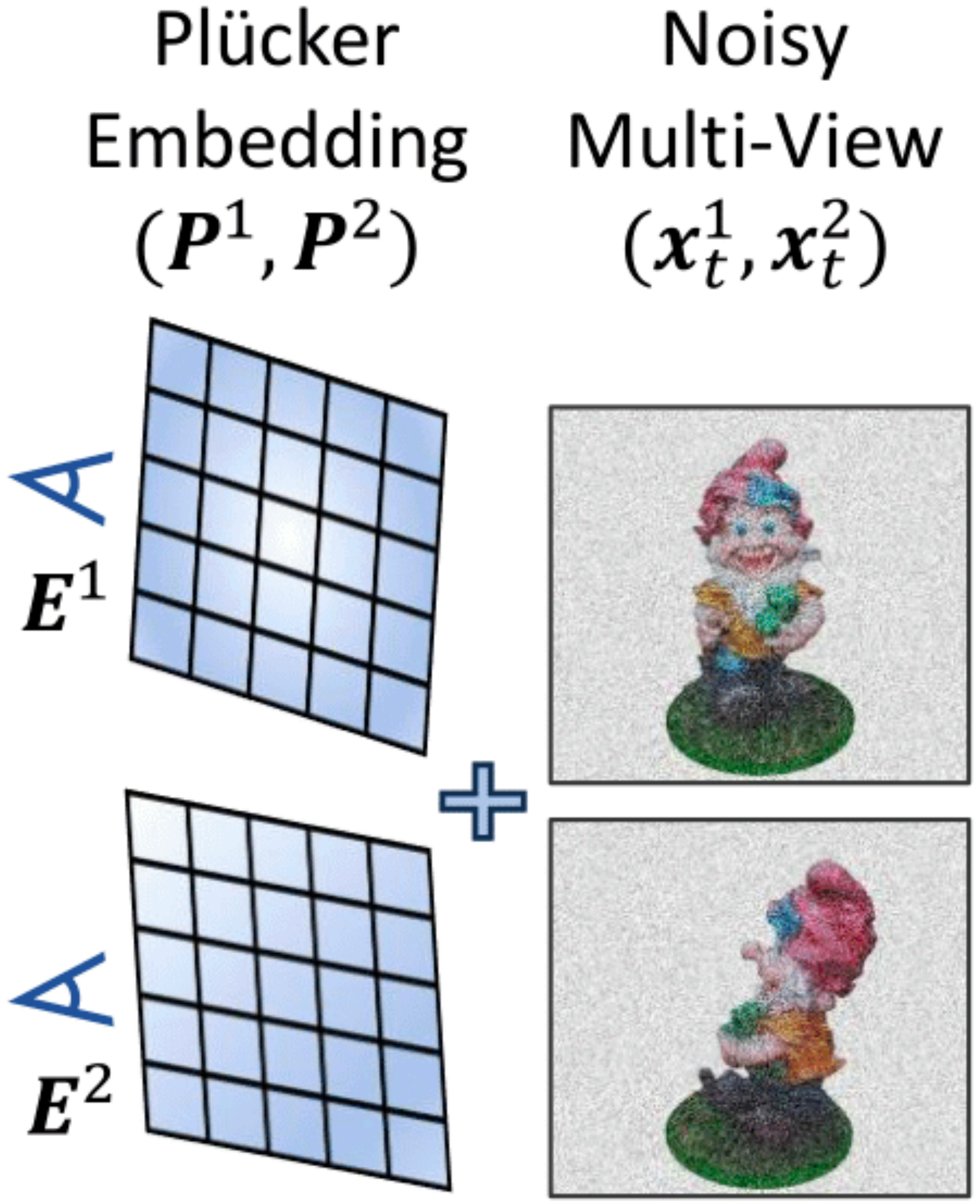
Plucker Coordinates

SPAD: Plucker Coordinates

To reason about direction of the rays, we can use plucker coordinates — as positional encoding.



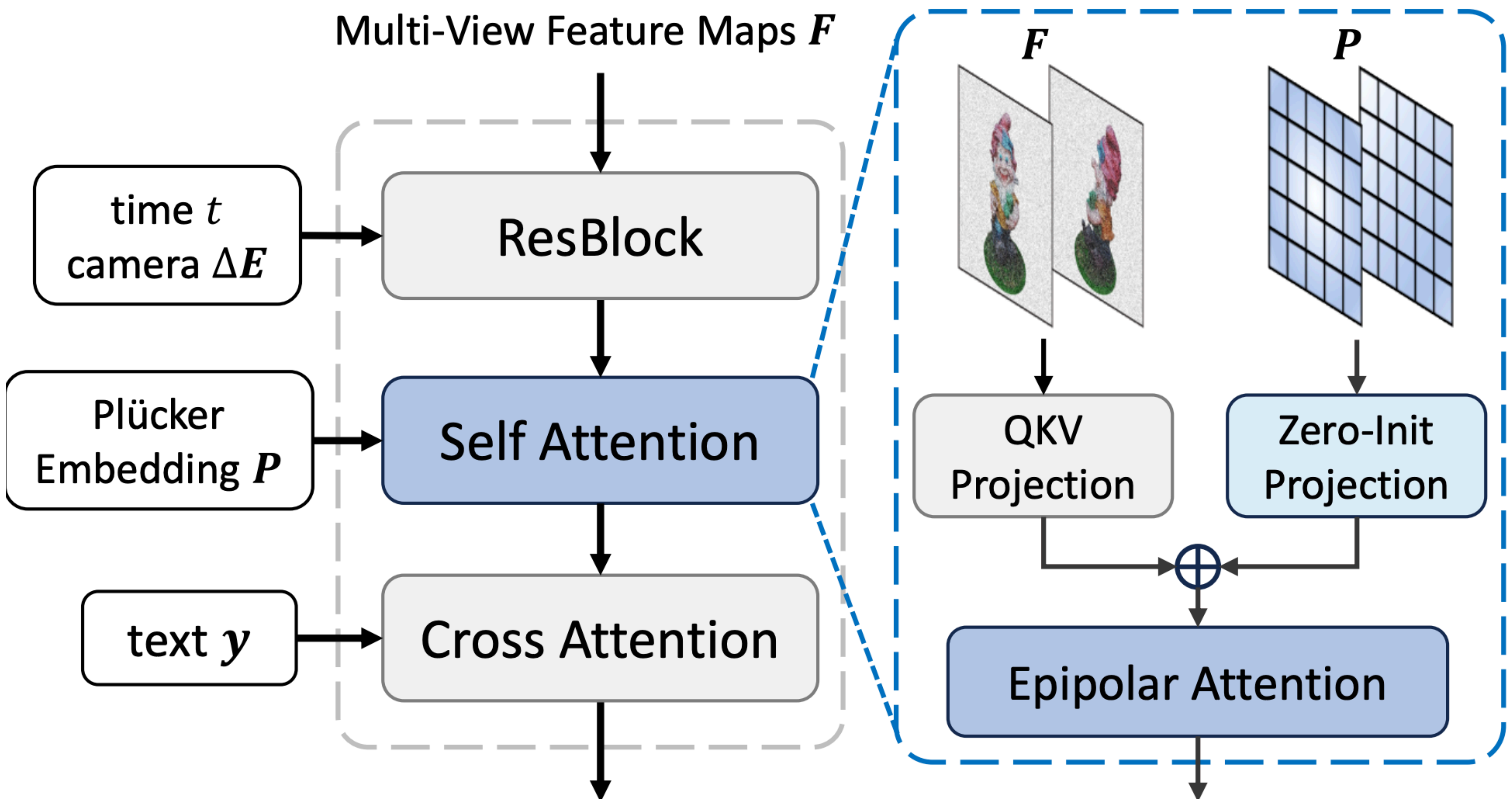
Plucker Coordinates



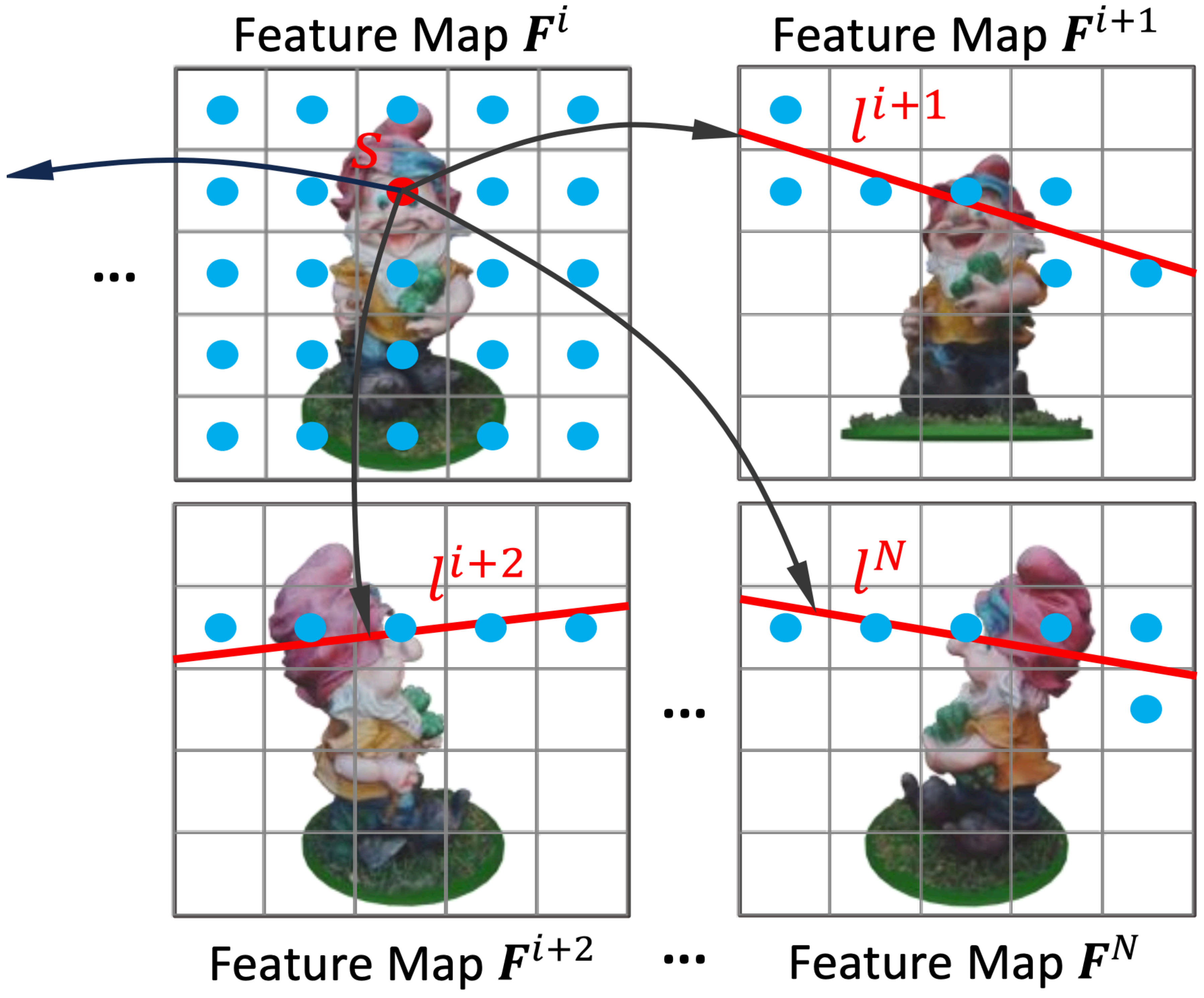
Positional Encoding

SPAD: Overall Modifications.

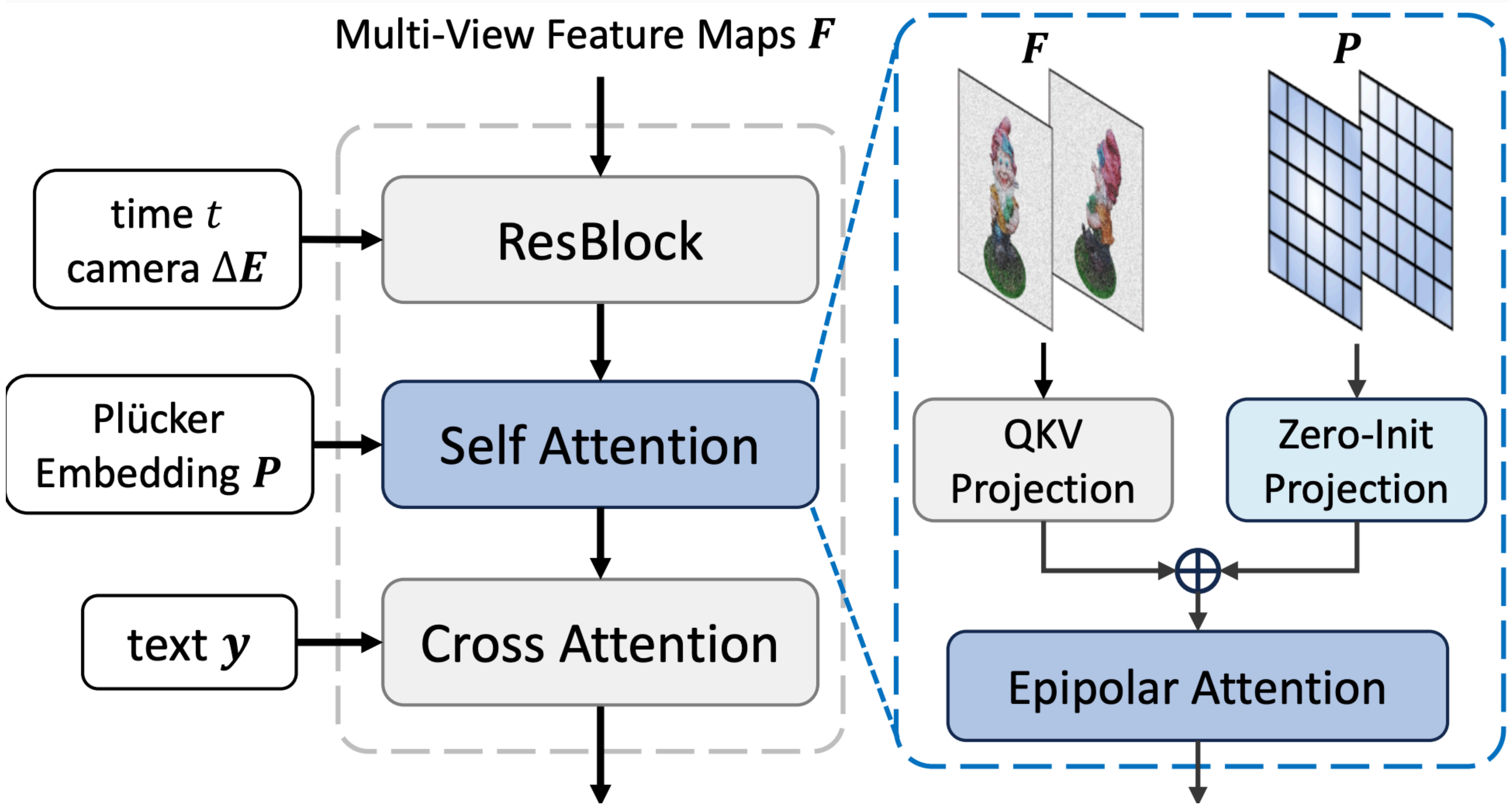
1. Replaced self-attention mask with epipolar mask
2. Insert plucker coordinates via zero-initialized MLP.



SPAD — Spatially Aware Multiview Diffuser



Epipolar Attention



UNet Attention Modification

***SPAD:* Training Details.**

- Training
  - 4-view models on 16-32 H100 GPUs for 100K steps.
  - Effective batch-size ~ 1700 samples.

- Data
  - Quality >> Quantity (finetuning) — trained on 25% of Objaverse (200K) assets filtered with the most like, view, and comment count.

SPAD: Quantitative Results (Text-to-Views)

SPAD outperforms MVDream / SyncDreamer — on **view quality** and **text-to-view alignment**.*

Method	IS ↑	CLIP-score ↑

SPAD: Quantitative Results (Text-to-Views)

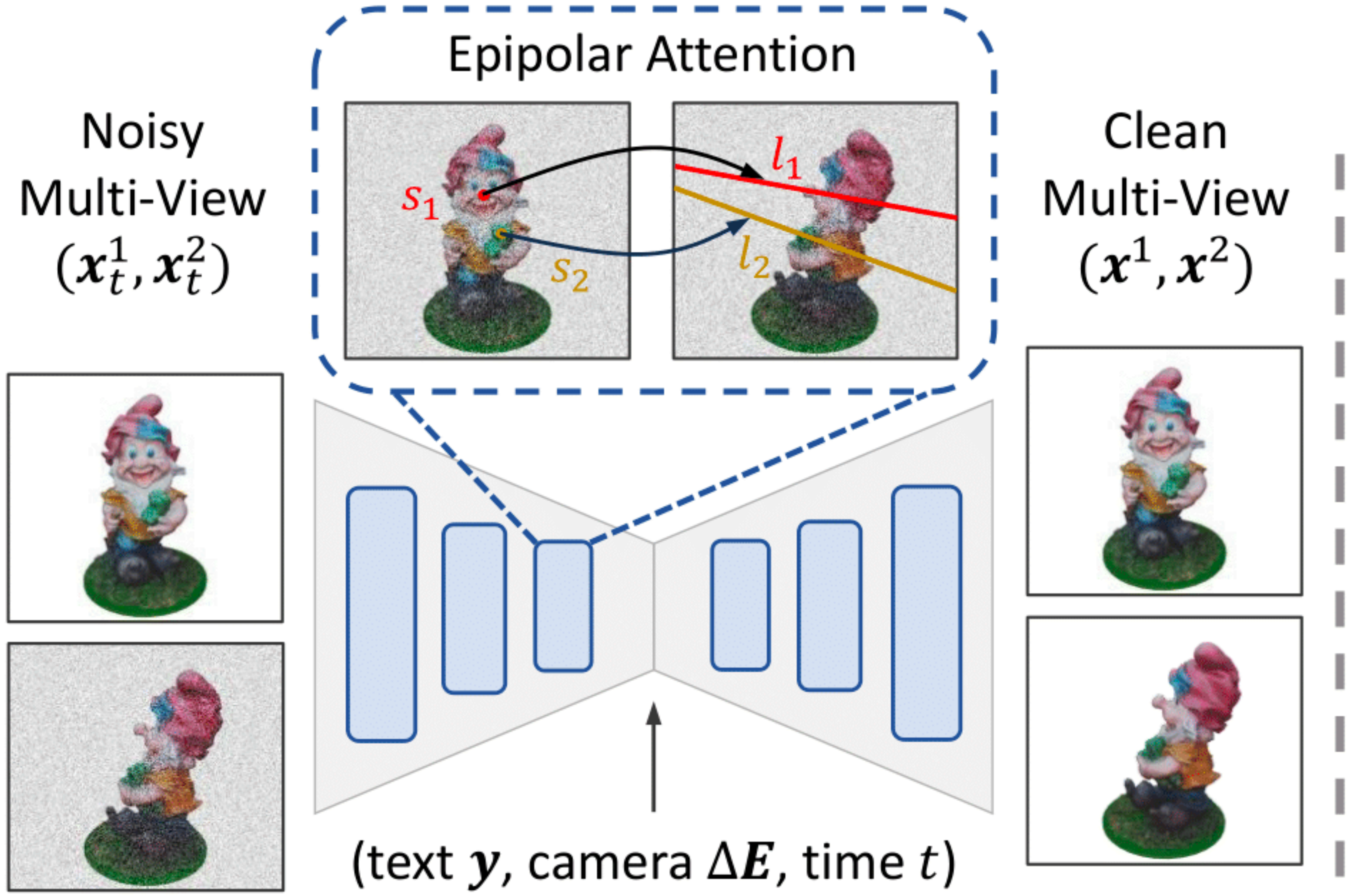
SPAD outperforms MVDream / SyncDreamer — on **view quality** and **text-to-view alignment**.*

Method	IS ↑	CLIP-score ↑
MVDream (v2.1) † [72]	13.36±0.87	30.22±3.83
MVDream (v1.5) † [72]	9.72±0.43	28.55±4.05
SyncDreamer ‡ [46]	11.69±0.24	27.76±4.84
Vanilla MV-DM	11.04±0.81	28.52±3.69
SPAD (**Ours**)	11.18±0.97	29.87±3.33

*When using the same base model SD1.5

SPAD: Quantitative Results (Image-to-Views)

We can **simply freeze one of the views (to input view)** to create a Image-to-Views (NVS) model.

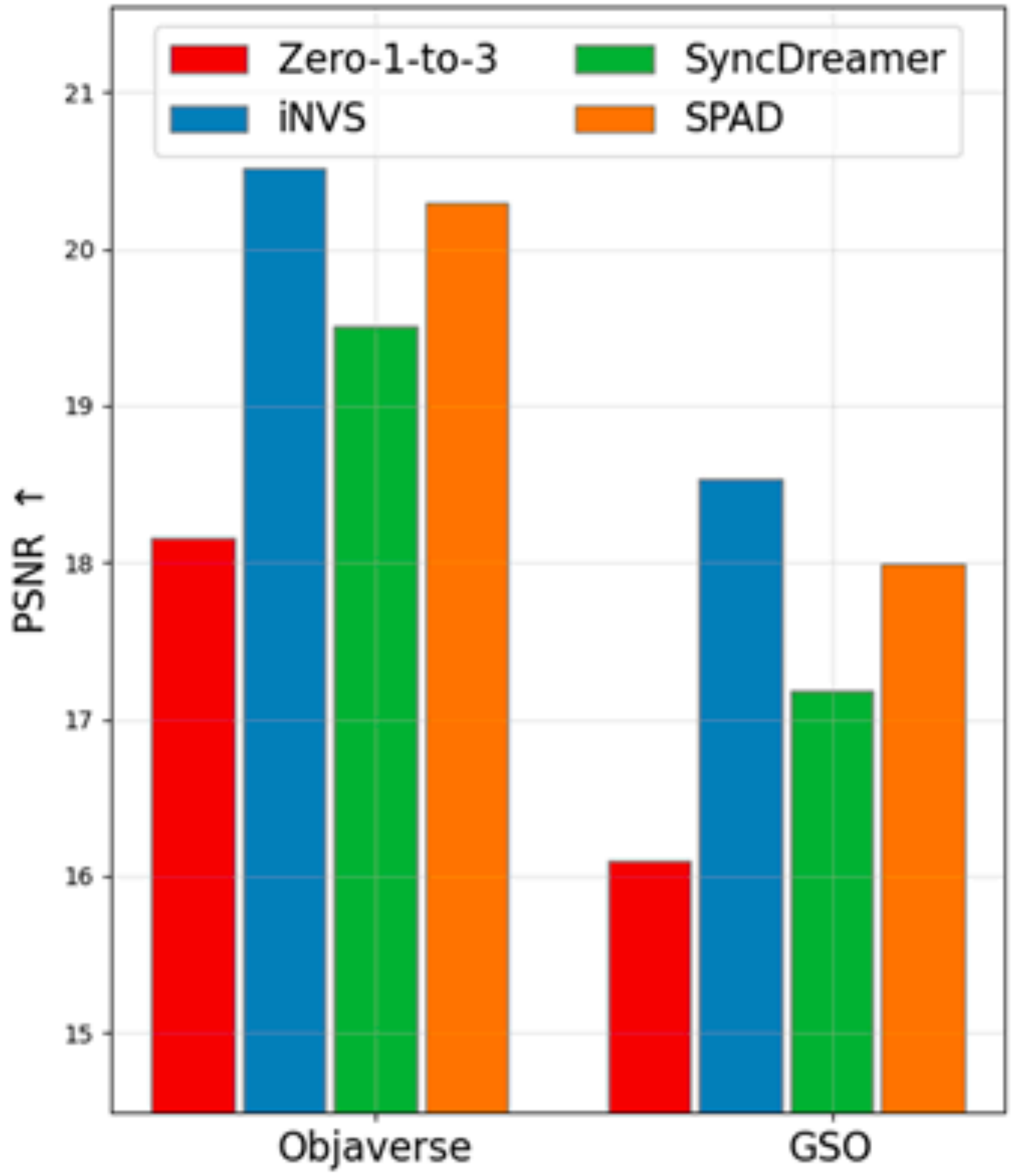


SPAD: Quantitative Results (Image-to-Views)

We use NVS metrics — PSNR / SSIM /LPIPS to compare SPAD with iNVS, SyncDreamer and Zero123.

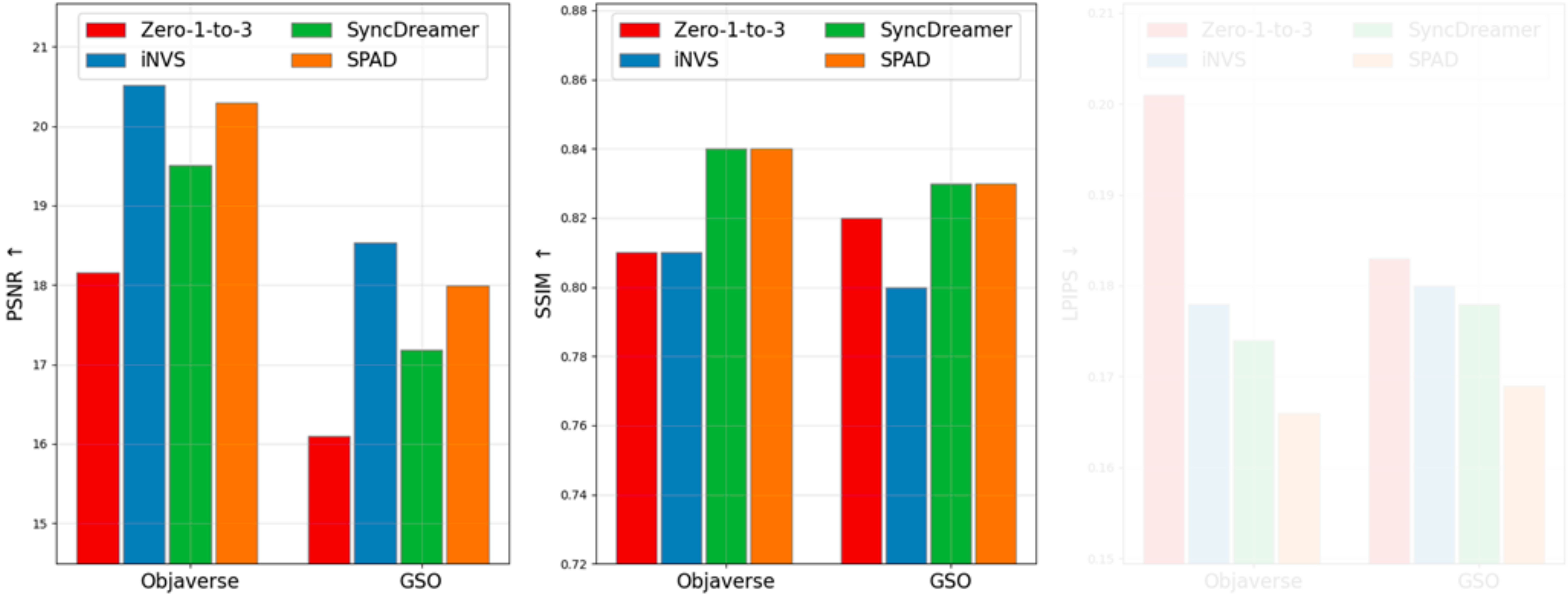
SPAD: Quantitative Results (Image-to-Views)

We use NVS metrics — PSNR / SSIM /LPIPS to compare SPAD with iNVS, SyncDreamer and Zero123.



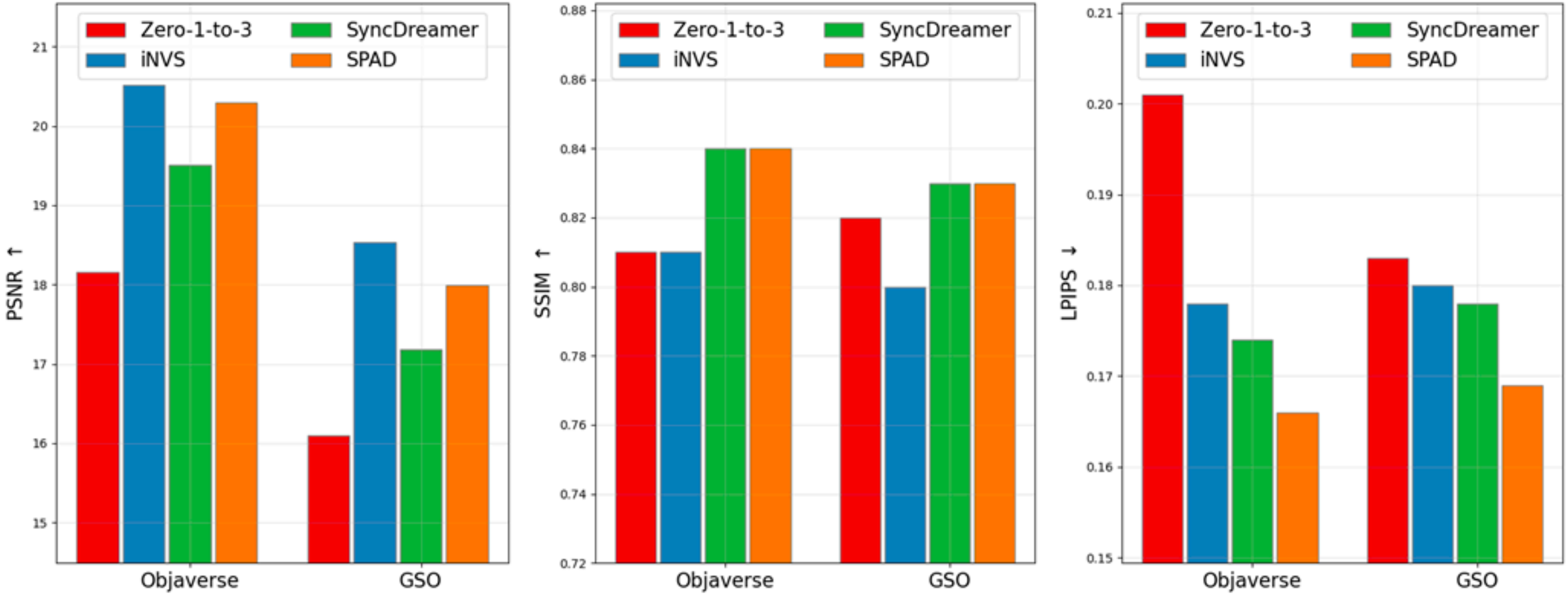
SPAD: Quantitative Results (Image-to-Views)

We use NVS metrics — PSNR / SSIM /LPIPS to compare SPAD with iNVS, SyncDreamer and Zero123.



SPAD: Quantitative Results (Image-to-Views)

We use NVS metrics — PSNR / SSIM /LPIPS to compare SPAD with iNVS, SyncDreamer and Zero123.



SPAD: Qualitative Results (Text-to-Views)

A DSLR photo of a pair of tan cowboy boots, studio lighting, product photography

A wooden chair

A cute steampunk elephant

SPAD: Qualitative Results (Text-to-Views)

A knight's armored metal helmet with gold trim and holes

A small robot with a glass container on its head, metal legs, and a glass top

F-15 Eagle, F-16 Fighter Jet, and F/A-18F Super Hornet aircraft

SPAD: Qualitative Results (Text-to-Views)

A medieval shield with a cross and wooden handle

A black futuristic space helmet with reflective surface

A small biplane flying in the air

A flying red dragon

SPAD: Qualitative Results (Text-to-Views)

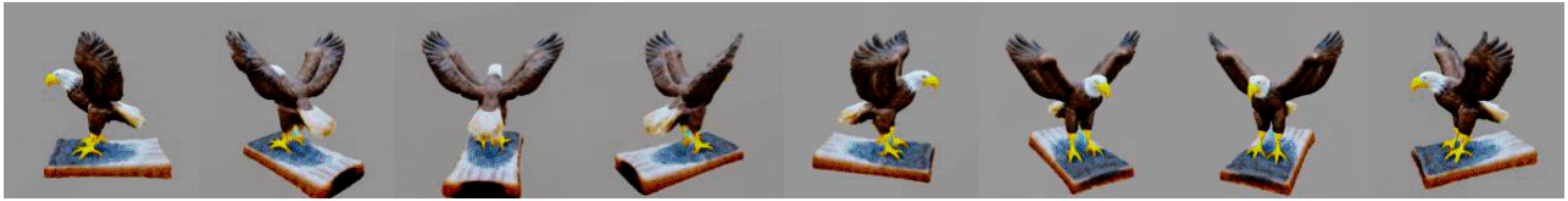
Yellow teapot with a hat on top

An owl with a cat head

A wooden-framed couch with purple upholstery

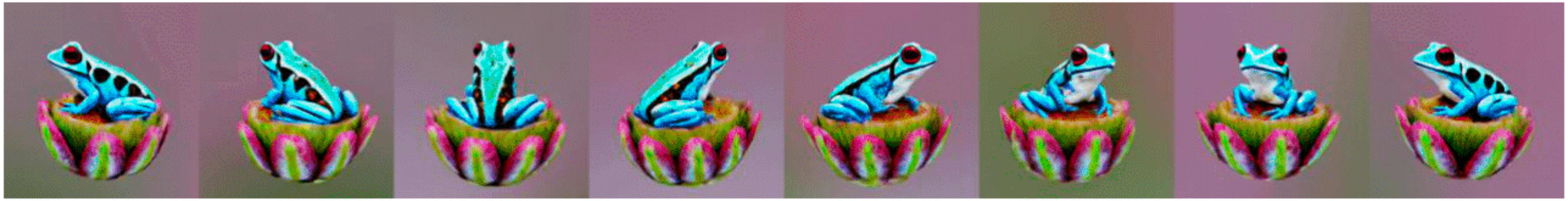
A small stone fountain and cistern with leaves, accompanied by a stone pillar, wall, and old building

SPAD: Text-to-3D Results (Multi-view SDS)



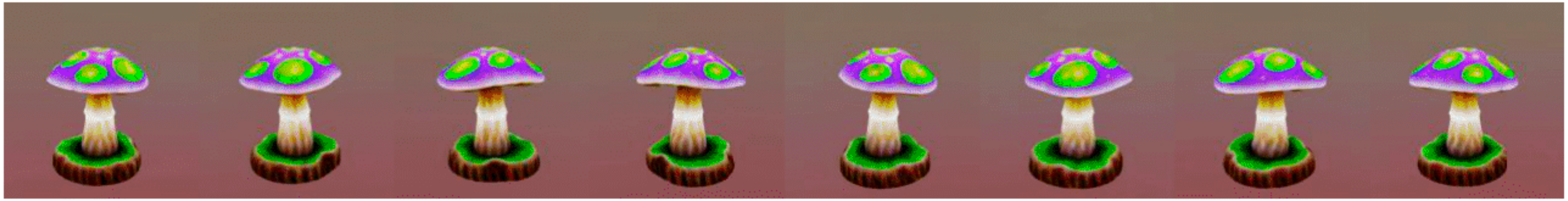
A bald eagle carved out of wood

A bichon frise wearing academic regalia

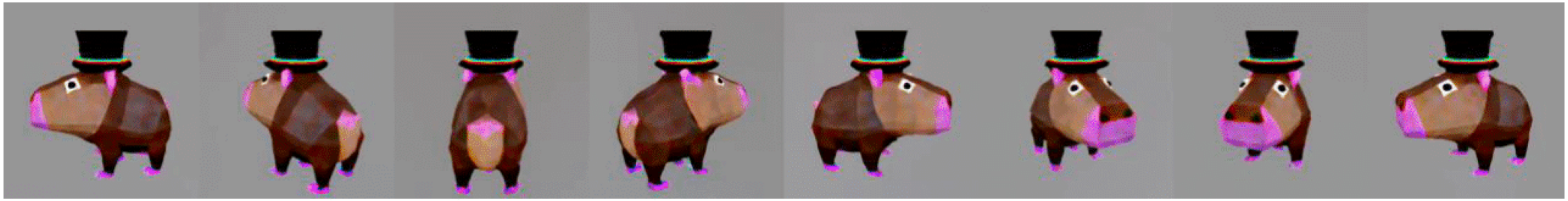


A blue poison-dart frog sitting on a water lily

SPAD: Text-to-3D Results (Multi-view SDS)



A brightly colored mushroom growing on a log



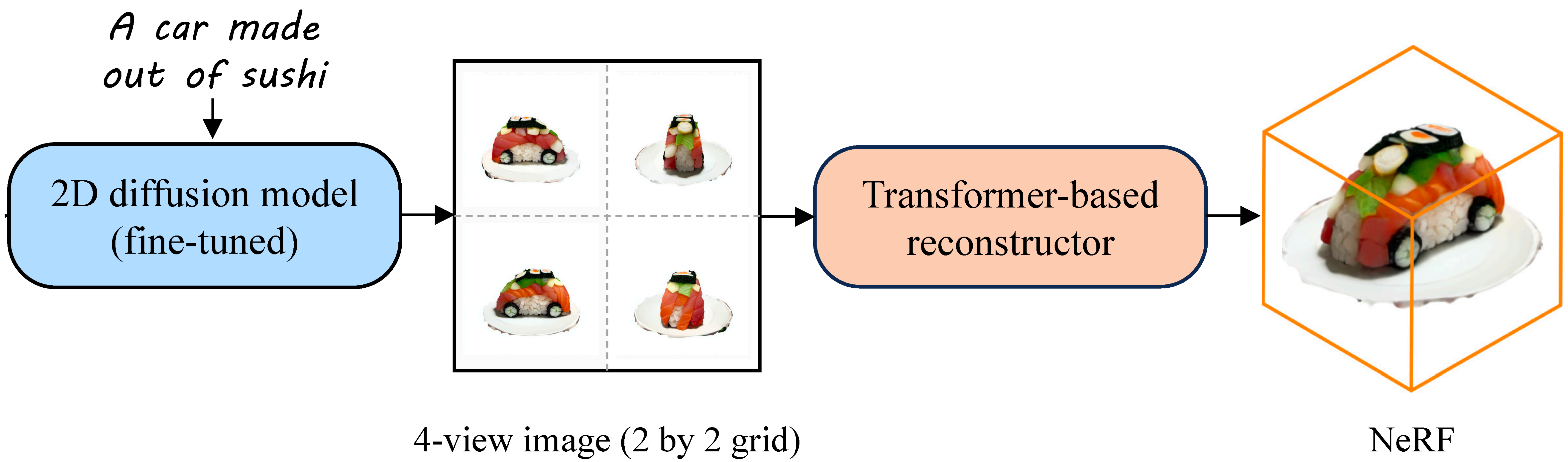
A capybara wearing a top hat, low poly

A beautiful dress made out of garbage bags, on a mannequin. Studio lighting, high quality, high resolution

SPAD: Text-to-3D Results (Multi-view SDS)

SPAD: Text-to-3D Results (Views-to-NeRF)

We push the views generated by SPAD into a transformer-based decoder that generates a NeRF (triplane) in a single forward pass.

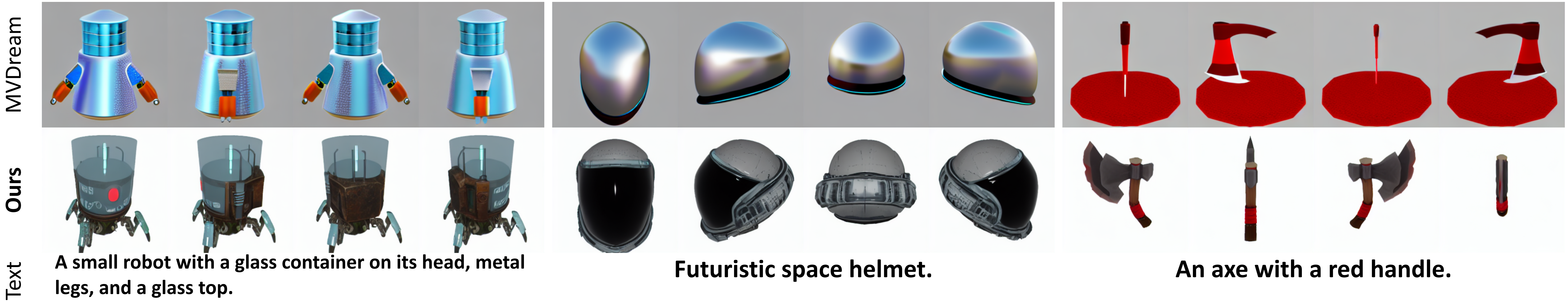


Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model [Jiahao Li, et al.]

SPAD: Text-to-3D Results (Views-to-NeRF)

SPAD: Text-to-3D Results (Views-to-NeRF)

SPAD: Comparison to MVDream

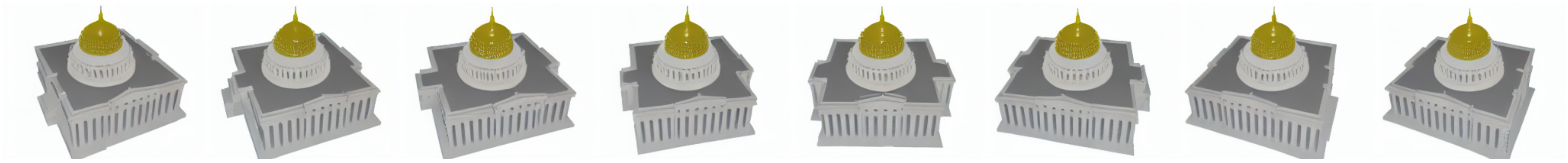


SPAD: Comparison to Zero123

SPAD: Close views in single inference

A red fidget spinner model.

A blue muscle car.



The US Capitol building with a white exterior and golden-yellow dome.

A white Ford F-150 King Ranch pickup truck.

Thanks for listening.

Webpage: https://yashkant.github.io/spad

Yash Kant[1], Ziyi Wu[1], Michael Vasilkovsky[2], Guocheng Qian[2,3], Jian Ren[2], Riza Alp Guler[2], Bernard Ghanem[3], *Sergey Tulyakov[2], *Igor Gilitschenski[1], *Aliaksandr Siarohin[2]

University of Toronto[1] Snap Research[2] KAUST[3]