Diffusion Models for Visual Content Creation

Niloy Mitra, Duygu Ceylan, Paul Guerrero,

Daniel Cohen-Or, Or Patashnik, Chun-Hao Huang, Minhyuk Sung

Part 5: Beyond Single Images

https://geometry.cs.ucl.ac.uk/courses/diffusion4ContentCreation_sigg24/

Presentation Schedule

Introduction to Diffusion Models

Guidance and Conditioning Sampling

Attention

Break

Personalization and Editing

Beyond Single Images

Diffusion Models for 3D Generation

This Section

- Creating “multiple images”
  - in 2D image domain → image montage
  - in 3D spatial domain → multi-views (*without* explicit 3D awareness)
  - in temporal dimension → videos

- Training-free / zero-shot setup
  - Repurposing existing pretrained image diffusion model

- Training setup
  - Video diffusion model

Individual Editing – Per Frame

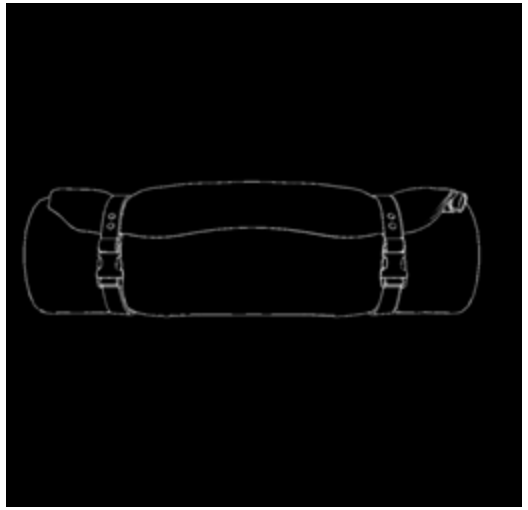
input

per-frame edit

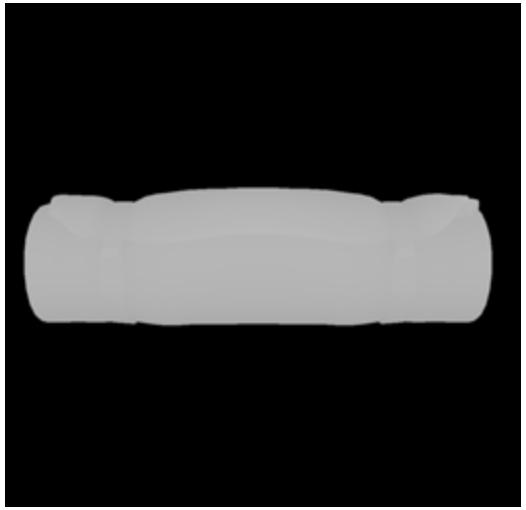
text + depth-conditioned SD

How to Synchronize Multiple Views/Frames?

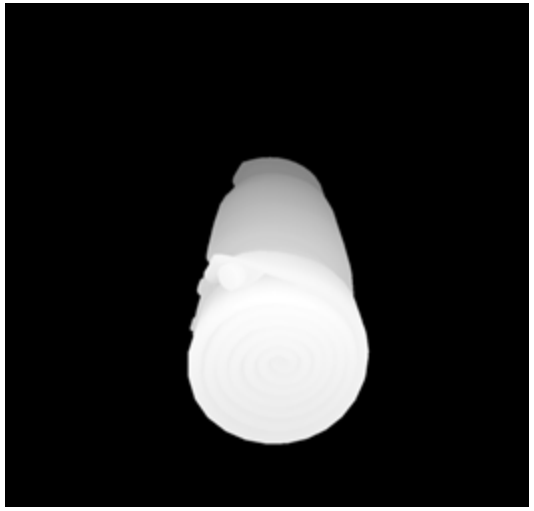
input



input



per-view edits



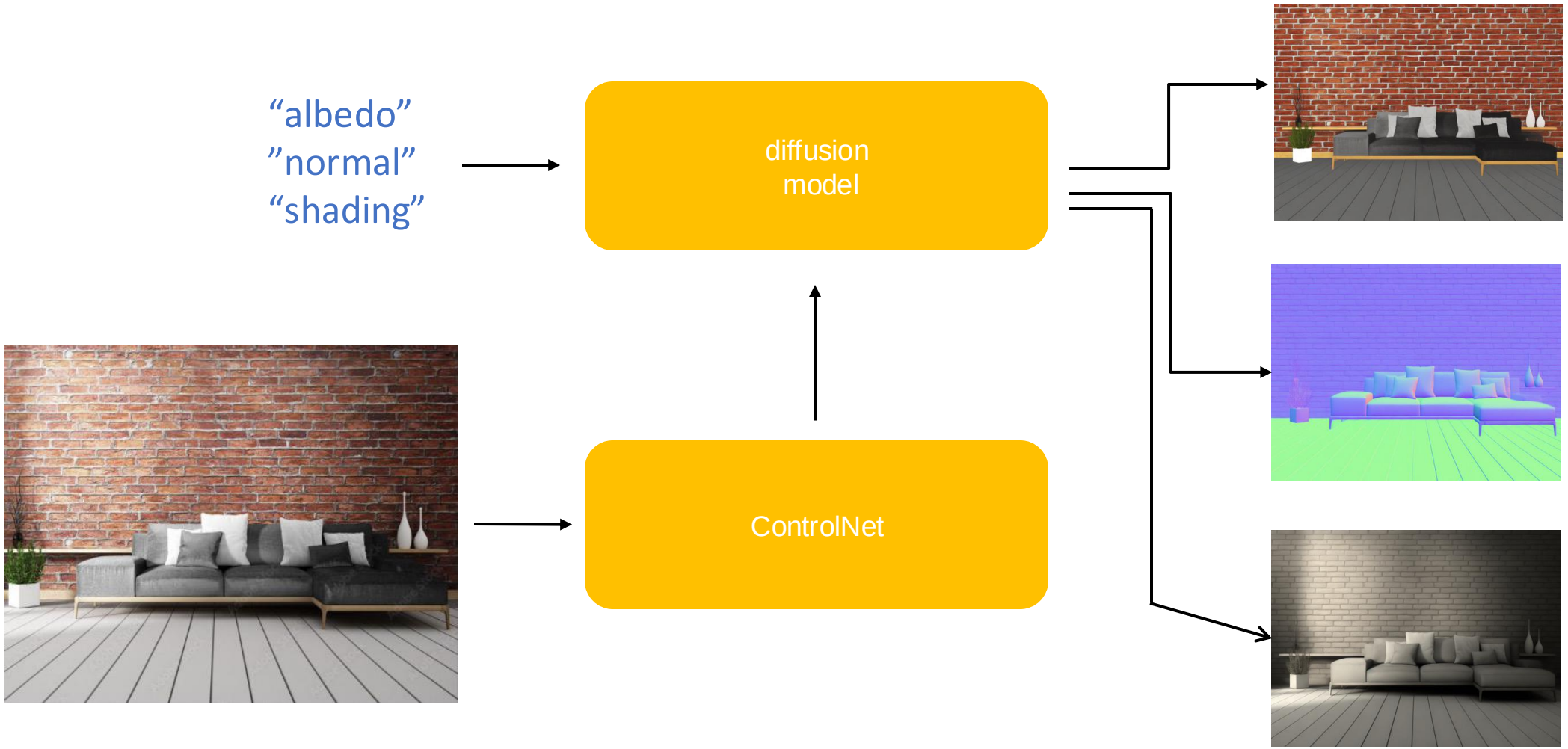
Design Space for Consistency

- Output multiple channels w/ pixel alignment
- Synchronize multiple images w/ over overlapping regions
- Synchronize multiple views via fixed (3D) geometry
- Synchronize multiple views w/ geometric prior
- Synchronize multiple frames

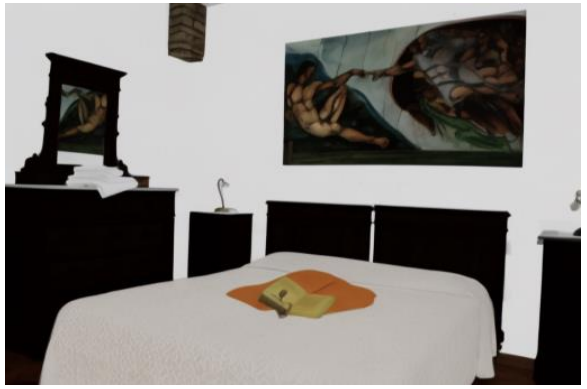
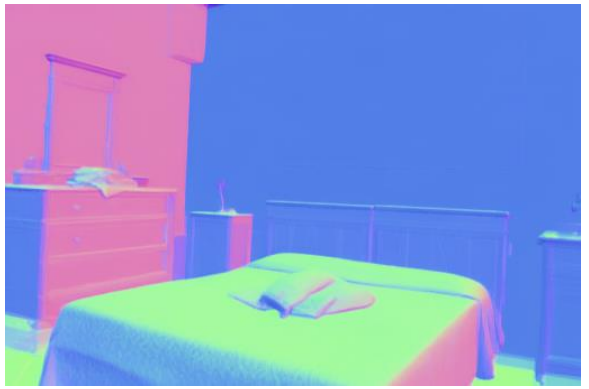
Design Space for Consistency

- **Output multiple channels w/ pixel alignment**
- Synchronize multiple images w/ over overlapping regions
- Synchronize multiple views via fixed (3D) geometry
- Synchronize multiple views w/ geometric prior
- Synchronize multiple frames

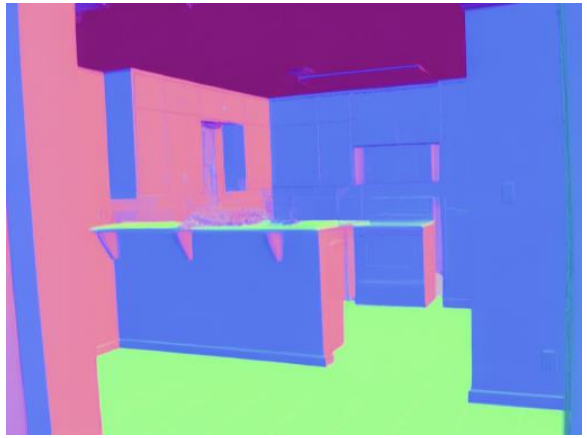
Joint Intrinsic Layers from Diffusion Models



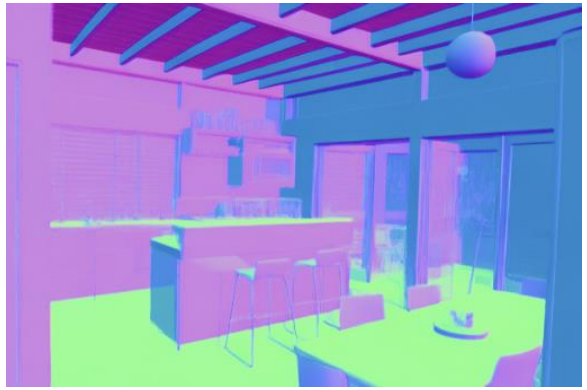
[IntrinsicDiffusion: Joint Intrinsic Layers from Latent Diffusion Models, Luo et al., SIGGRAPH 2024]



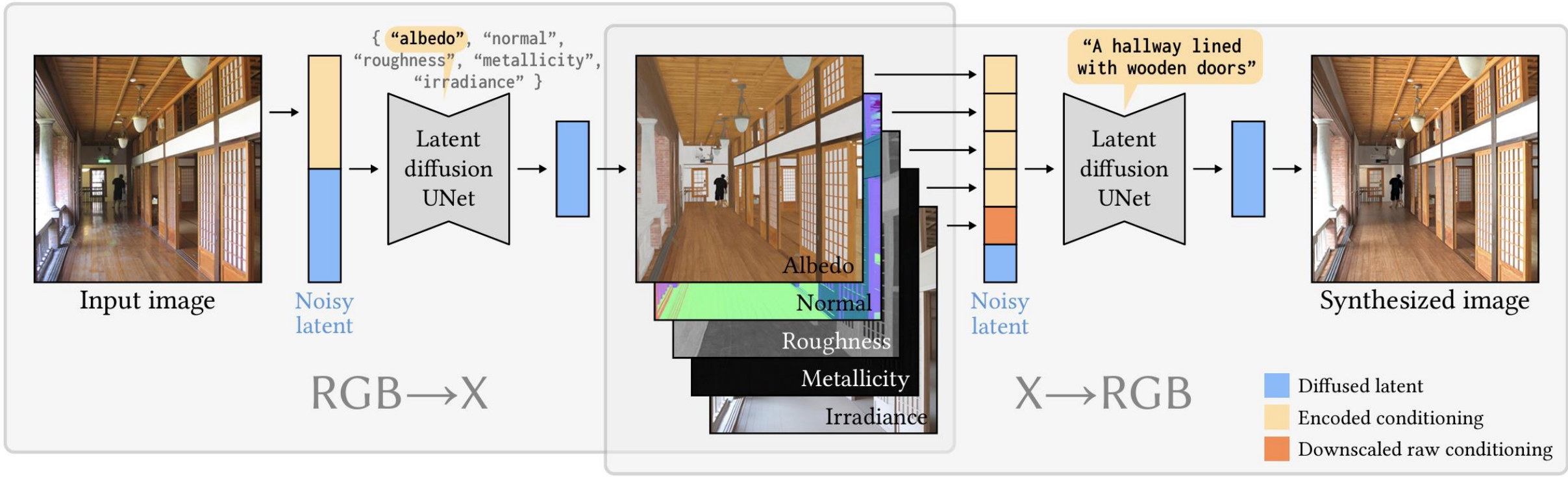







input RGB albedo surface normal shading

RGBX: Multi-channel Input or Output

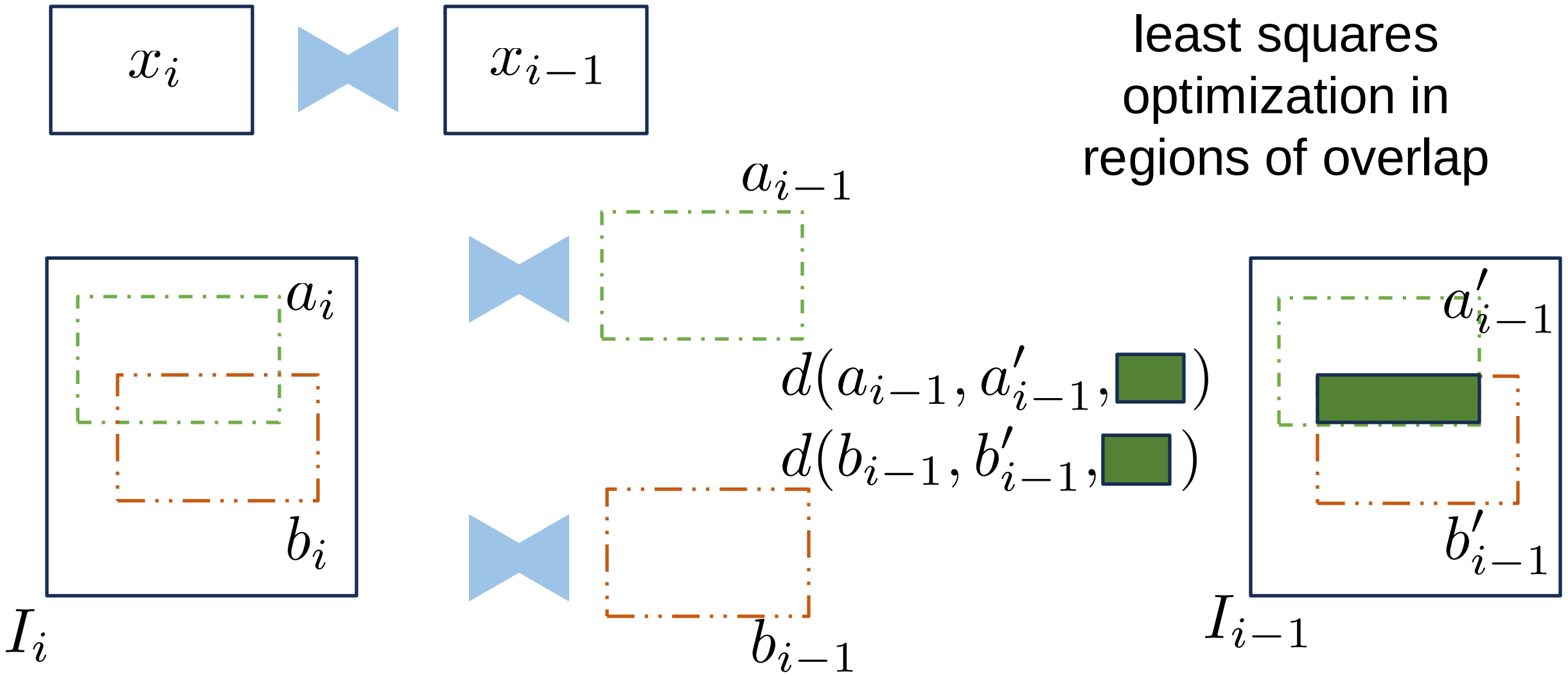


[RGB↔X:Image decomposition and synthesis using material- and lighting-aware diffusion models, Zeng et al., SIGGRAPH 2024]

Design Space for Consistency

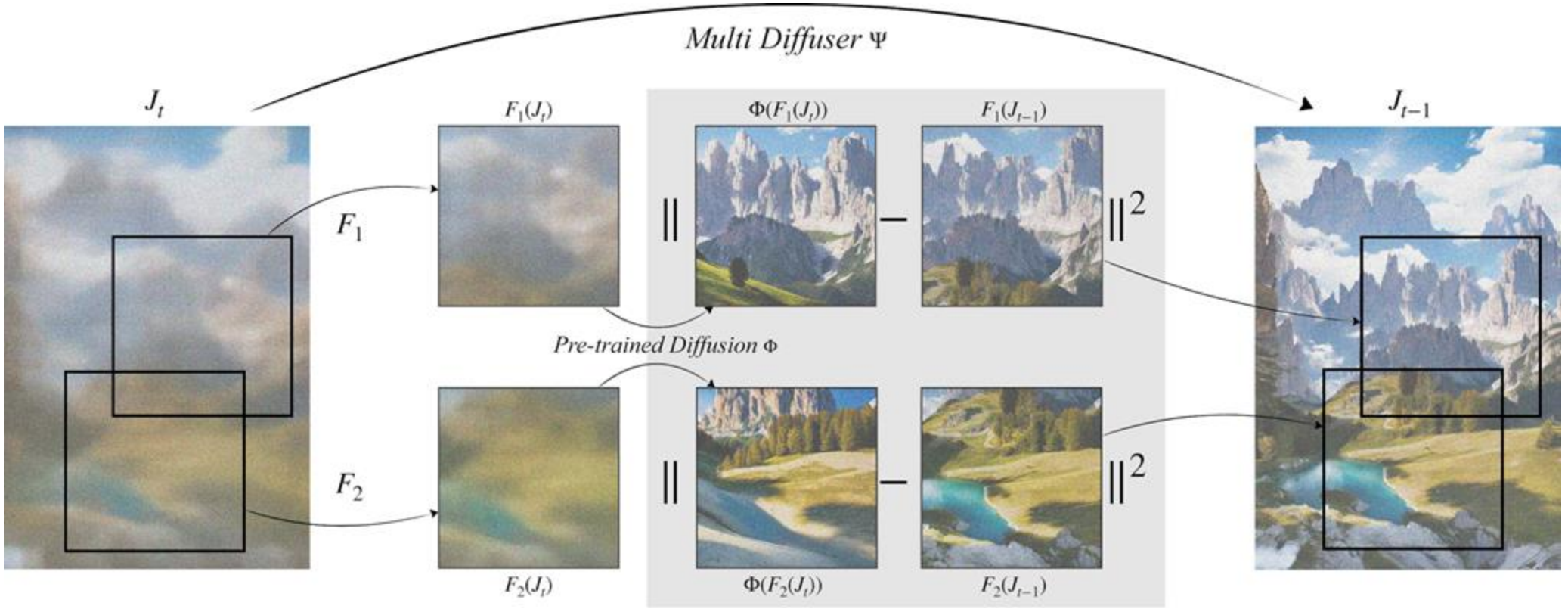
- Output multiple channels w/ pixel alignment
- **Synchronize multiple images w/ over overlapping regions**
- Synchronize multiple views via fixed (3D) geometry
- Synchronize multiple views w/ geometric prior
- Synchronize multiple frames

Synchronizing Colors or Latent Features



Synchronizing Colors or Latent Features

- Averaging noisy latents of the same pixels



[Bar-Tal et al., MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation, *ICML'23*]

Loss-guided Denoising

$\mathcal{D}(\mathbf{x}_t)$ noisy image at t=45

$\mathcal{D}(\phi_\theta(\mathbf{x}_t, t))$ predicted x0 at t=45

$\mathcal{D}(\mathbf{x}_0)$ final denoised x0

$$\hat{\mathbf{x}}_t^{(i)} = \mathbf{x}_t^{(i)} - w\nabla_{\mathbf{x}_t^{(i)}}\mathcal{L}\left(\mathcal{D}(\phi_\theta(\mathbf{x}_t^{(i)}, t)), \mathcal{D}(\phi_\theta(\mathbf{x}_t^{(0)}, t))\right)$$

Loss $\mathcal{L}$: LPIPS score on the predicted denoised x0

Lee et al., SyncDiffusion: Coherent Montage via Synchronized Joint Diffusions, *NeurIPS'23*

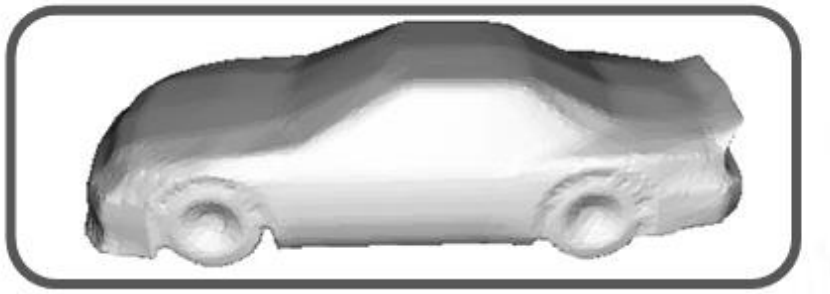
without loss-guided denoising

with loss-guided denoising

Design Space for Consistency

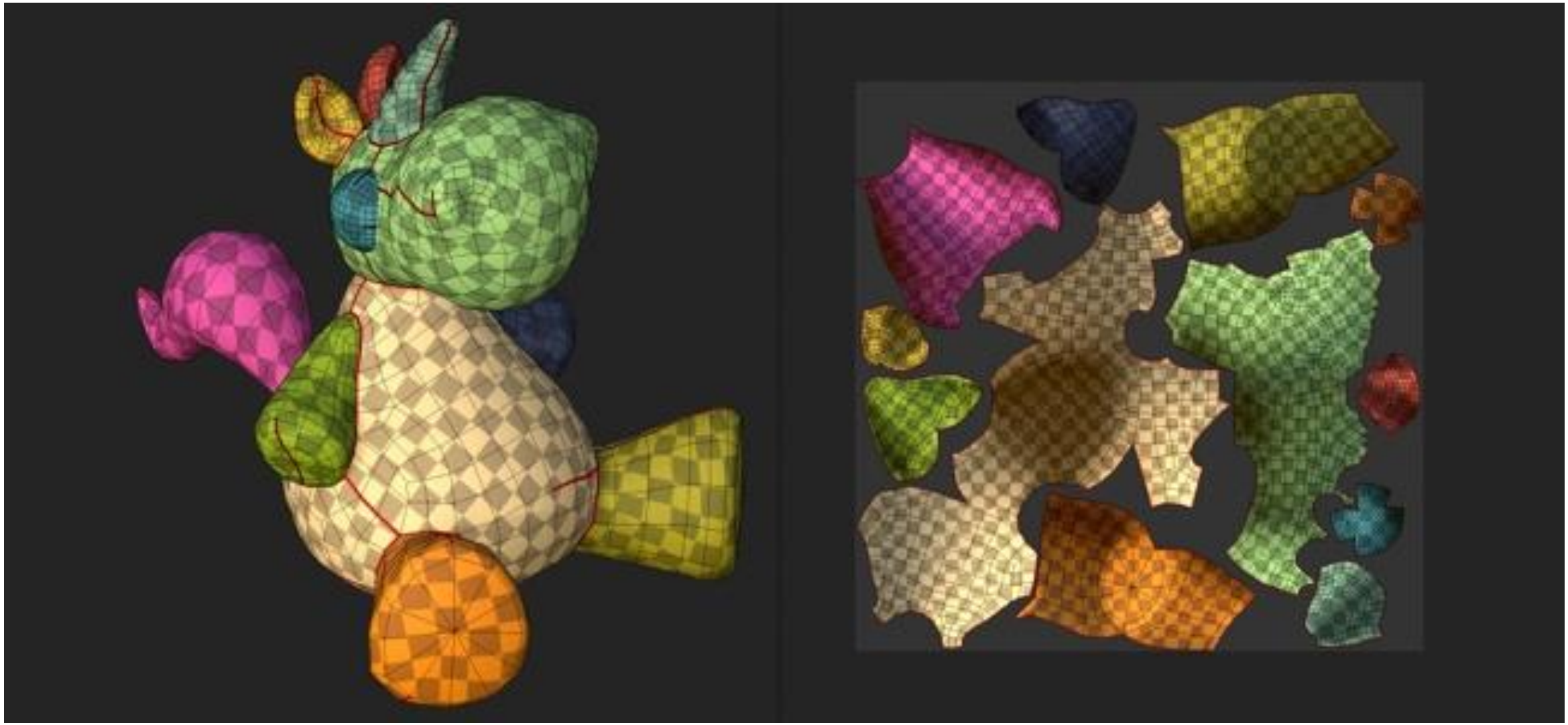
- Output multiple channels w/ pixel alignment
- Synchronize multiple images w/ over overlapping regions
- **Synchronize multiple views via fixed (3D) geometry**
- Synchronize multiple views w/ geometric prior
- Synchronize multiple frames

Correspondence via Fixed Geometry

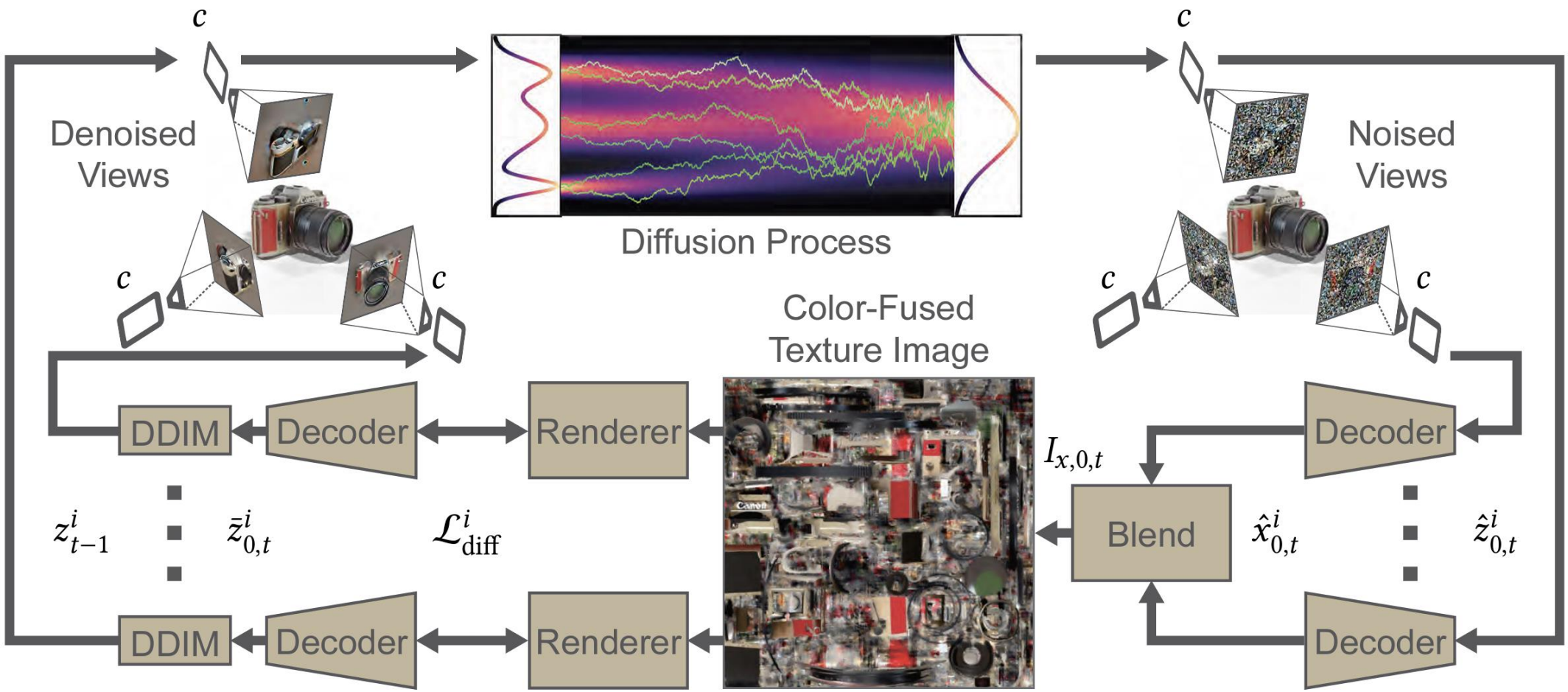


Zhang et al., TexPainter: Generative Mesh Texturing with Multi-view Consistency, SIGGRAPH 2024

Correspondence via Given UV Maps



Correspondence via Fixed Geometry



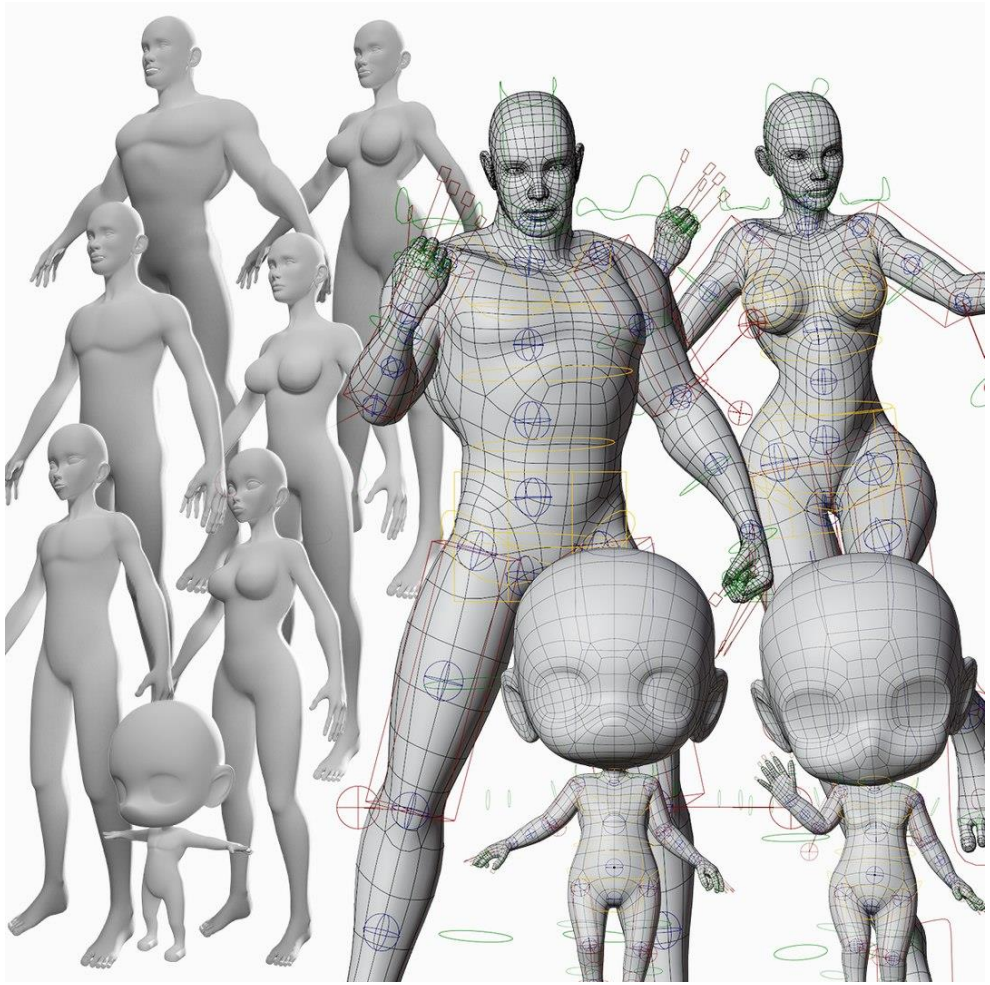
Zhang et al., TexPainter: Generative Mesh Texturing with Multi-view Consistency, SIGGRAPH 2024

Correspondence via Given Texture Space

motorbike, Ducati Hypermotard 939

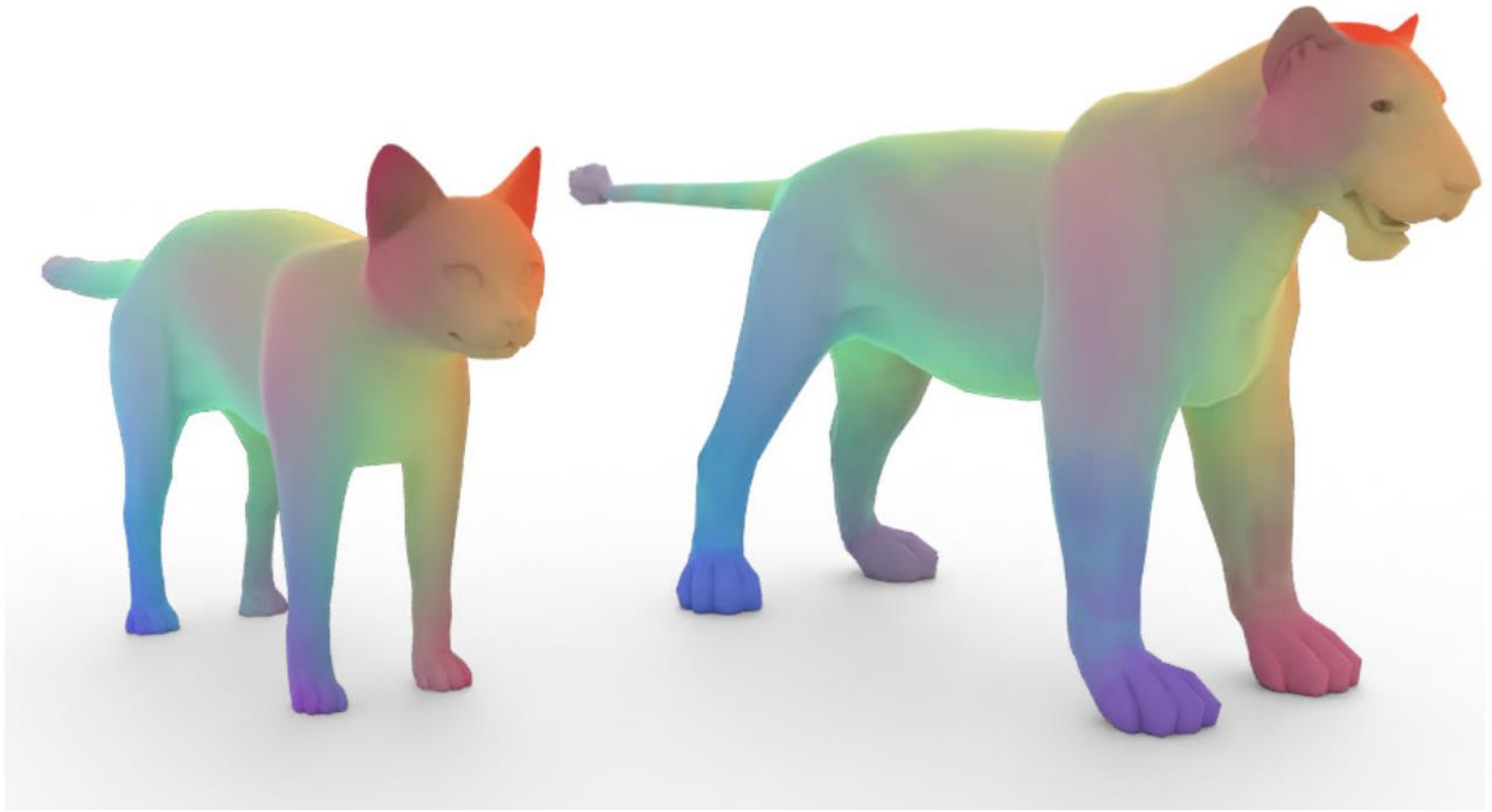
Other related works, TEXTure, Fantasia3D, TexFusion, …

Meshes in the Wild

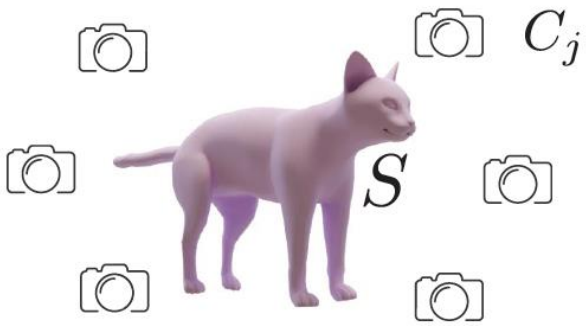


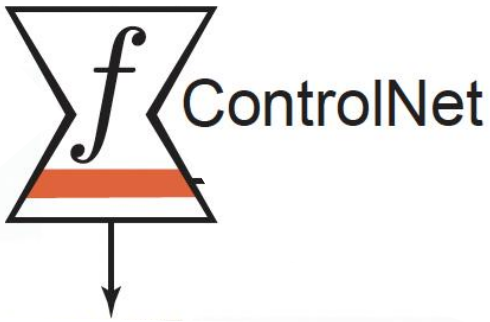
[Meshes from TurboSquid]

Diff3D can Decorate them w/ Features



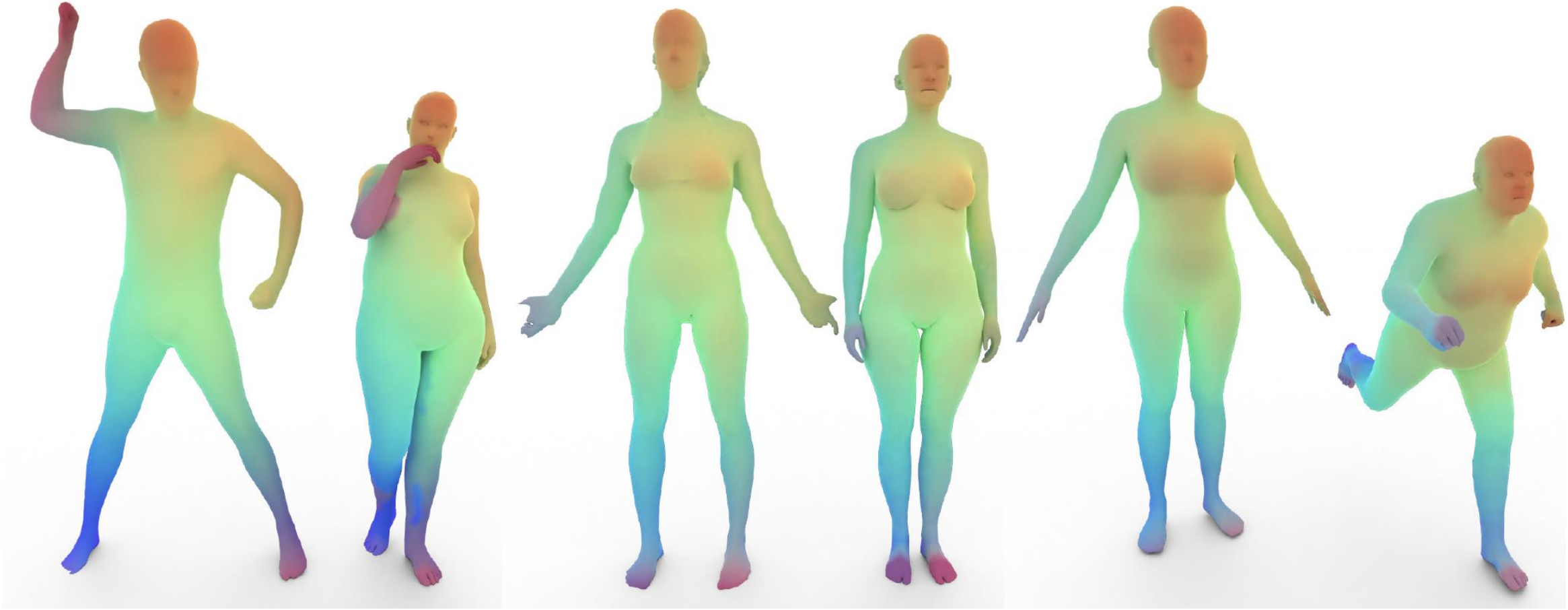
Method Overview





Dutt et al., Diffusion 3D Features (Diff3F): Decorating Untextured Shapes with Distilled Semantic Features, *CVPR'24*

Result Gallery



Diffusion Features

Dutt et al., Diffusion 3D Features (Diff3F): Decorating Untextured Shapes with Distilled Semantic Features, *CVPR'24*

Design Space for Consistency

- Output multiple channels w/ pixel alignment
- Synchronize multiple images w/ over overlapping regions
- Synchronize multiple views via fixed (3D) geometry
- **Synchronize multiple views w/ geometric prior**
- Synchronize multiple frames

Combining 2D & 3D Data

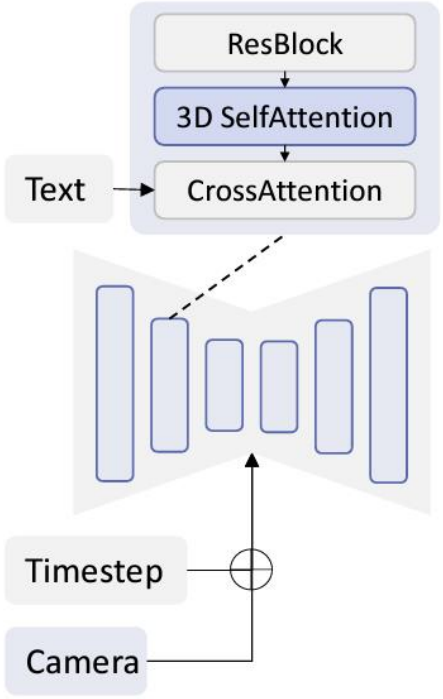
MVImage

3D model

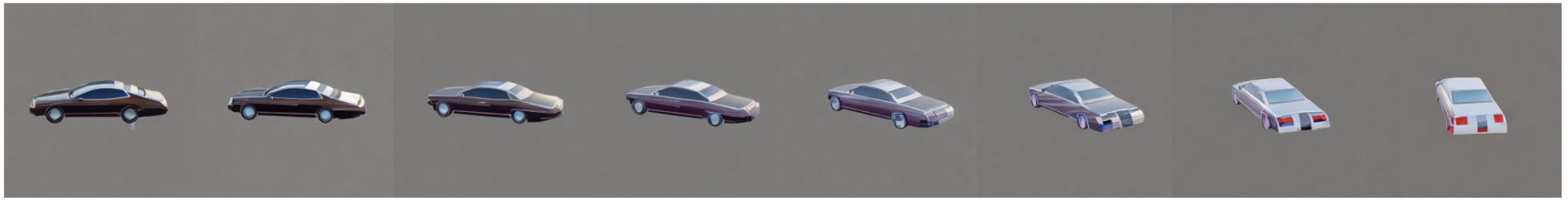
Rendered images

Training Loss

Multi-view Generation



Multi-view Diffusion UNet



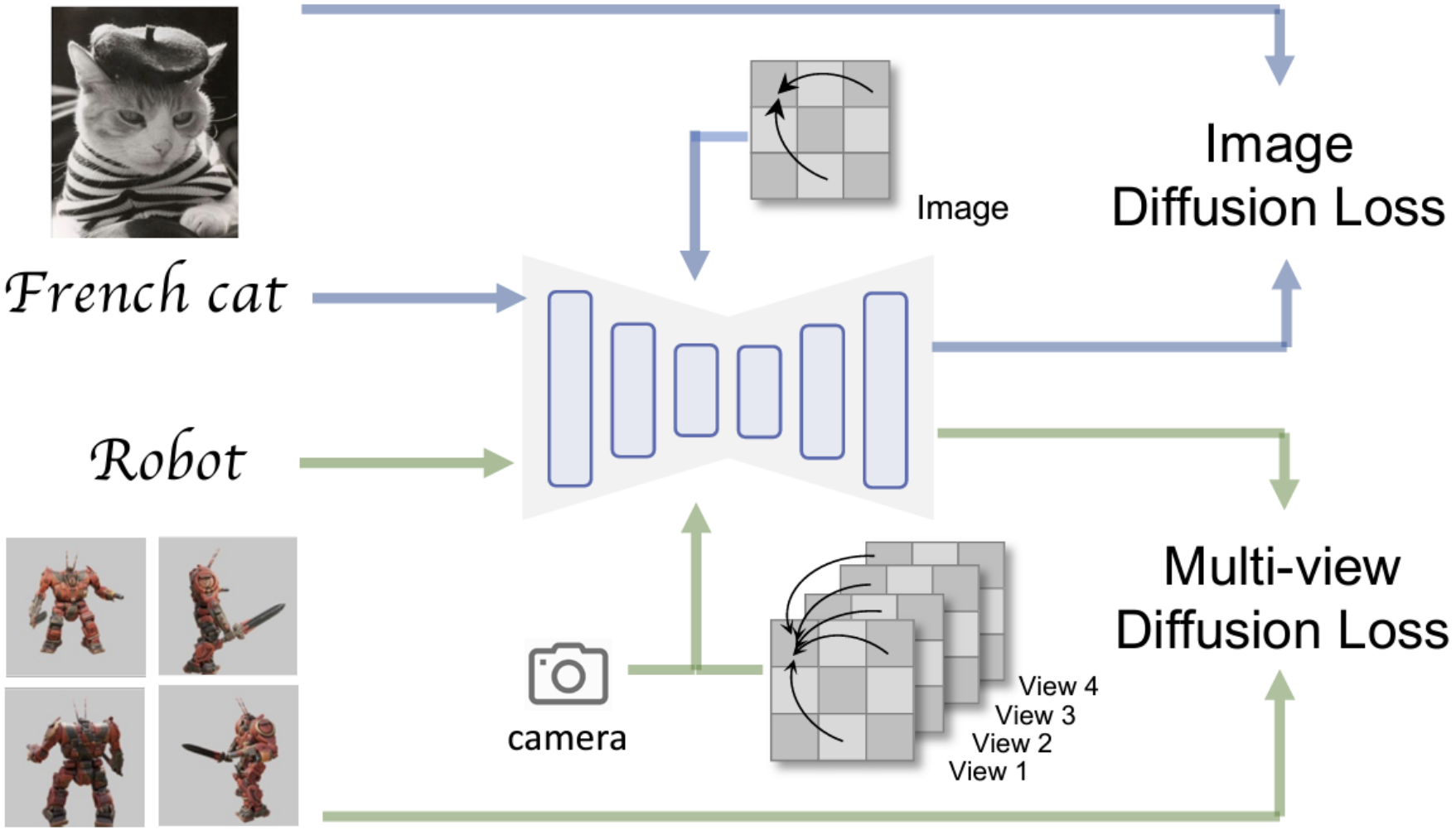
(a) Temporal Attention

MVImageNet Pseudocode

Algorithm 1: Pseudocode for MVDream training

Data: $\mathcal{X}, \mathcal{X}_{mv}$
for $i \leftarrow 1$ *to* $n-1$ **do**
  sample $mode \sim U(0,1)$;
  if $mode \leq 0.7$ **then**
    select a random 3D sample from $\mathcal{X}_{mv}$;
    $\mathbf{x} \leftarrow$ 4 random orthogonal views out of 32 views;
    $\mathbf{c} \leftarrow$ camera extrinsics;
  else
    $\mathbf{x} \leftarrow$ 4 random images from $\mathcal{X}$;
    $\mathbf{c} \leftarrow \varnothing$;
  $y \leftarrow$ text descriptions;
  sample $t \sim U(0, 1000)$;
  $\mathbf{x}_t \leftarrow$ add_noise$(\mathbf{x}, t)$;
  forward_and_backward$(\theta, \mathbf{x}, \mathbf{x}_t, y, \mathbf{c}, t)$

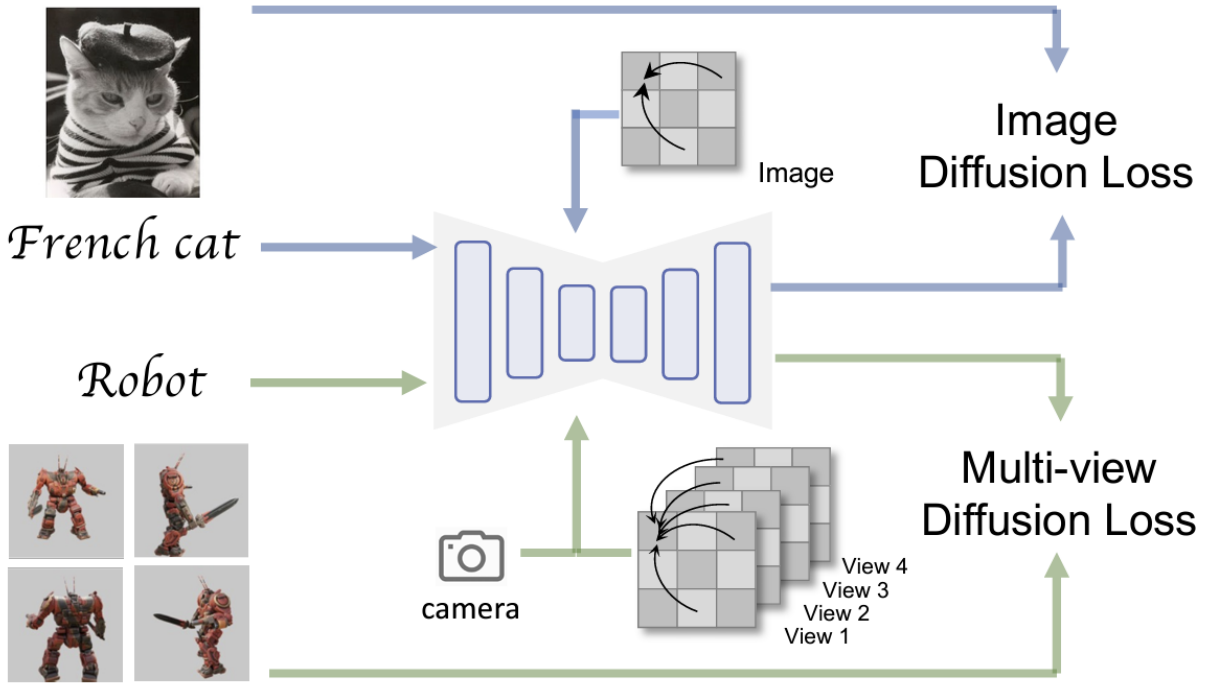
Training


French cat
Robot
Image
Image
Diffusion Loss
camera
View 4
View 3
View 2
View 1
Multi-view
Diffusion Loss

Training Details

- $x_t \in \mathbb{R}^{F \times H \times W \times C}$
- $c \in \mathbb{R}^{F \times 16}$
- $x_0 \in \mathbb{R}^{F \times H \times W \times C}$
- Inflated 3D self-attention
- $B \times F \times H \times W \times C \rightarrow B \times FHW \times C$

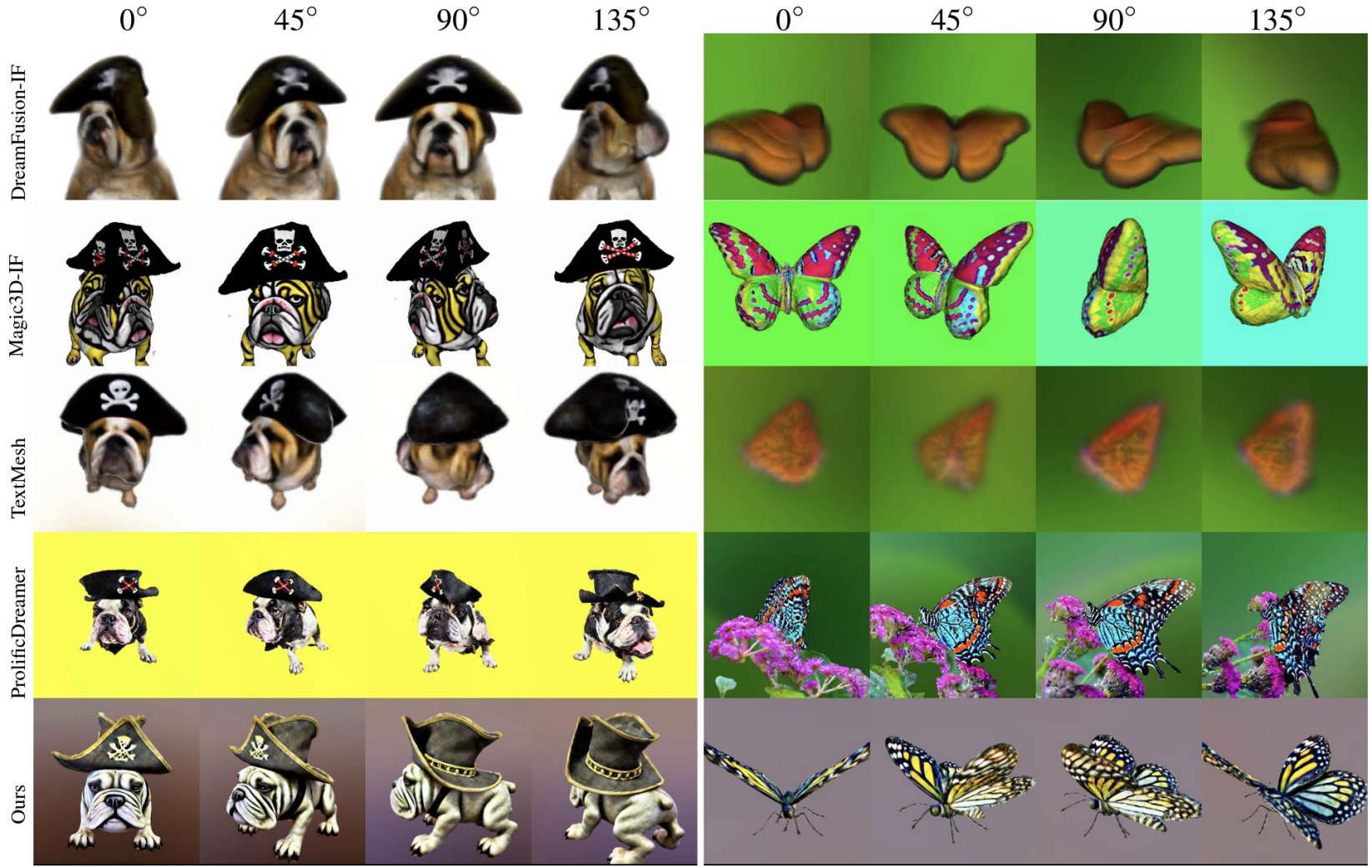
$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$



$$\mathcal{L}_{MV}(\theta, \mathcal{X}, \mathcal{X}_{mv}) = \mathbb{E}_{\mathbf{x},y,\mathbf{c},t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t; y, \mathbf{c}, t)\|_2^2\right]$$

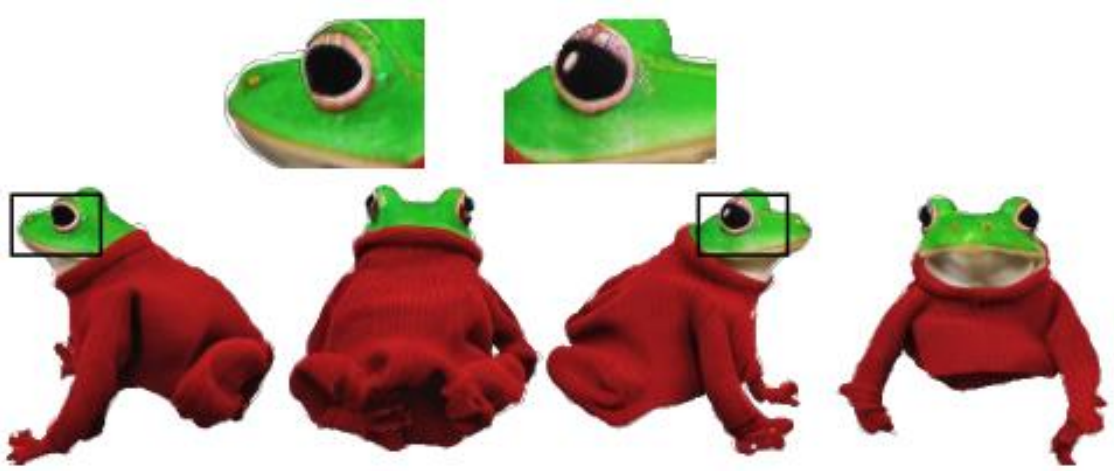
Example Results

View-consistency



3D only vs 2D/3D data for Training

Applied on Multi-view Generation



Tsalicoglou et al., Textmesh: Generation of realistic 3D meshes from text prompts, arXiv'23

Zero123++

A Single Image to Consistent Multi-view Diffusion Base Model
(stacking 6 predefined views)

Code

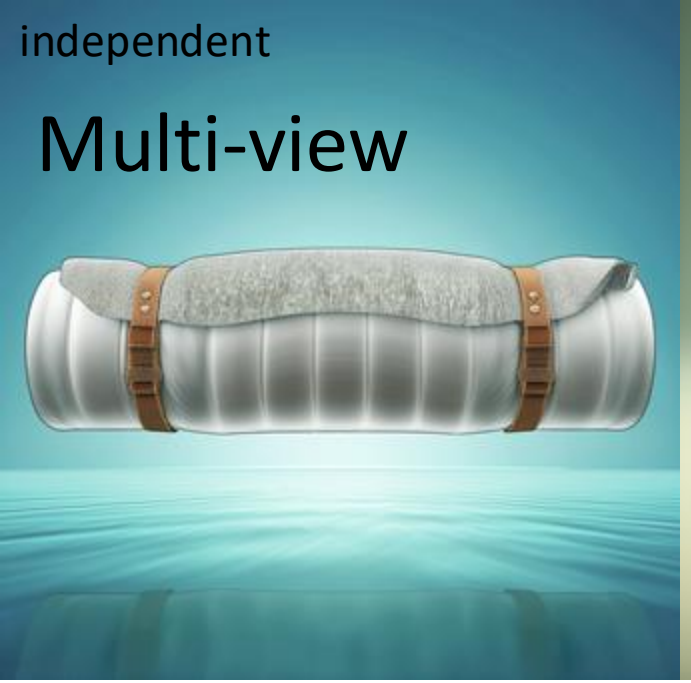
https://github.com/SUDO-AI-3D/zero123plus

Demo

https://github.com/SUDO-AI-3D/zero123plus

independent

Multi-view



grid-view

batch

MultiDiffusion

Fusing Diffusion Paths for Controlled Image Generation
(averaging noisy latents)

Code

https://github.com/omerbt/MultiDiffusion

SyncDiffusion

Coherent Montage via Synchronized Joint Diffusions
(loss-guided denoising)

Code

https://github.com/KAIST-Visual-AI-Group/SyncDiffusion

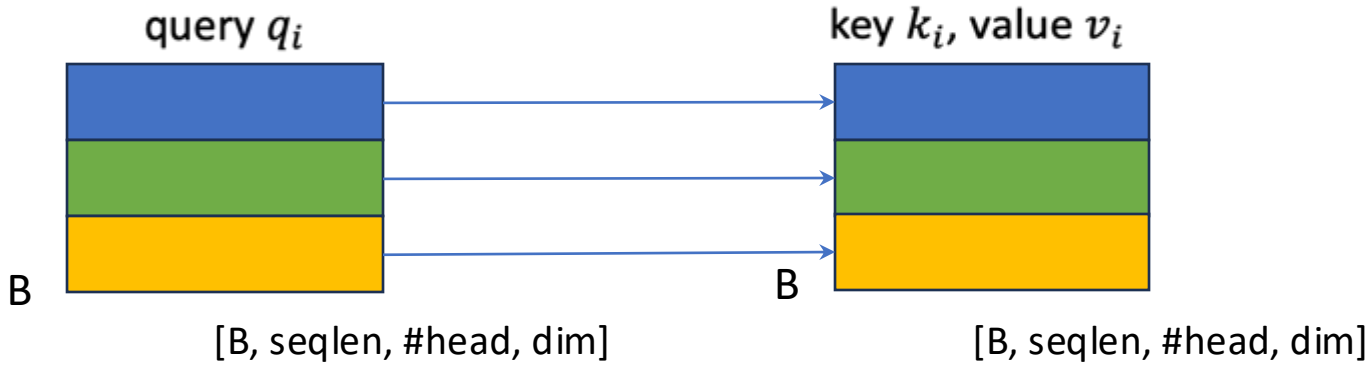
Design Space for Consistency

- Output multiple channels w/ pixel alignment
- Synchronize multiple images w/ over overlapping regions
- Synchronize multiple views via fixed (3D) geometry
- Synchronize multiple views w/ geometric prior
- **Synchronize multiple frames**

Self-attention in UNet

$$Att(q_i, k_i, v_i) = softmax\left(\frac{q_i k_i^{\mathrm{T}}}{\sqrt{d}}\right) v_i$$

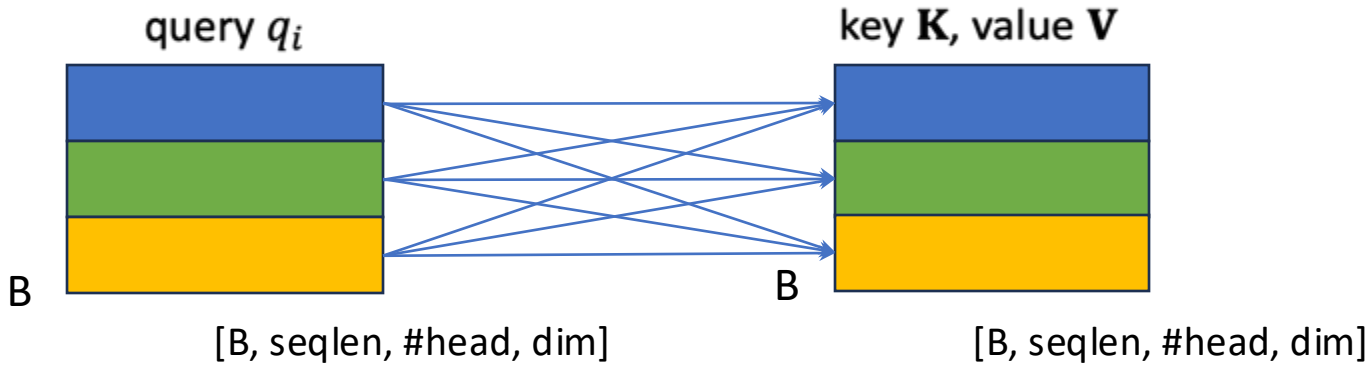
- Independent generation: each sample in the batch attends to only itself.



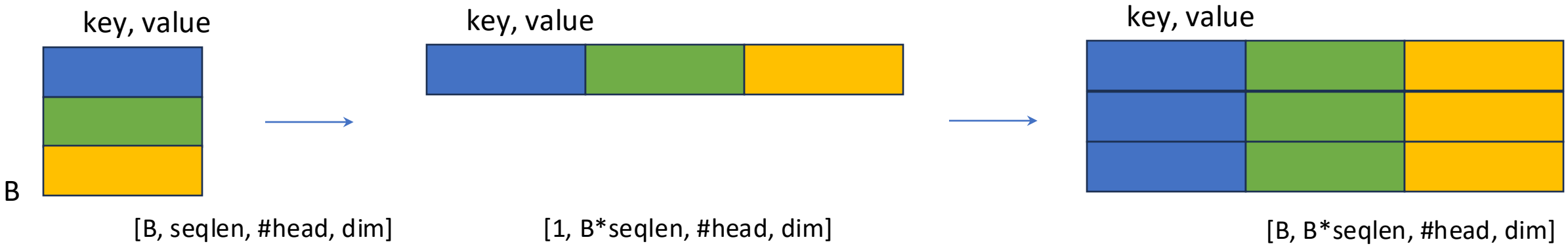
Repurposing Self Attention

$$Att(q_i, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{q_i \mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V},$$
$$\text{where } \mathbf{K} = [k_1, \ldots k_i, \ldots], \mathbf{V} = [v_1, \ldots v_i, \ldots]$$

- Batch generation: each sample attends to every samples.

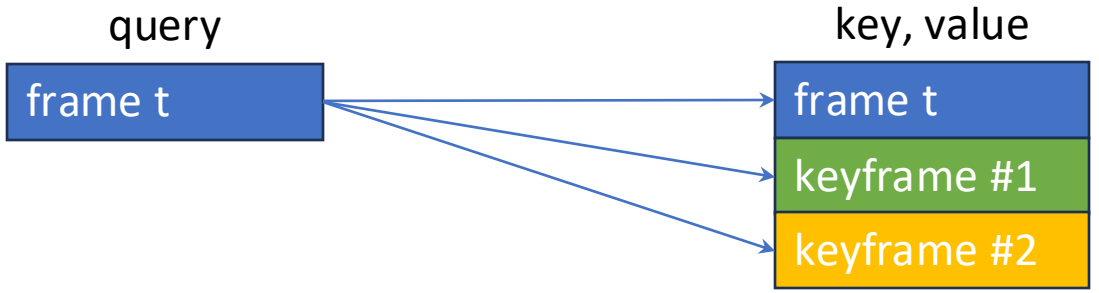


Pseudo Code




```
bs = k.shape[0]
k, v = (t.reshape(1, -1, attn.heads,attn.dim_head) for t in (k, v))
k = k.expand(bs, -1, -1, -1)
v = v.expand(bs, -1, -1, -1)
```

Training-free/zero-shot Video Stylization



- Each frame attends to pre-defined “keyframes”
  - Pix2Video [1]: keyframes = [frame 1, frame t-1]
  - FateZero [2]: keyframes = [middle frame]
  - Text2Video-Zero [3]: keyframes = [frame 1]

[1] Ceylan et al., Pix2Video: Video Editing using Image Diffusion, *ICCV’23*
[2] Qi et al., FateZero: Fusing Attentions for Zero-shot Text-based Video Editing, *ICCV’23*
[3] Khachatryan et al., Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators, *ICCV’23*

Text-guided Video Stylization

a group of chocolate pigs looking for food

Ceylan et al., Pix2Video: Video Editing using Image Diffusion, *ICCV'23*

Another Example

a Swarovski blue crystal swan on the lake

Ceylan et al., Pix2Video: Video Editing using Image Diffusion, *ICCV'23*

Pix2Video

Video Editing using Image Diffusion

Code

https://github.com/duyguceylan/pix2video

FateZero

Fusing Attentions for Zero-shot Text-based Video Editing

Code

https://github.com/ChenyangQiQi/FateZero

Text2Video-Zero

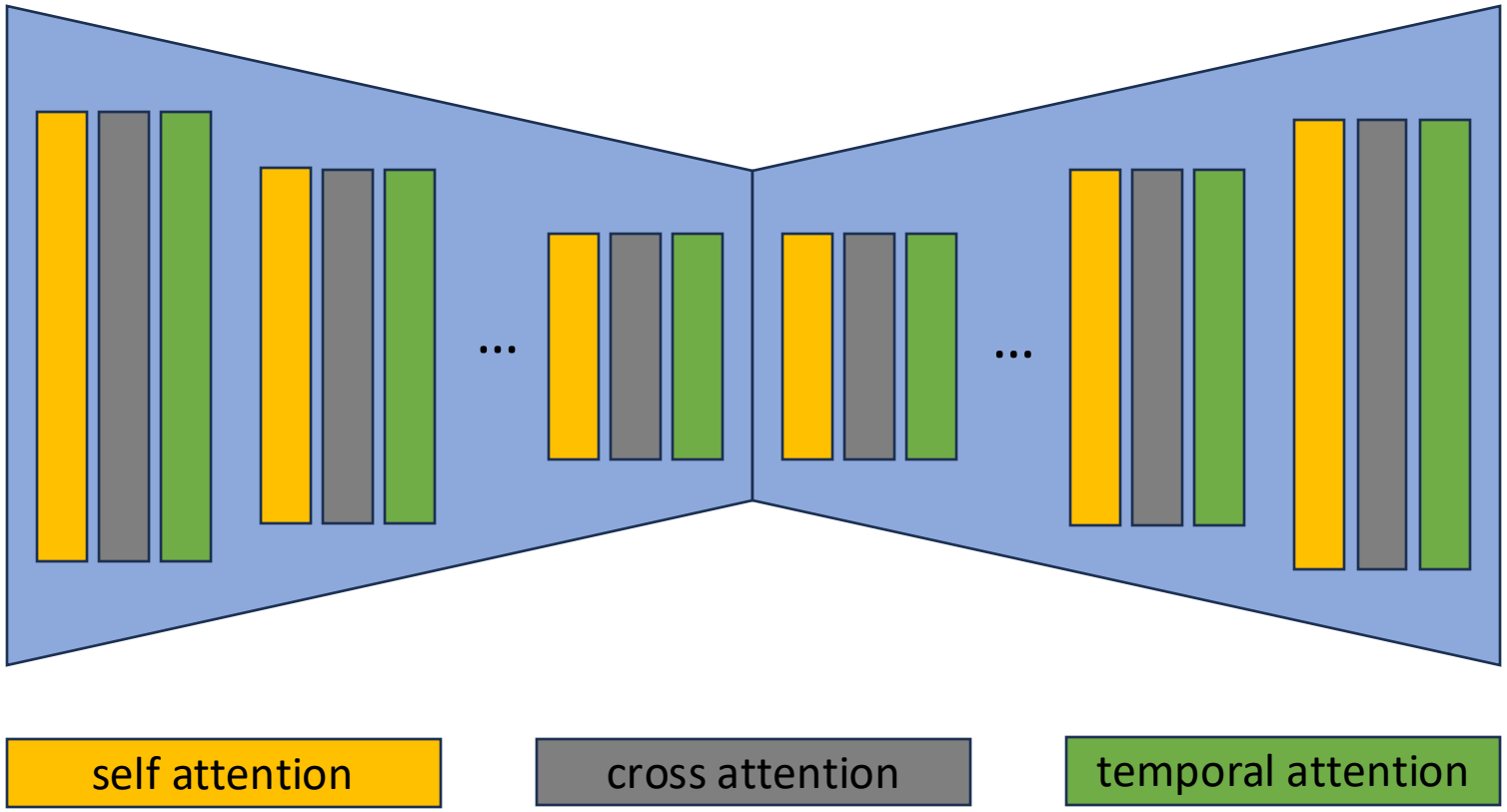
Text-to-Image Diffusion Models are Zero-Shot Video Generators

Code

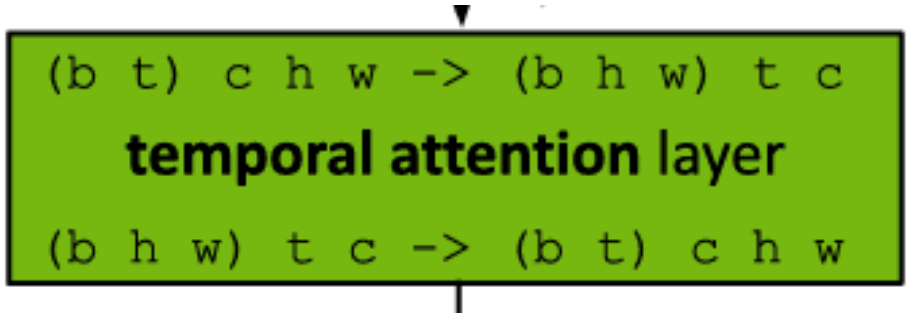
https://github.com/Picsart-AI-Research/Text2Video-Zero

Video Diffusion Model

- Adding a temporal attention layer after spatial cross attention



Inflated Temporal Attention Layer



Blattmann et al., Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models, *CVPR'23*

Blattmann et al., Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets, *arXiv'23*

Guo et al., AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning, *ICLR'24*

Code

https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/transformers/transformer_temporal.py

Presentation Schedule

Introduction to Diffusion Models

Guidance and Conditioning Sampling

Attention

Break

Personalization and Editing

Beyond Single Images

Diffusion Models for 3D Generation