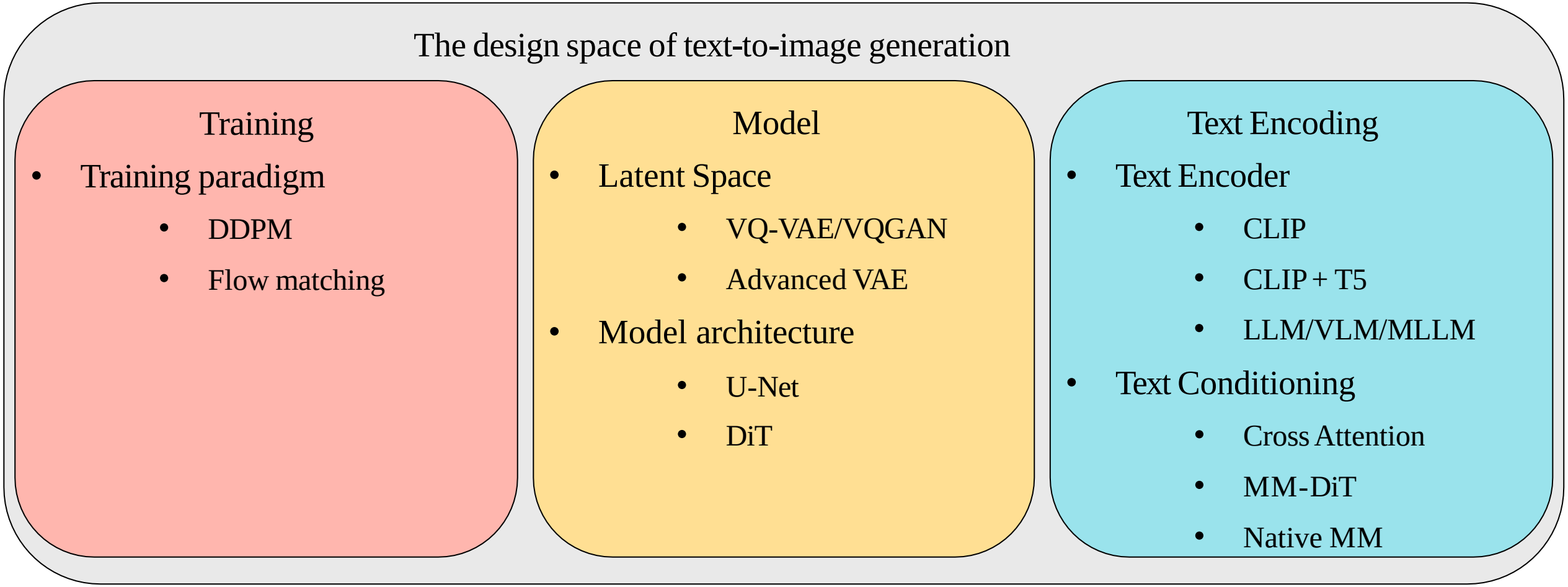
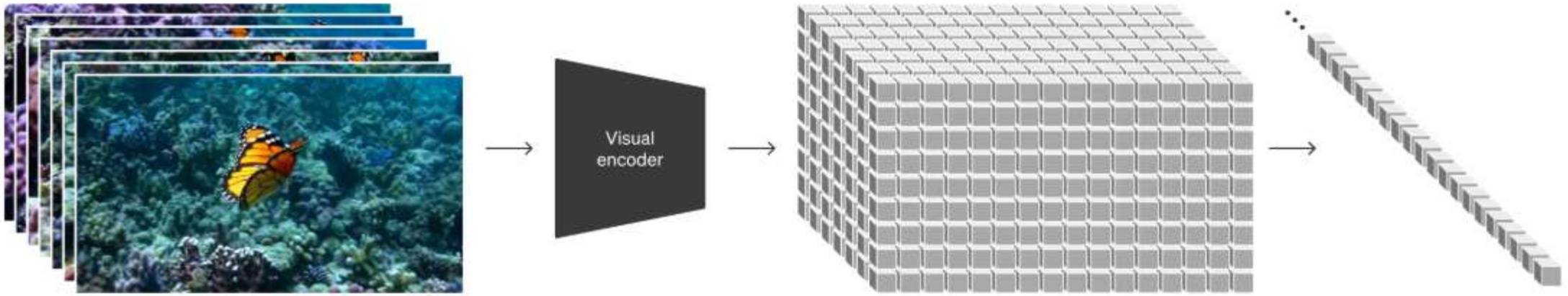
Recap.

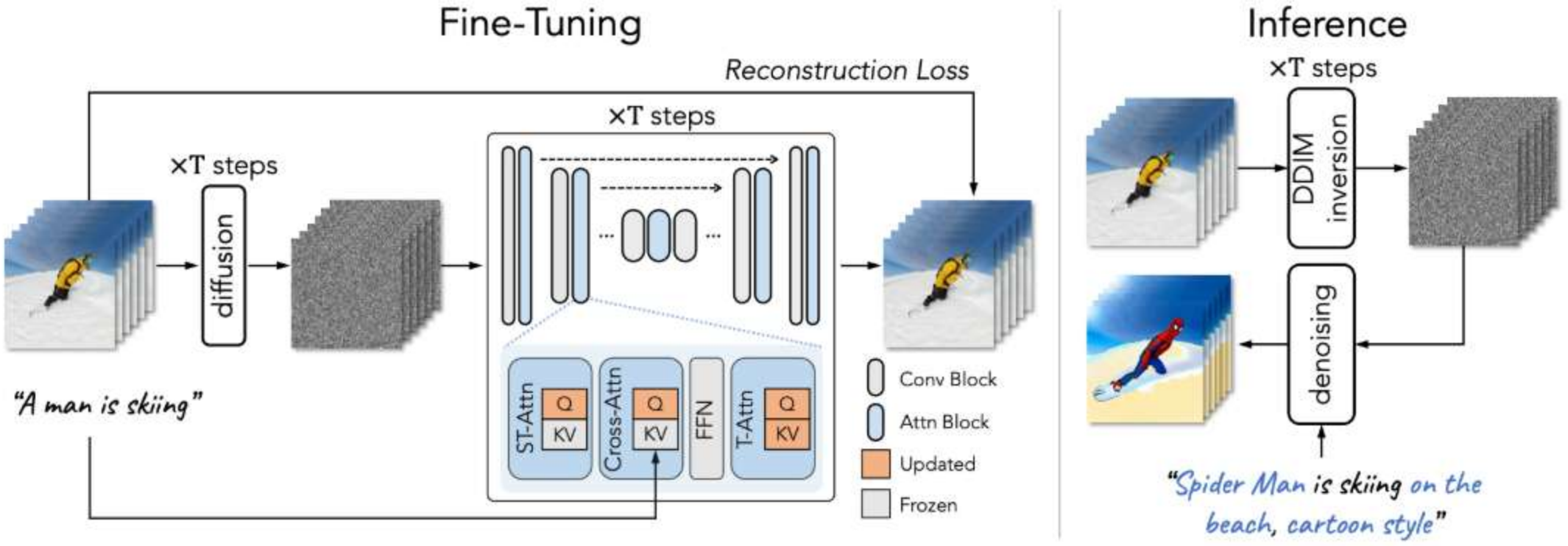

The design space of text-to-image generation
Training
Training paradigm
DDPM
Flow matching
Model
Latent Space
VQ-VAE/VQGAN
Advanced VAE
Model architecture
U-Net
DiT
Text Encoding
Text Encoder
CLIP
CLIP + T5
LLM/VLM/MLLM
Text Conditioning
Cross Attention
MM-DiT
Native MM

Sora: DiT for Video Generation



https://openai.com/research/video-generation-models-as-world-simulators

Adapting Image Models to Generate Videos

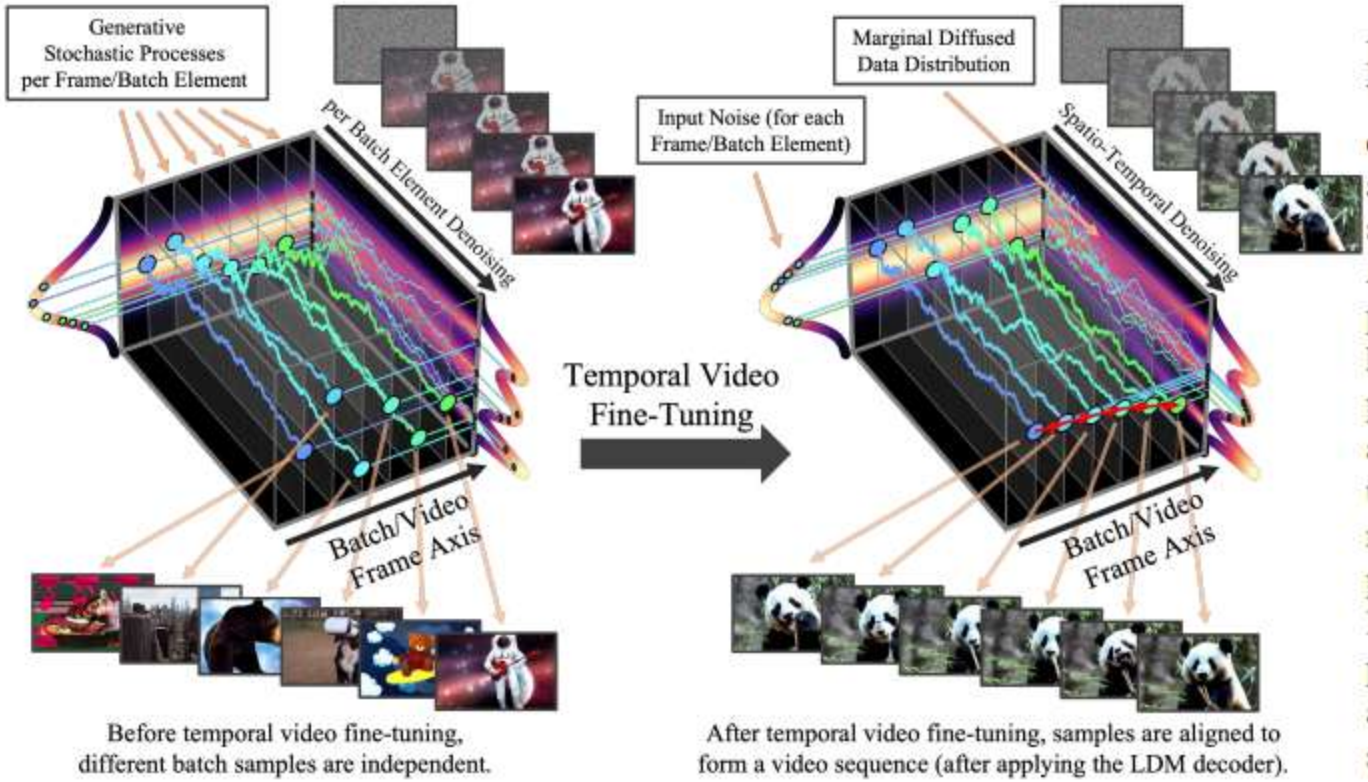


$$\mathbf{Q} = \mathbf{W}^Q \mathbf{z}_{v_i}, \quad \mathbf{K} = \mathbf{W}^K [\mathbf{z}_{v_1}, \mathbf{z}_{v_{i-1}}], \quad \mathbf{V} = \mathbf{W}^V [\mathbf{z}_{v_1}, \mathbf{z}_{v_{i-1}}]$$

$$\mathbf{O} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \cdot \mathbf{V}$$

Tune-A-Video https://arxiv.org/abs/2212.11565

Adapting Image Models to Generate Videos



Before temporal video fine-tuning, different batch samples are independent.

After temporal video fine-tuning, samples are aligned to form a video sequence (after applying the LDM decoder).

Figure 2. **Temporal Video Fine-Tuning.** We turn pre-trained image diffusion models into temporally consistent video generators. Initially, different samples of a batch synthesized by the model are independent. After temporal video fine-tuning, the samples are temporally aligned and form coherent videos. The stochastic generation process before and after fine-tuning is visualised for a diffusion model of a one-dim. toy distribution. For clarity, the figure corresponds to alignment in pixel space. In practice, we perform alignment in LDM's latent space and obtain videos after applying LDM's decoder (see Fig. 3). We also video fine-tune diffusion model upsamplers in pixel or latent space (Sec. 3.4).

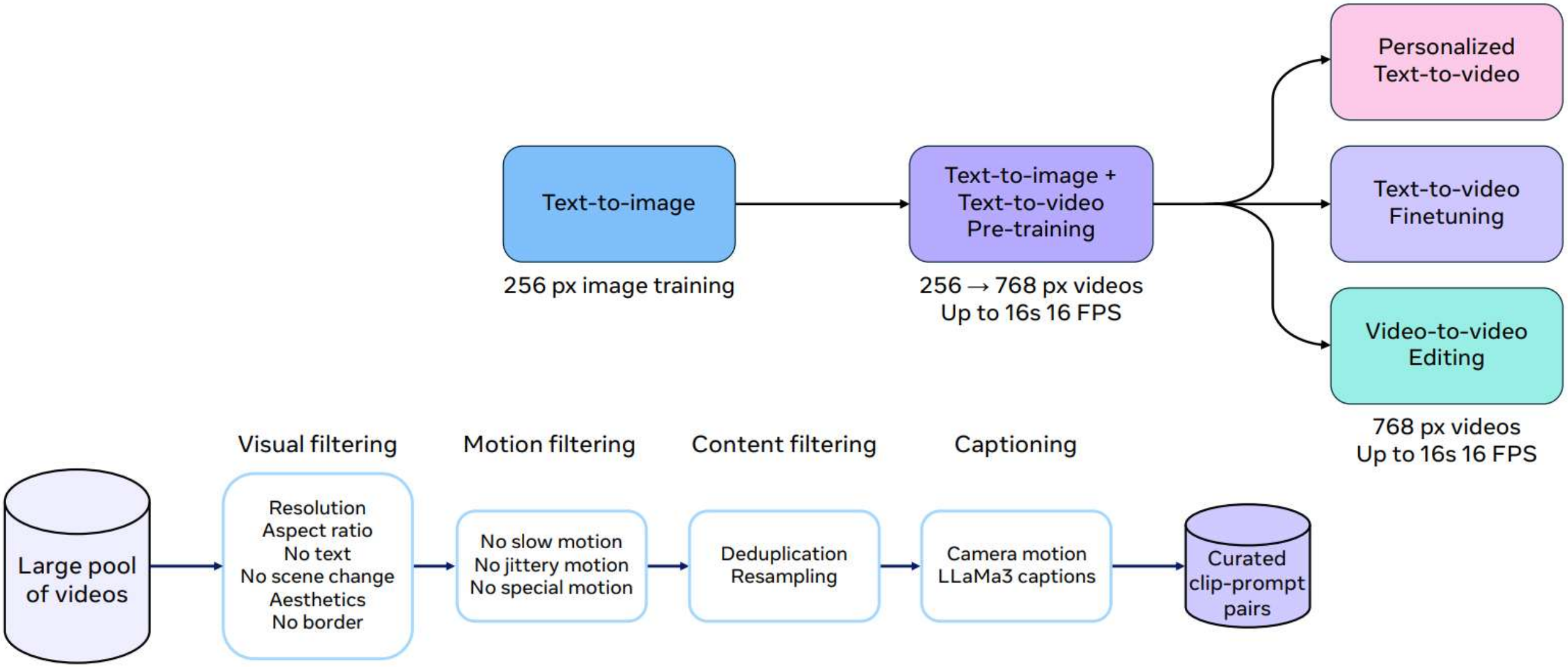
Video LDM https://arxiv.org/abs/2304.08818

Genie: Generative Interactive Environments

“Remarkably, Genie learns not only which parts of an observation are generally controllable, but also infers diverse latent actions that are consistent across the generated environments. Note here how the same latent actions yield similar behaviors across different prompt images. ”

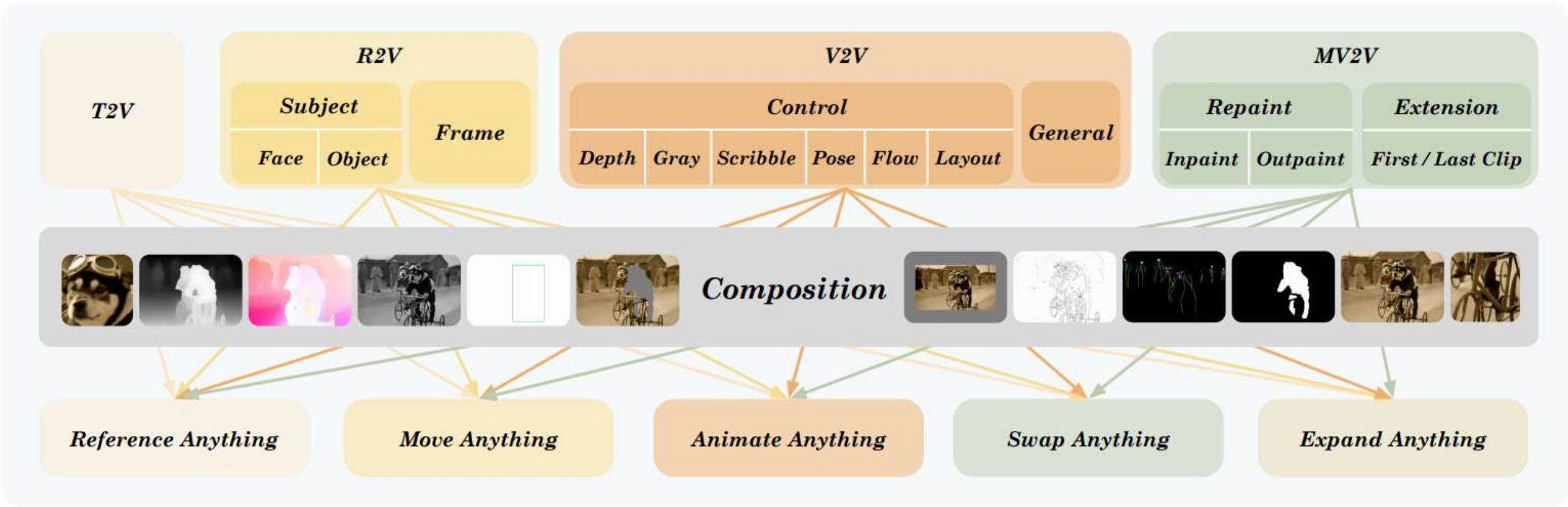
https://sites.google.com/view/genie-2024/home

Movie Gen:
A Cast of Media Foundation Models



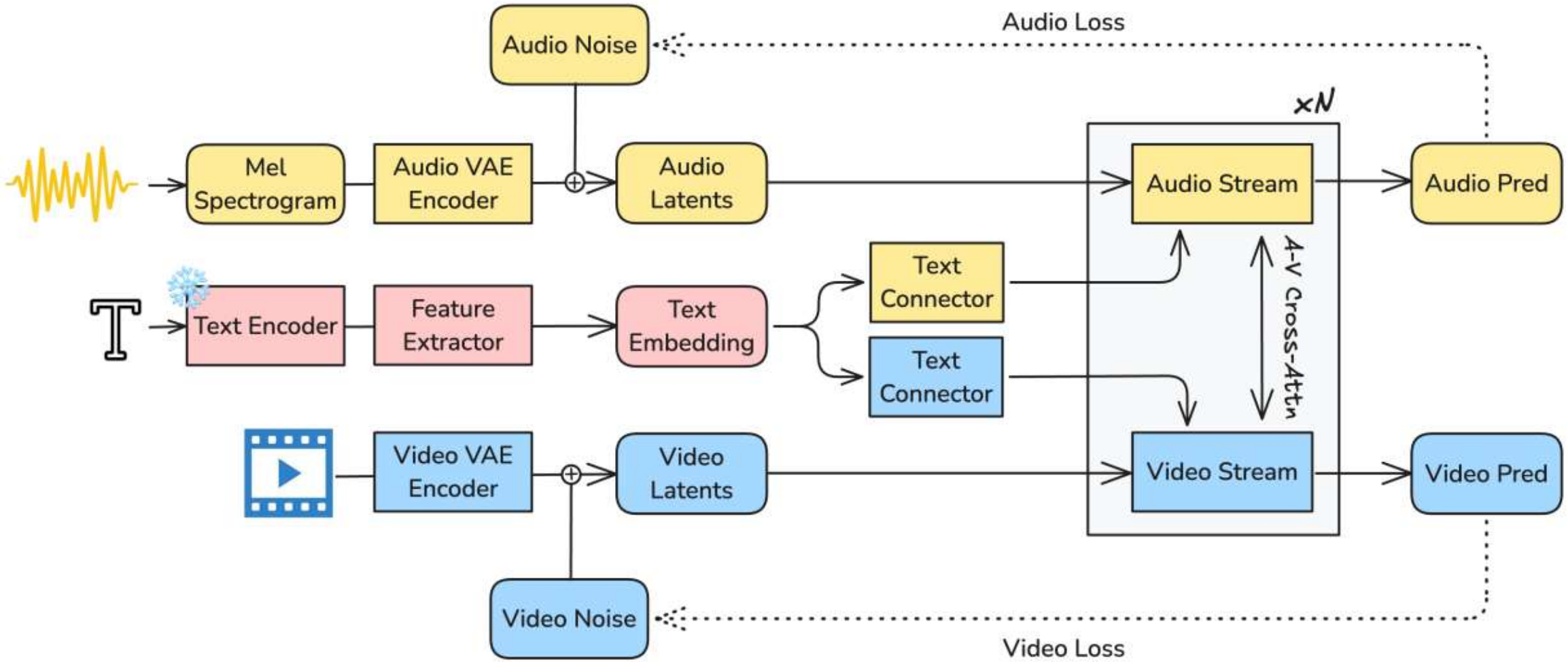
https://ai.meta.com/static-resource/movie-gen-research-paper

VACE



https://arxiv.org/pdf/2503.07598

LTX



https://arxiv.org/pdf/2601.03233

3D Generation

Text-to-3D structure generation

x: generated 3D structures

y: text prompt

Figure credit: Tang, et al. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. ECCV 2024

Image-to-3D structure generation

x: generated 3D structures

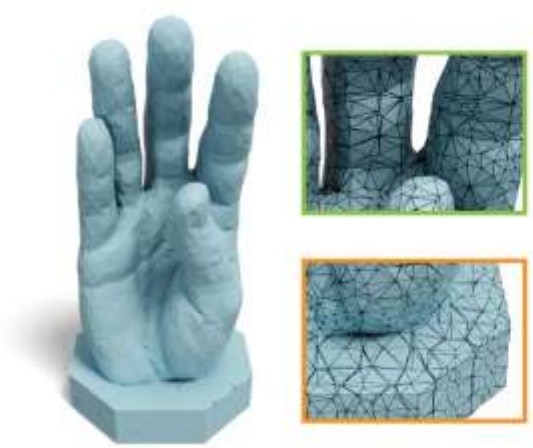
y: image prompt

https://stability.ai/news/stable-point-aware-3d

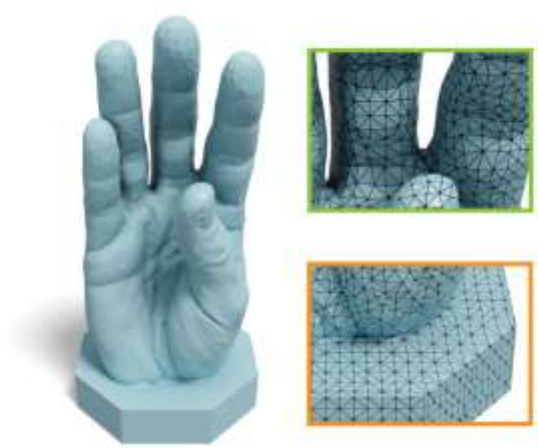
Surface-Based Rendering: Mesh

Isosurfacing w/ gradients

Marching Cubes 15k tris



DMTet 15k tris



FlexiCubes 13k tris

Reference 91k tris

Applications w/ FlexiCubes

3D reconstruction from images

Generative 3D modeling

Animated 3D reconstruction

Tet-mesh physics simulation

Developability

https://arxiv.org/pdf/2308.05371

Surface-Based Rendering: Mesh

DMTet

FlexiCubes

Motorbike

Chair

Car

GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images
Gao, Shen, Wang, Chen, Yin, Li, Litany, Gojcic, Fidler, NeurIPS 2022

https://arxiv.org/pdf/2308.05371

Surface-Based Rendering: Mesh

End-to-end optimization

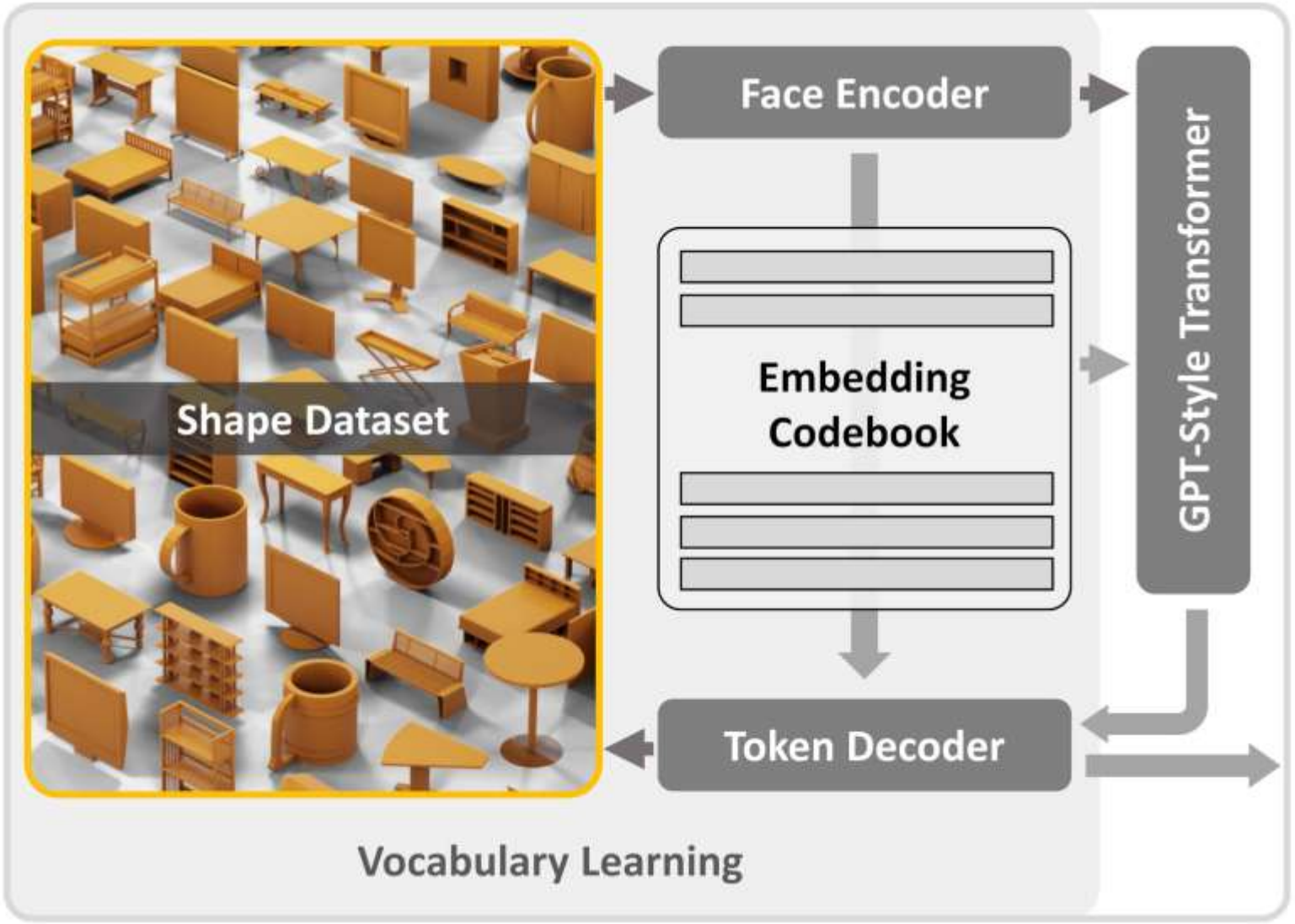
Reference

T-pose optimization

Appearance-Driven Automatic 3D Model Simplification
Hasselgren et. al. Eurographics Symposium on Rendering. 2021

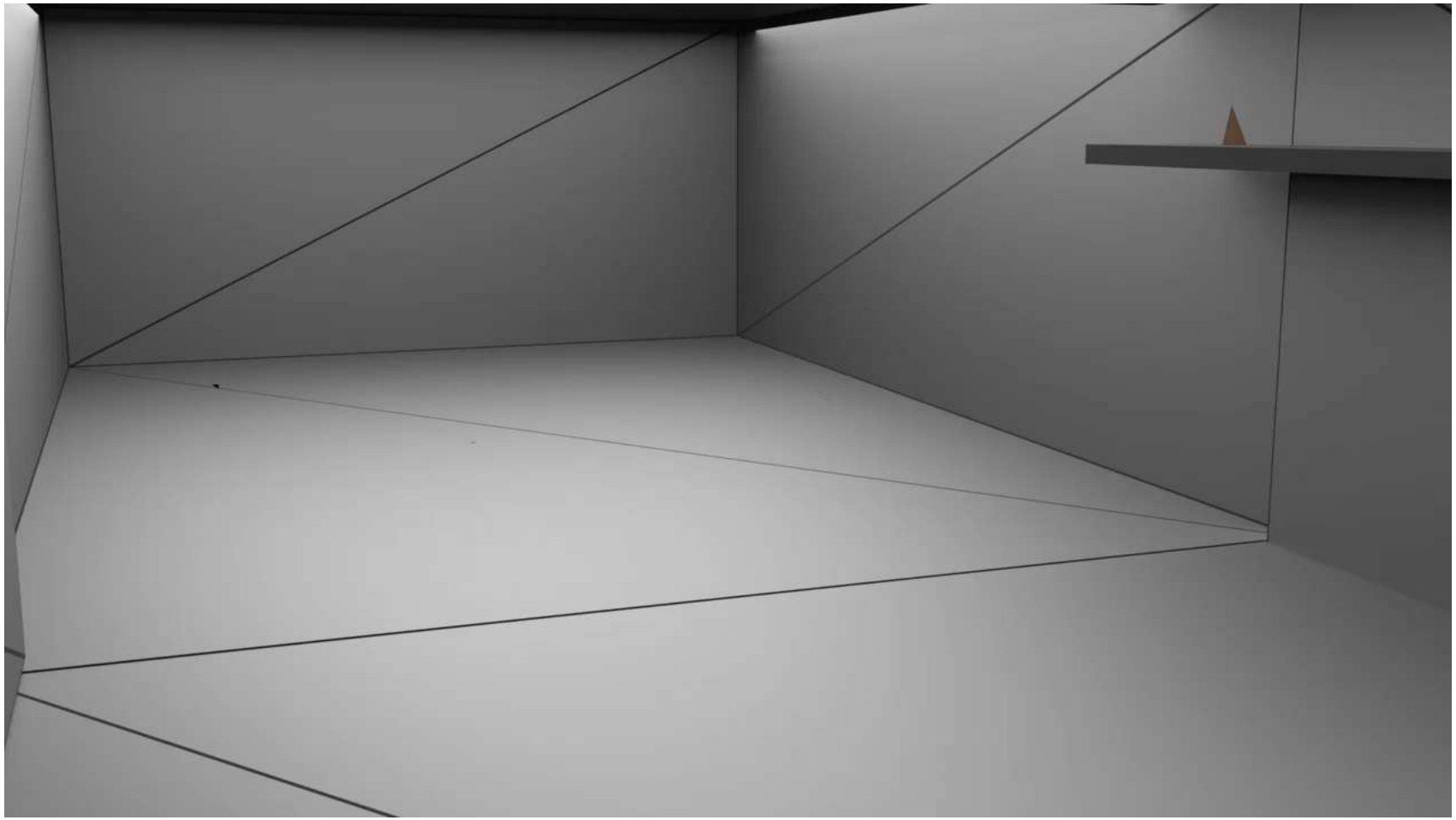
https://arxiv.org/pdf/2308.05371

Surface-Based Rendering: Mesh



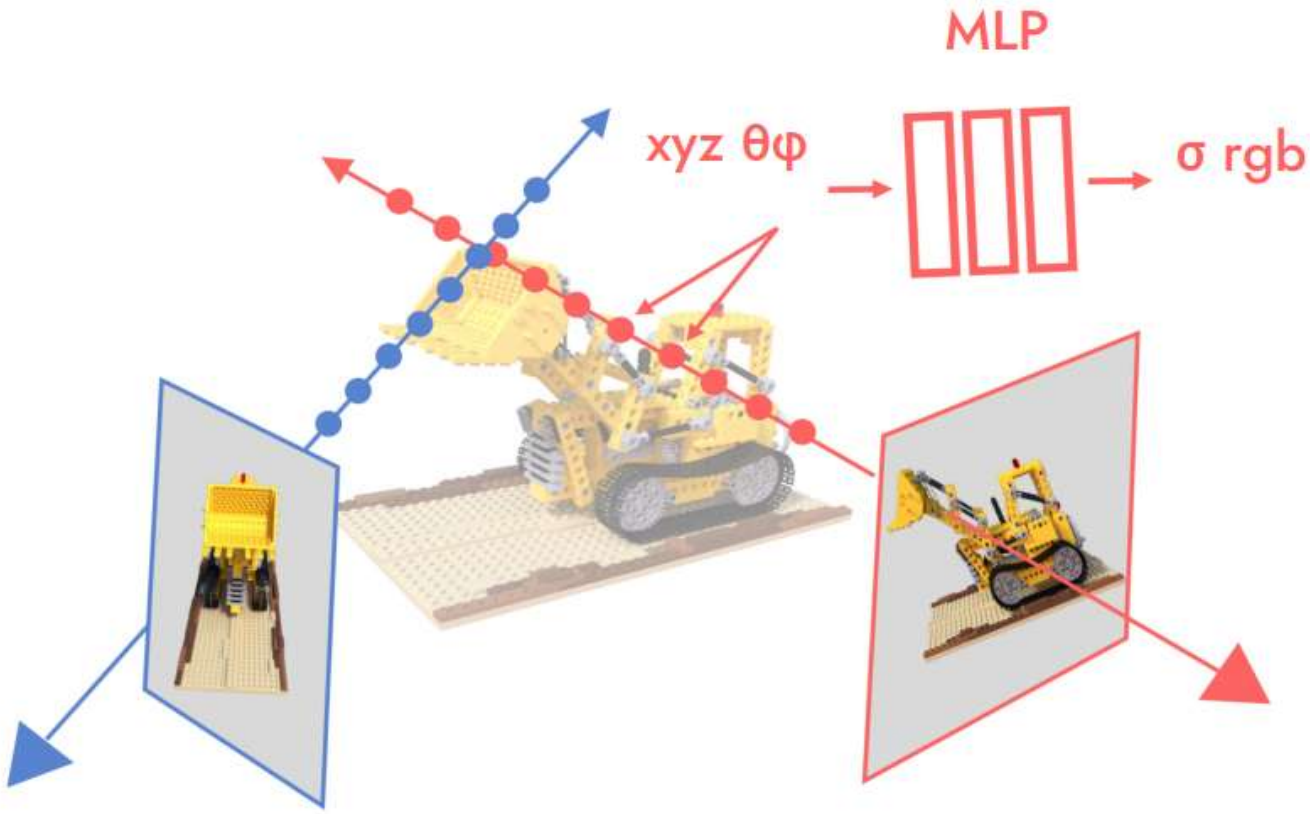
https://arxiv.org/abs/2311.15475

Surface-Based Rendering: Mesh

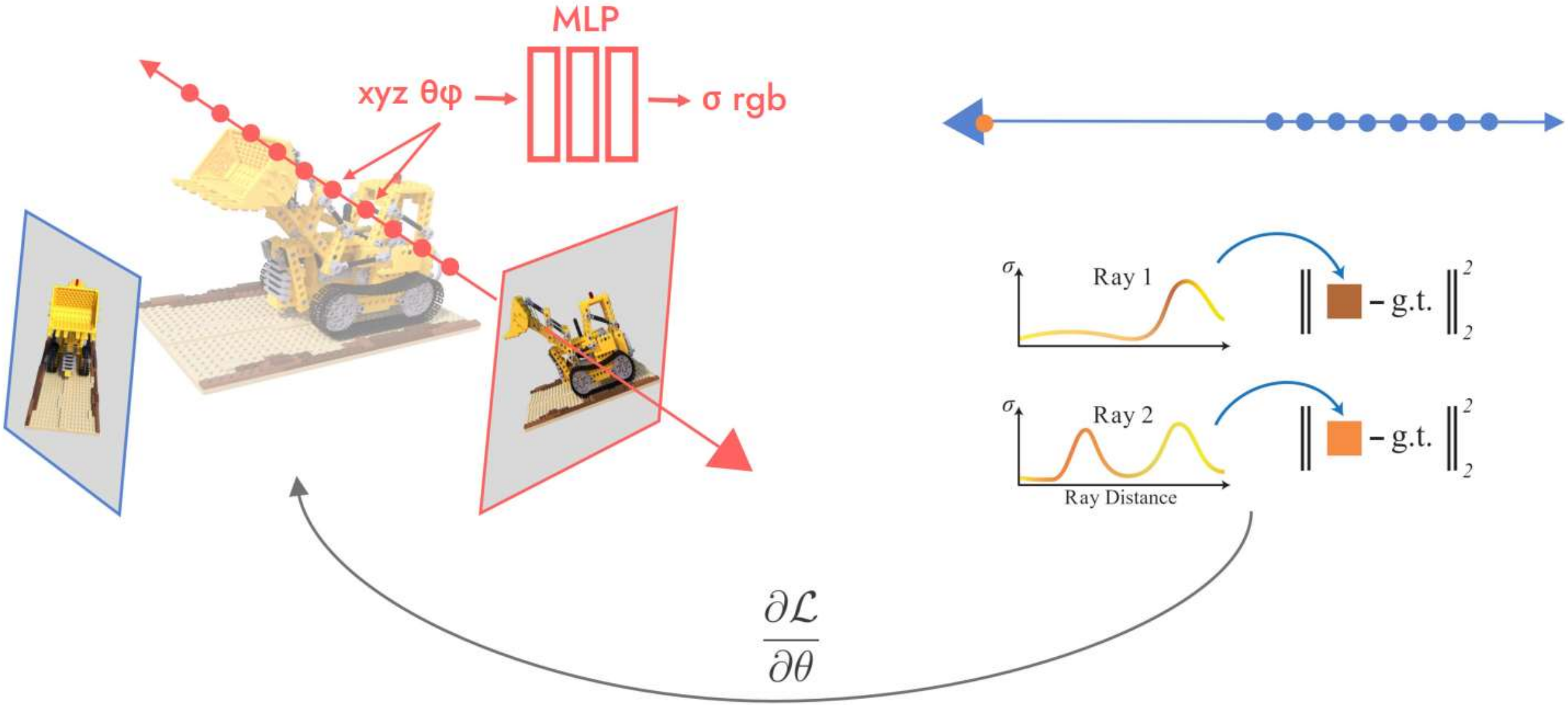


https://arxiv.org/abs/2311.15475

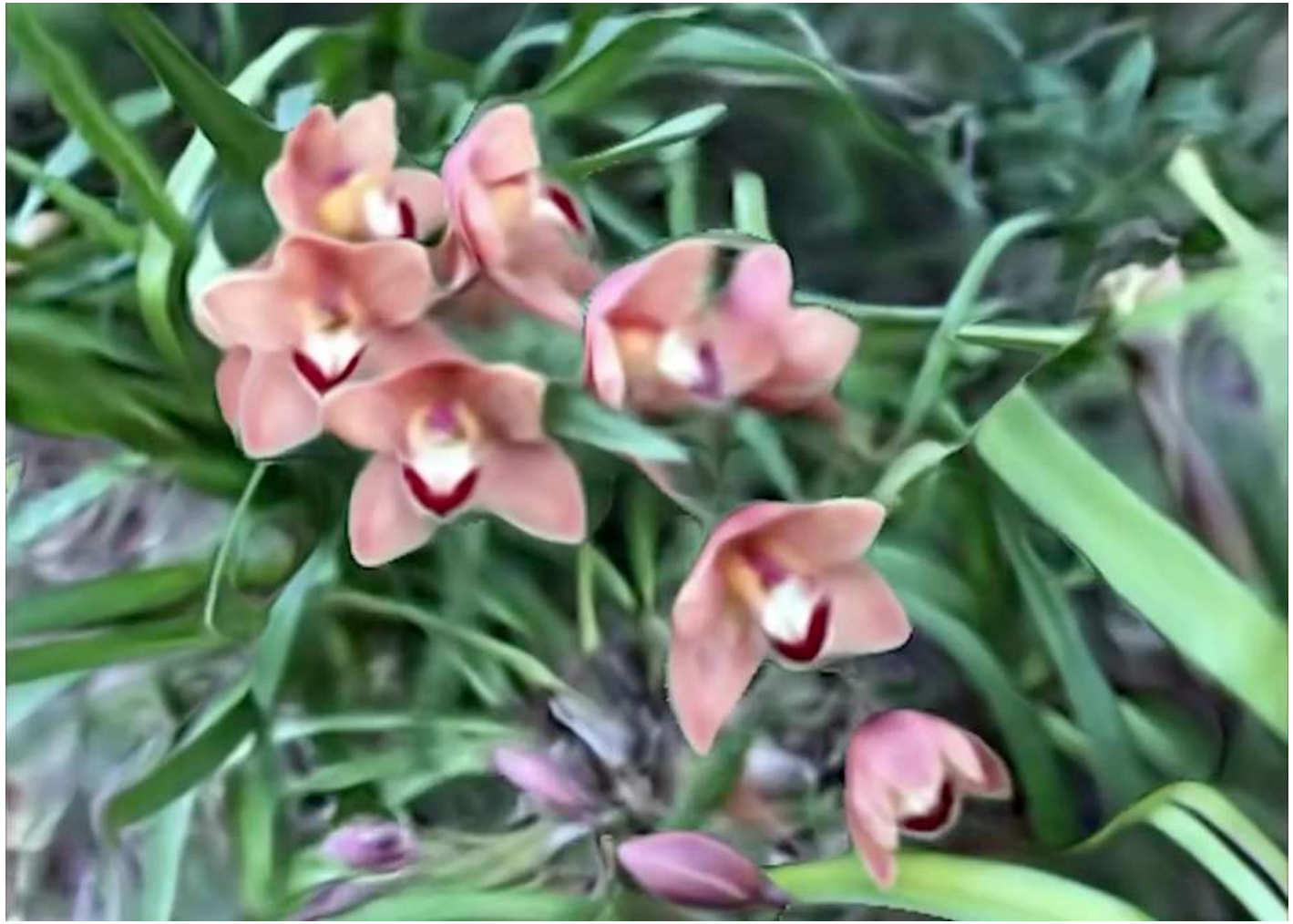
Volume-Based Rendering: NeRF/3DGS



Volume-Based Rendering: NeRF/3DGS



Volume-Based Rendering: NeRF/3DGS

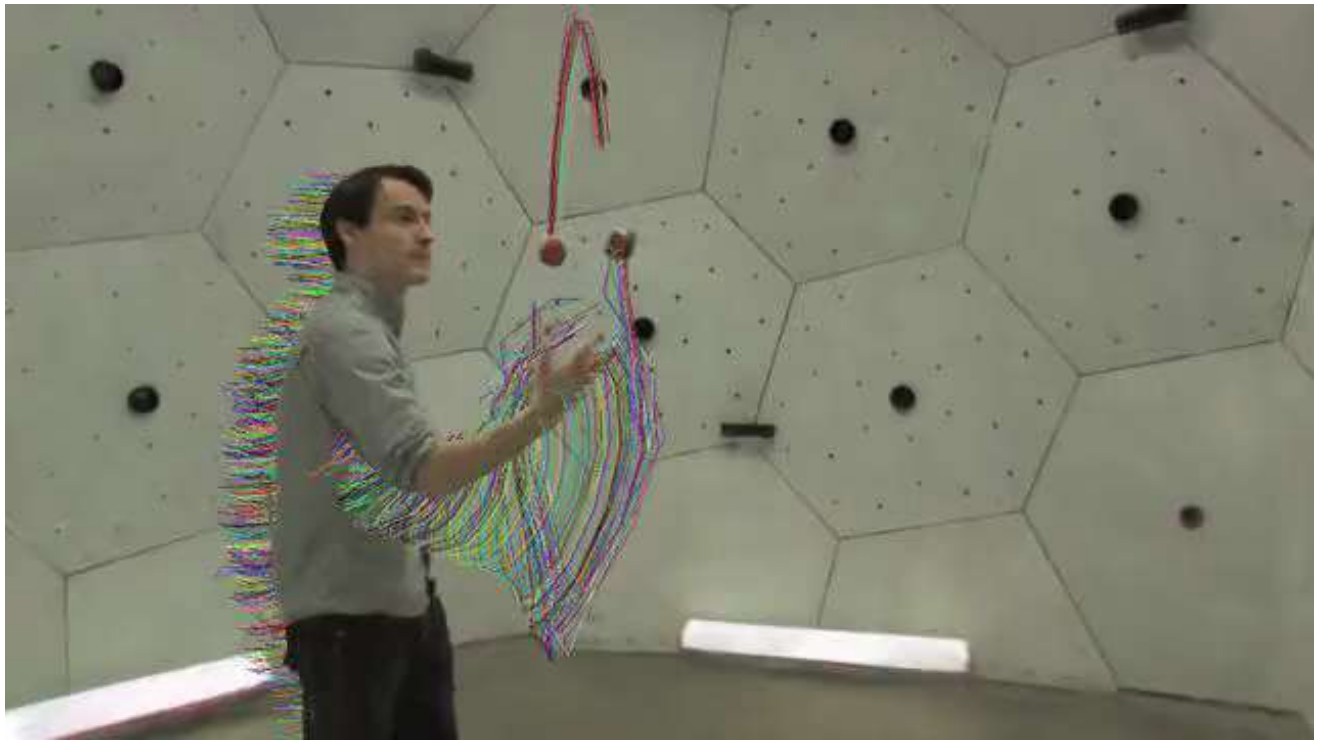


Scenes from NeRF (Mildenhall et al. 2021)

Volume-Based Rendering: NeRF/3DGS

https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

Volume-Based Rendering: NeRF/3DGS

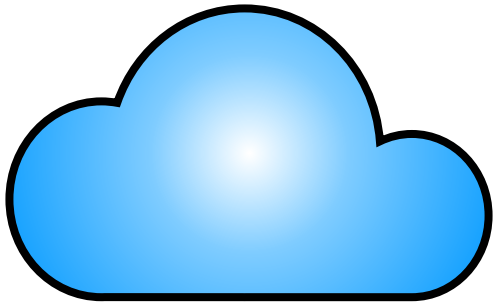


https://dynamic3dgaussians.github.io/

Surface-Based vs. Volume-Based

Aspect	NeRF / 3DGS (Volume-based)	Mesh (Surface-based)
Modeling Paradigm	Perception-driven learning of unknown 3D structures	Explicit modeling of known structures
Input Requirement	Real-world images/videos (multi-view), camera poses	Explicit meshes, CAD models, or known geometry
Topology Prior	Not required; topology-free	Required; predefined or manually designed
Representation Type	Continuous fields (density, radiance)	Discrete surface (vertices, edges, faces)
Learning Objective	Learn geometry and appearance from data	Represent or edit known 3D shapes
Optimization Type	Gradient-based, differentiable rendering	Structure-based, difficult to optimize end-to-end
Expressiveness	Implicit, generalizable to various categories	Explicit, high-fidelity but category-specific
Editability	Difficult (implicit representation)	Easy to edit (explicit geometry and topology)
Preferred Use Case	AI/CV: 3D perception, reconstruction, novel view synthesis	CG/CAD: design, simulation, character animation

Volume Rendering



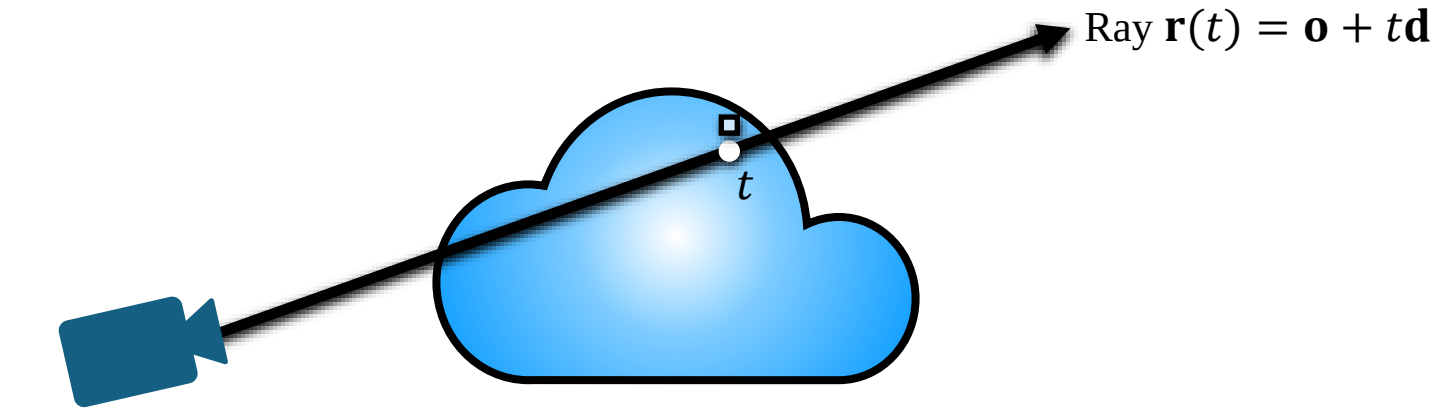
Scene is a cloud of tiny colored particles

Max and Chen 2010, Local and Global Illumination in the Volume Rendering Integral

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Volume Rendering

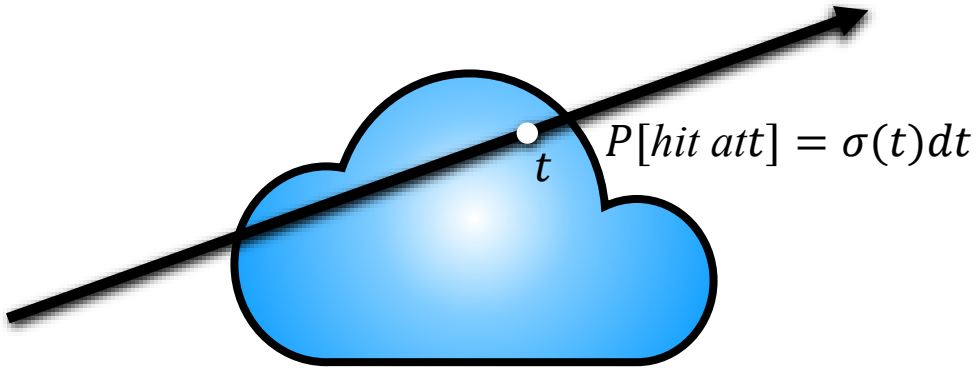


If a ray traveling through the scene hits a particle at t, we return its color $\mathbf{c}(t)$

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Volume Rendering

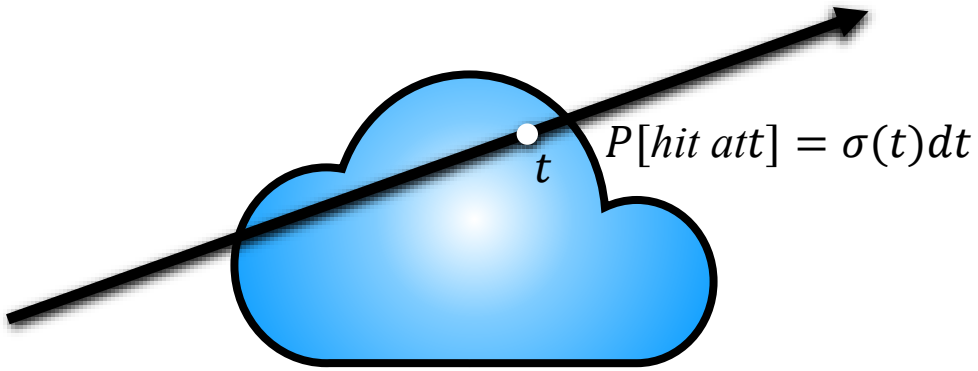


Probability that ray stops in a small interval around t is $\sigma(t)dt$.
σ is known as the Volume Density.

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Volume Rendering



Probability that ray stops in a small interval around t is $\sigma(t)dt$.
σ is known as the Volume Density.

Our field thus has the function signature:

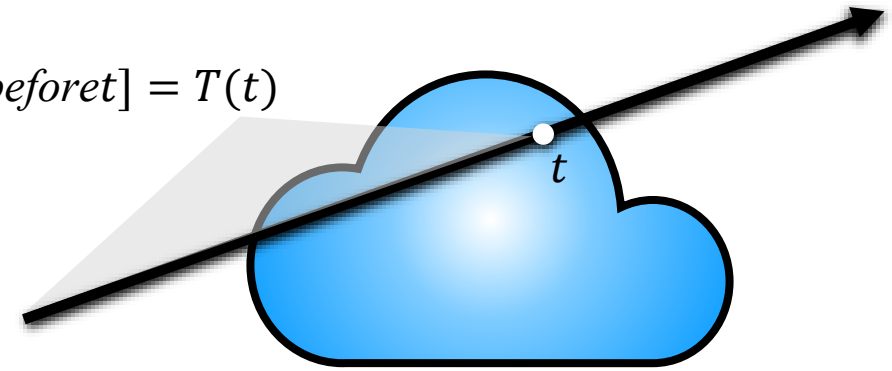
$$\Phi: \mathbb{R}^3 \rightarrow \mathbb{R}^3 \times \mathbb{R}^+; \Phi(\mathbf{x}) = (\sigma, \mathbf{c})$$

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Volume Rendering

$P[\textit{no hits before } t] = T(t)$

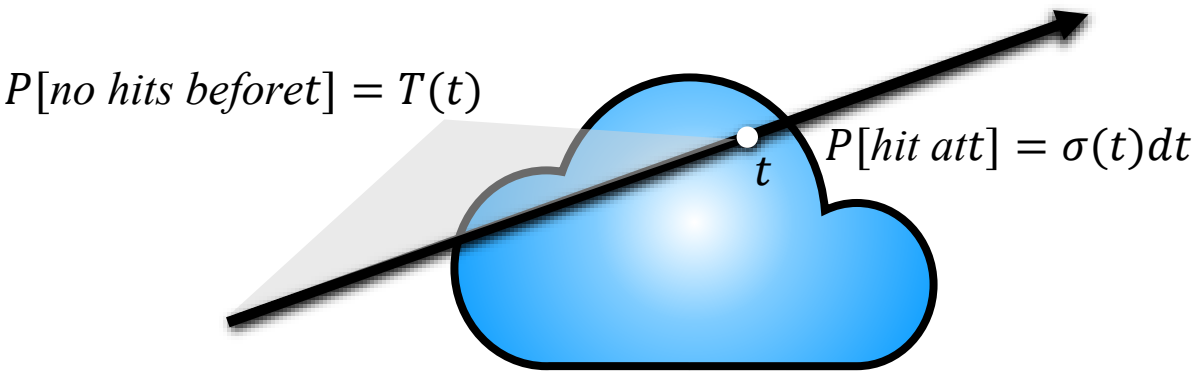


To determine if t is the first hit, need to know $T(t)$:
probability that the ray didn't hit any particles earlier.
$T(t)$ is called Transmittance.

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Volume Rendering



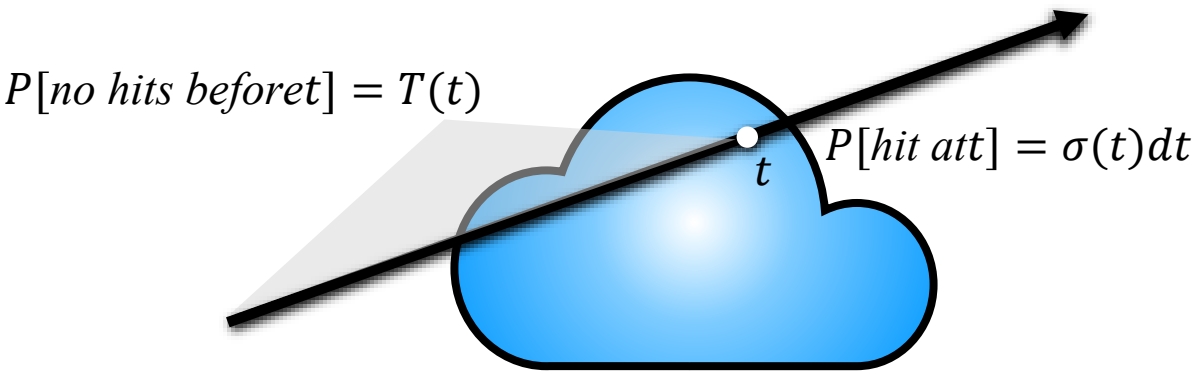
σ and T are related via

$$P[no\ hit\ before t + dt] = P[no\ hit\ before t] \times P[no\ hit\ att]$$

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Volume Rendering



σ and T are related via

$$P[T(t+dt)] = P[T(t)] \times P[1-\sigma(t)dt]$$

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Volume Rendering

No hits before t is equal to integral over density up until t.

$$T(t) = \exp\left(-\int_0^t \sigma(a)da\right)$$

$1 - T(t)$ can be seen as cumulative distribution function of probability that ray hits something before reaching t.

Then $T'(t) = T(t)\sigma(t)$ is probability that ray stops *exactly* at t.

Neural Radiance Fields, Mildenhall et al. 2021

Volume Rendering

Then $T'(t) = T(t)\sigma(t)$ is probability that ray stops *exactly* at t.

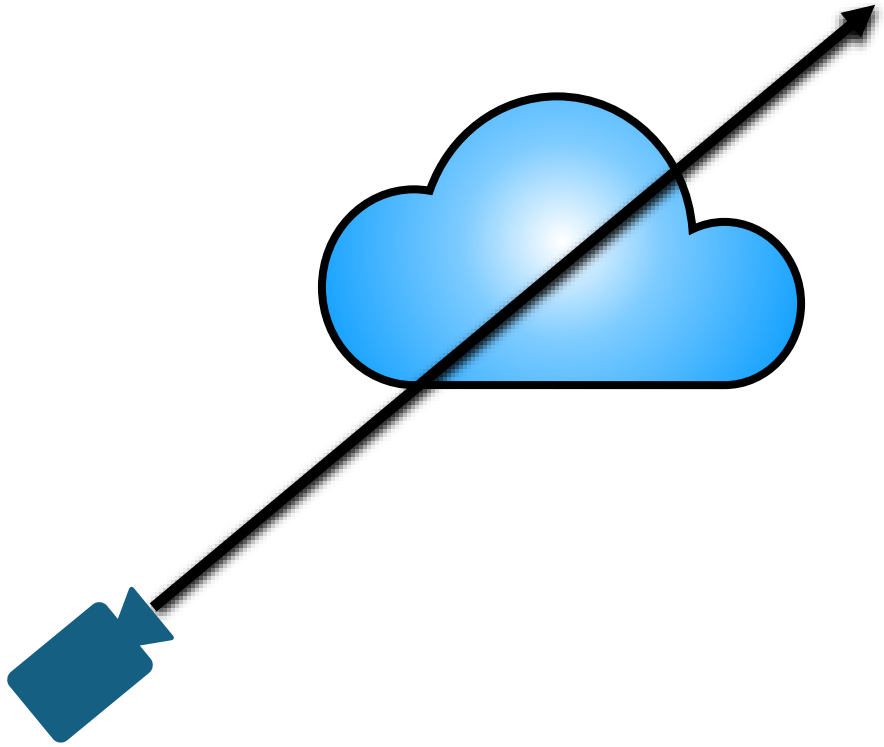
So the expected color returned by the ray will be

$$\int_{t_0}^{t_1} T(t)\sigma(t)\mathbf{c}(t)dt$$

Note the nested integral!

Neural Radiance Fields, Mildenhall et al. 2021

Approximating the nested integral

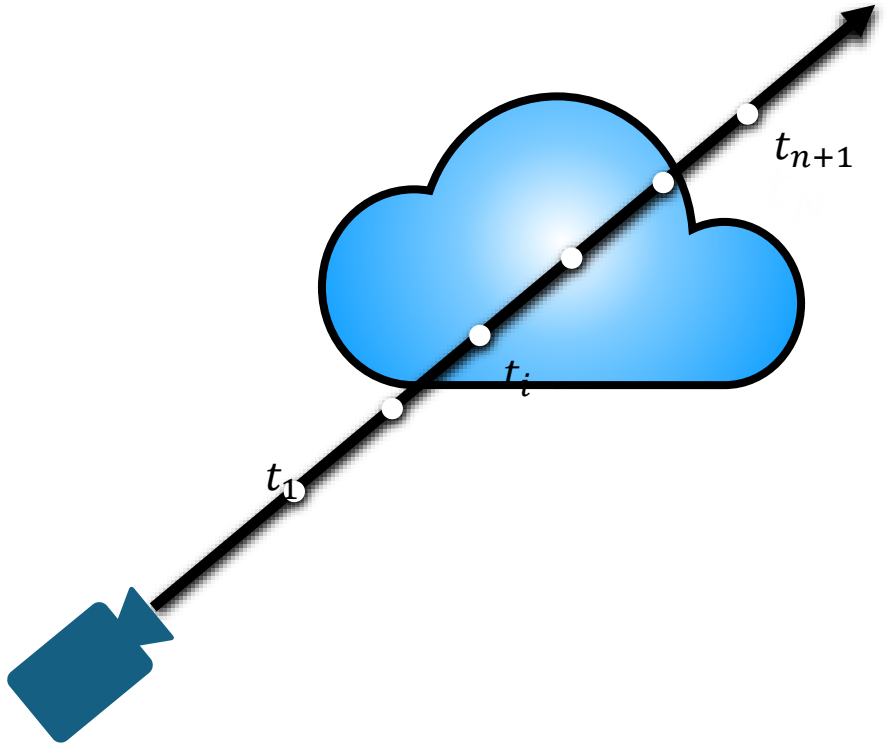


We use quadrature to approximate the nested integral,

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Approximating the nested integral

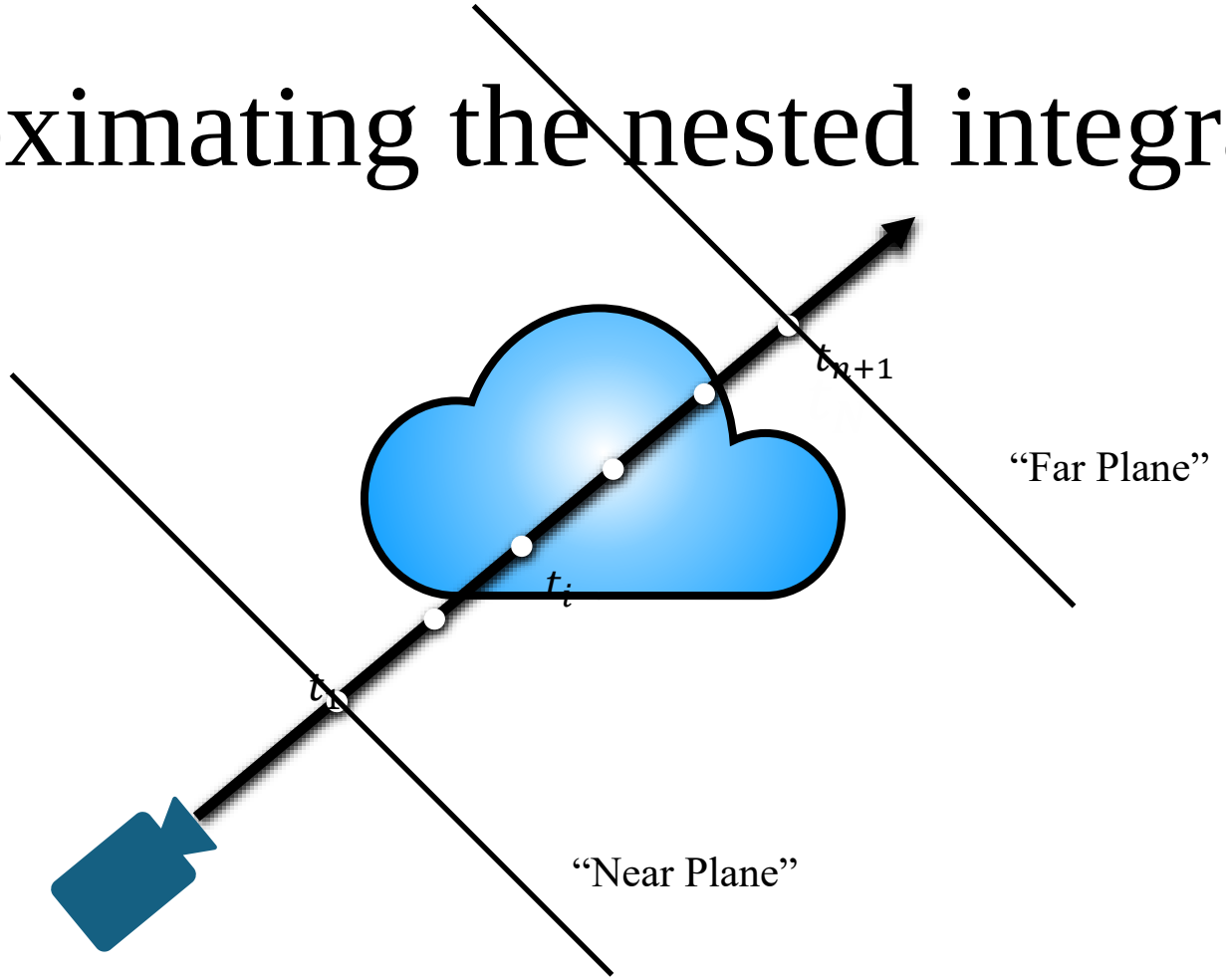


We use quadrature to approximate the nested integral,
splitting the ray up into n segments with endpoints $\{t_1, t_2, \ldots, t_{n+1}\}$

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

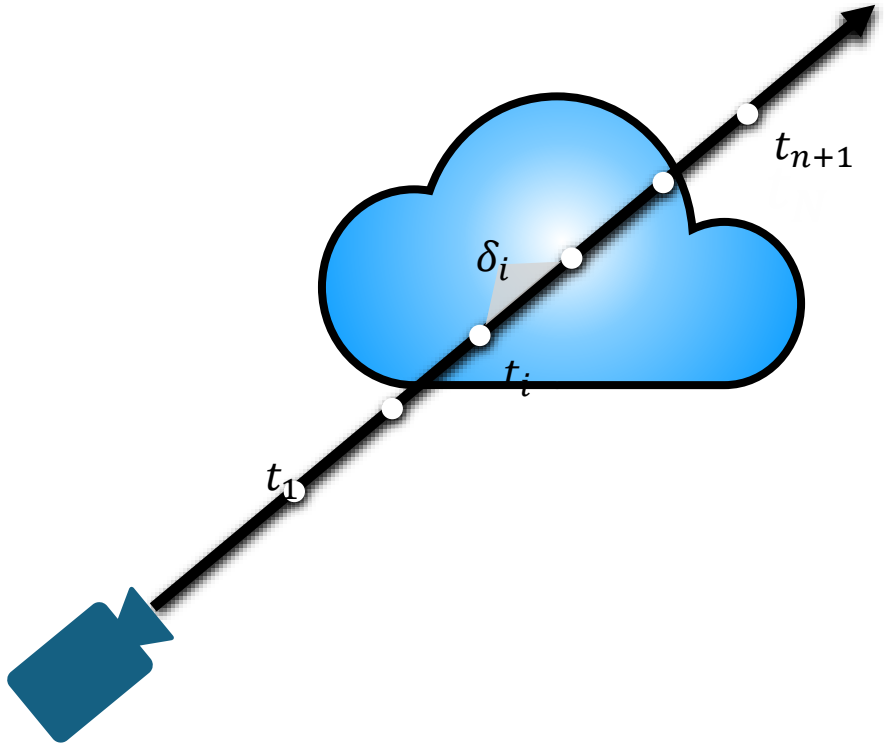
Approximating the nested integral



We use quadrature to approximate the nested integral,
splitting the ray up into n segments with endpoints $\{t_1, t_2, \ldots, t_{n+1}\}$

Slide credit: Ben Mildenhall

Approximating the nested integral

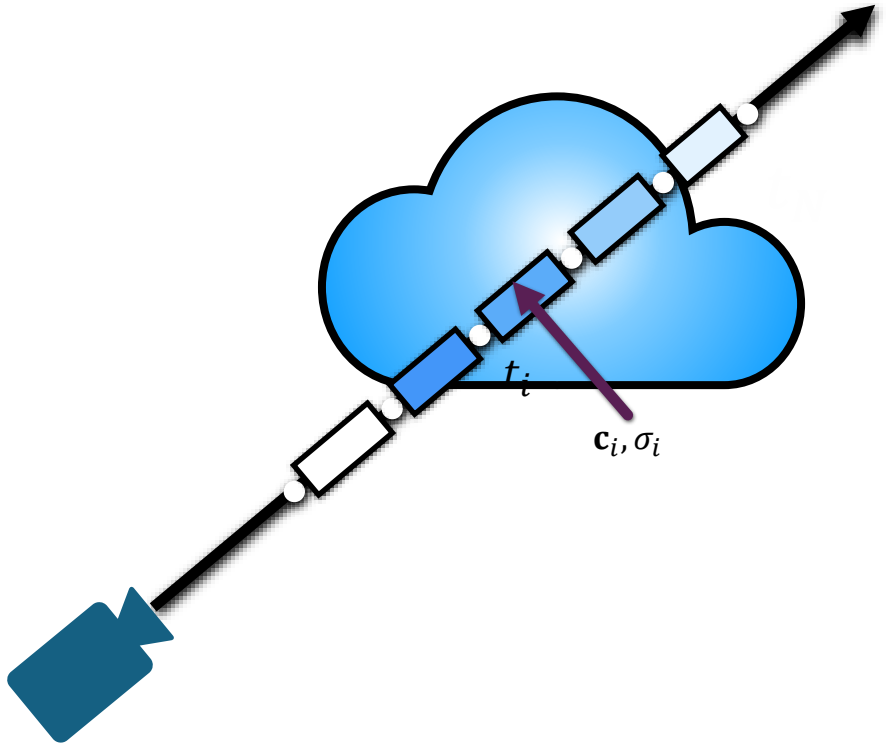


We use quadrature to approximate the nested integral,
splitting the ray up into n segments with endpoints $\{t_1, t_2, \ldots, t_{n+1}\}$
with lengths $\delta_i = t_{i+1} - t_i$

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Approximating the nested integral



We assume volume density and color are roughly constant within each interval

Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

Summary: volume rendering integral estimate

Rendering model for ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$:

$$\mathbf{c} \approx \sum_{i=1}^{n} T_i \alpha_i \mathbf{c}_i$$

colors

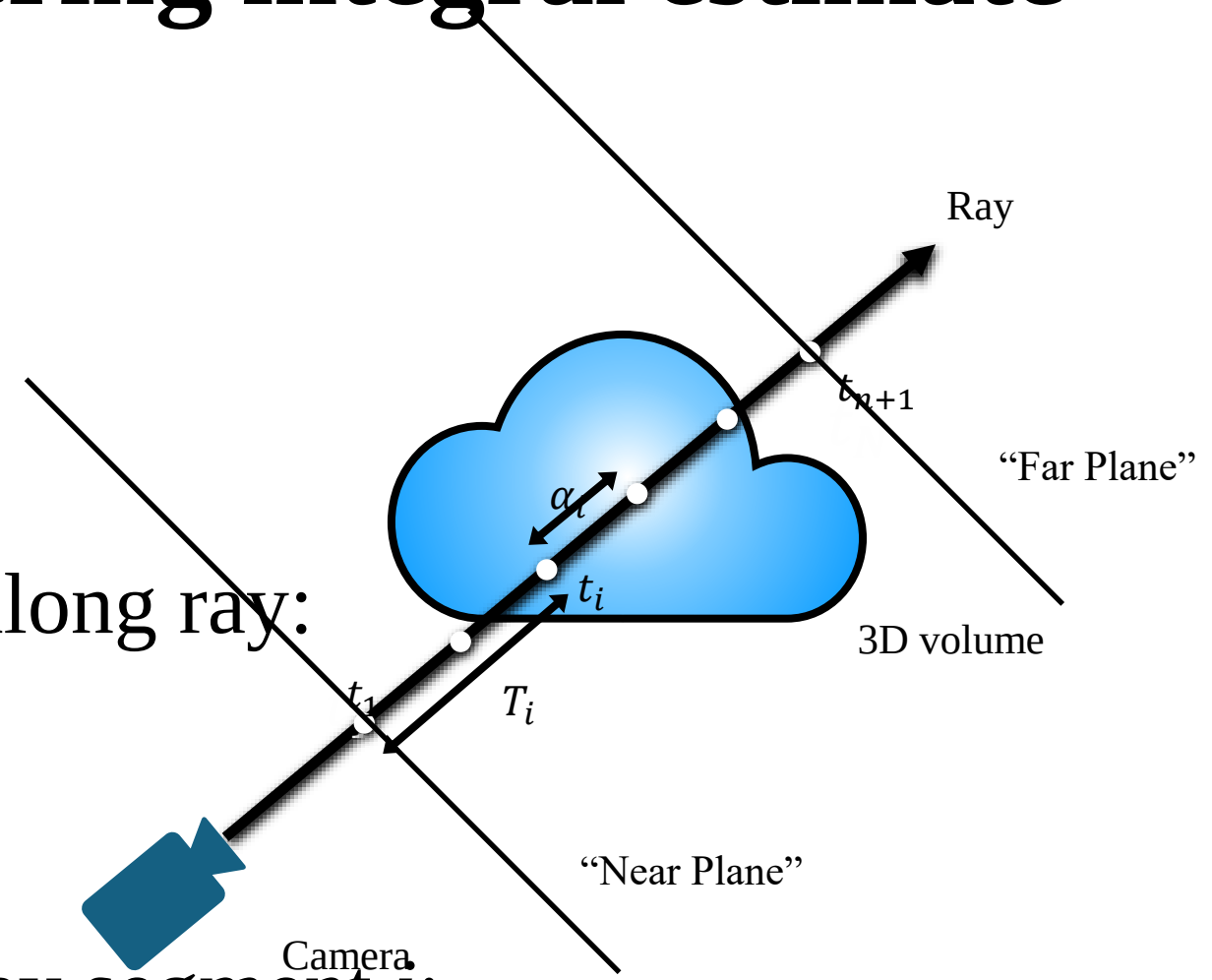
weights

How much light is blocked earlier along ray:

$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$$

How much light is contributed by ray segment i:

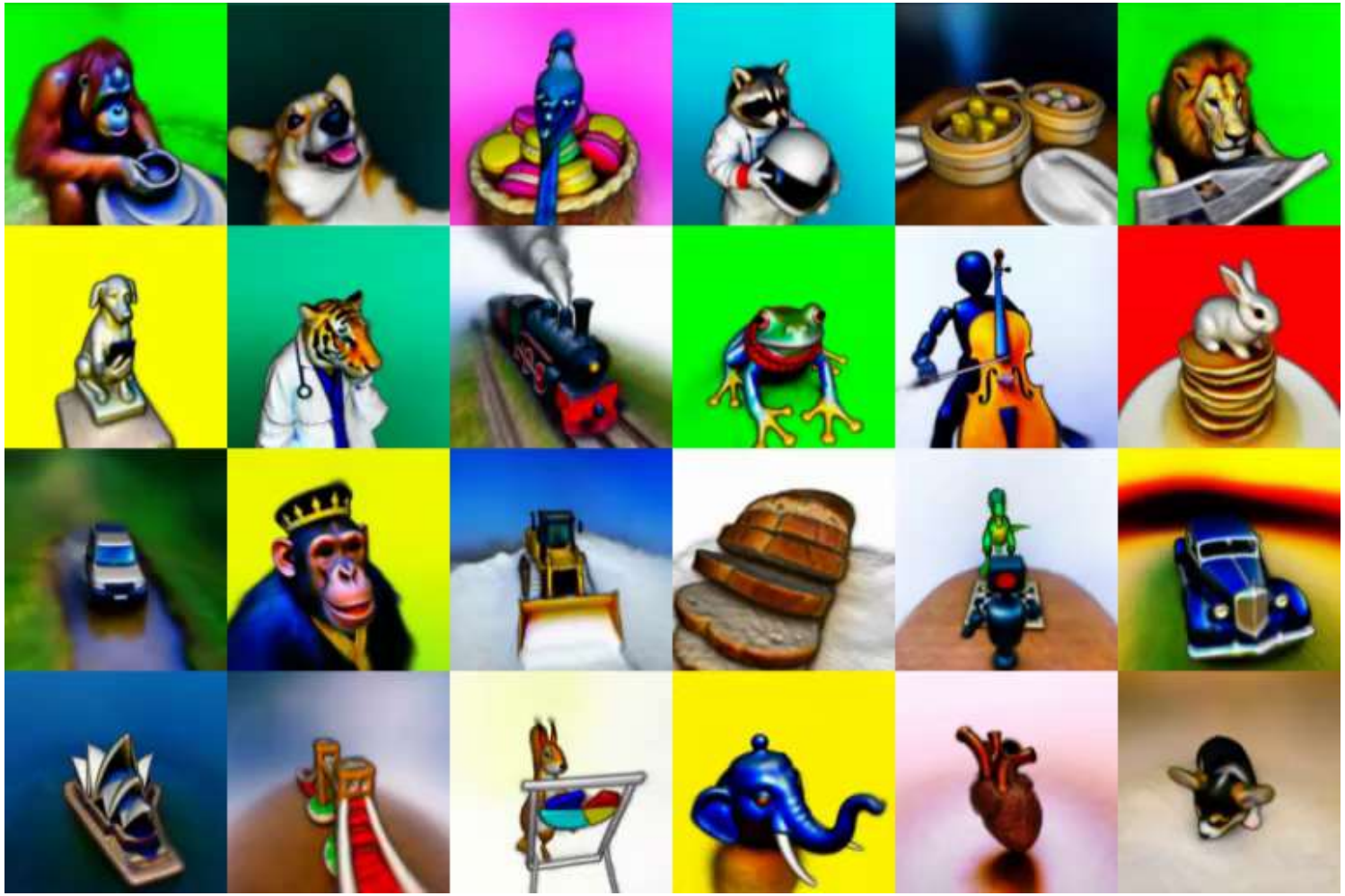
$$\alpha_i = 1 - \exp(-\sigma_i \delta_i)$$



Neural Radiance Fields, Mildenhall et al. 2021

Slide credit: Ben Mildenhall

DreamFusion: Text-to-3D using 2D Diffusion



https://arxiv.org/pdf/2209.14988

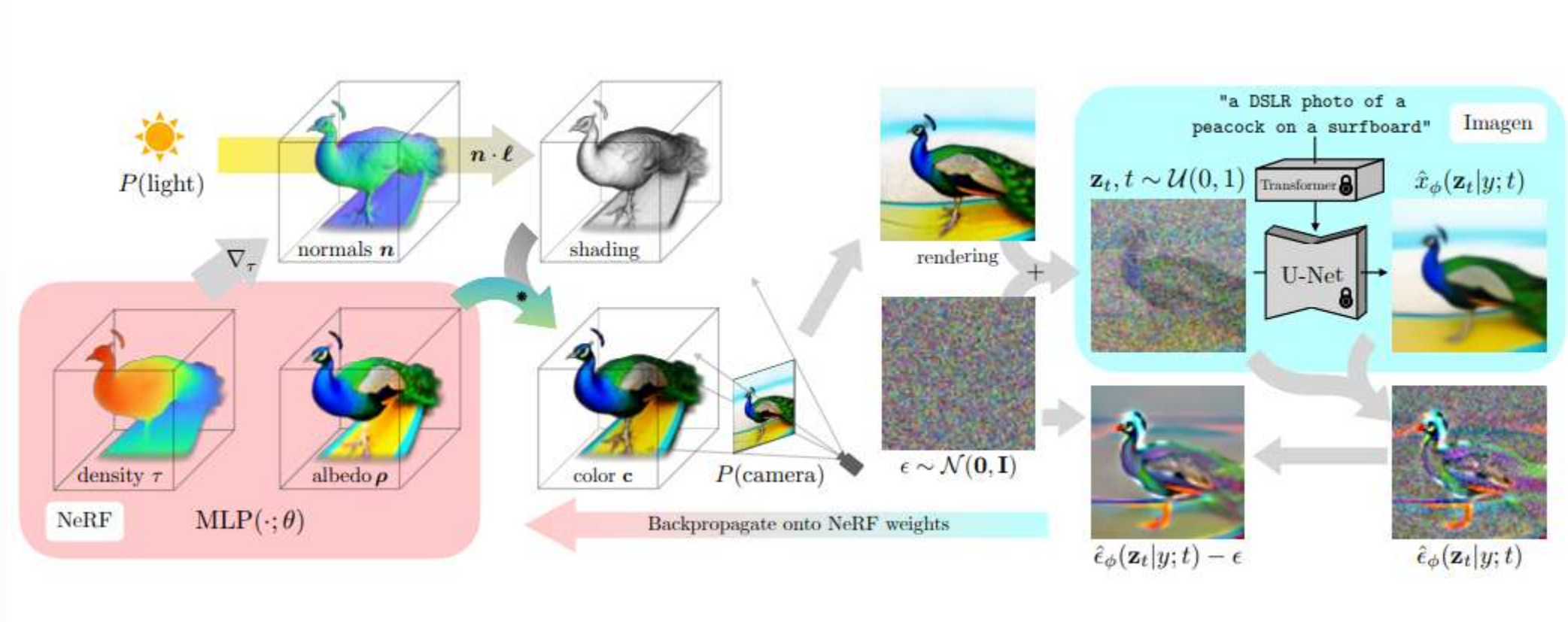
DreamFusion: Text-to-3D using 2D Diffusion

"a DSLR photo of a peacock on a surfboard"

DreamFusion
Automatic text-to-3D

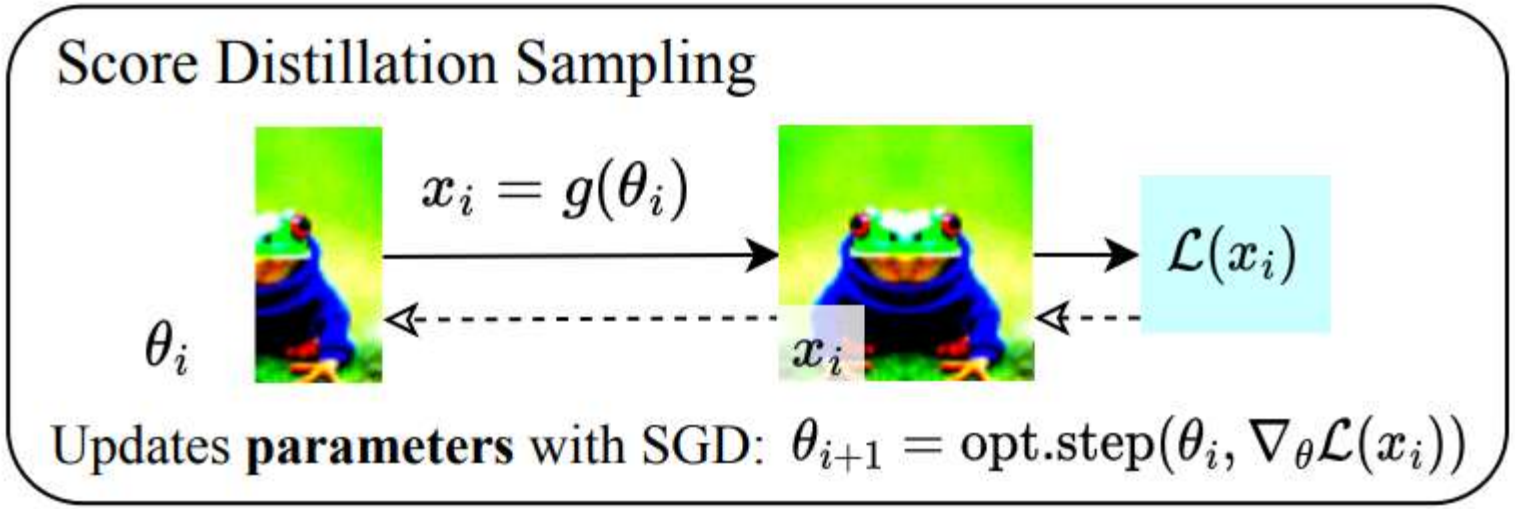
https://arxiv.org/pdf/2209.14988

DreamFusion: Text-to-3D using 2D Diffusion



https://arxiv.org/pdf/2209.14988

DreamFusion: Text-to-3D using 2D Diffusion



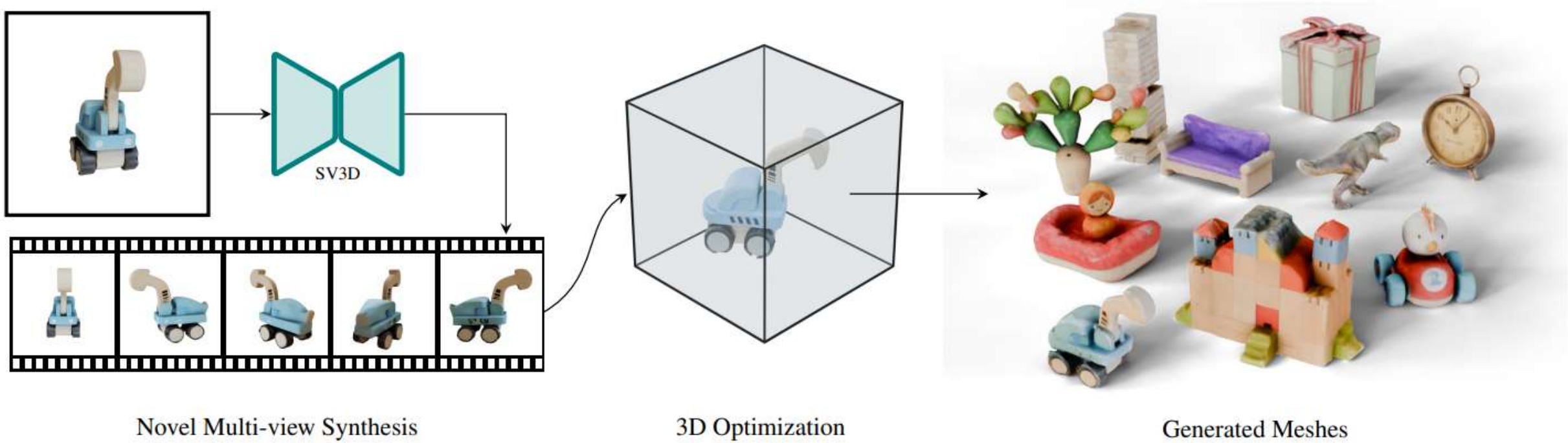
$$\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[w(t)\|\epsilon_\phi(\alpha_t\mathbf{x} + \sigma_t\epsilon; t) - \epsilon\|_2^2\right]$$

https://arxiv.org/pdf/2209.14988

SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion

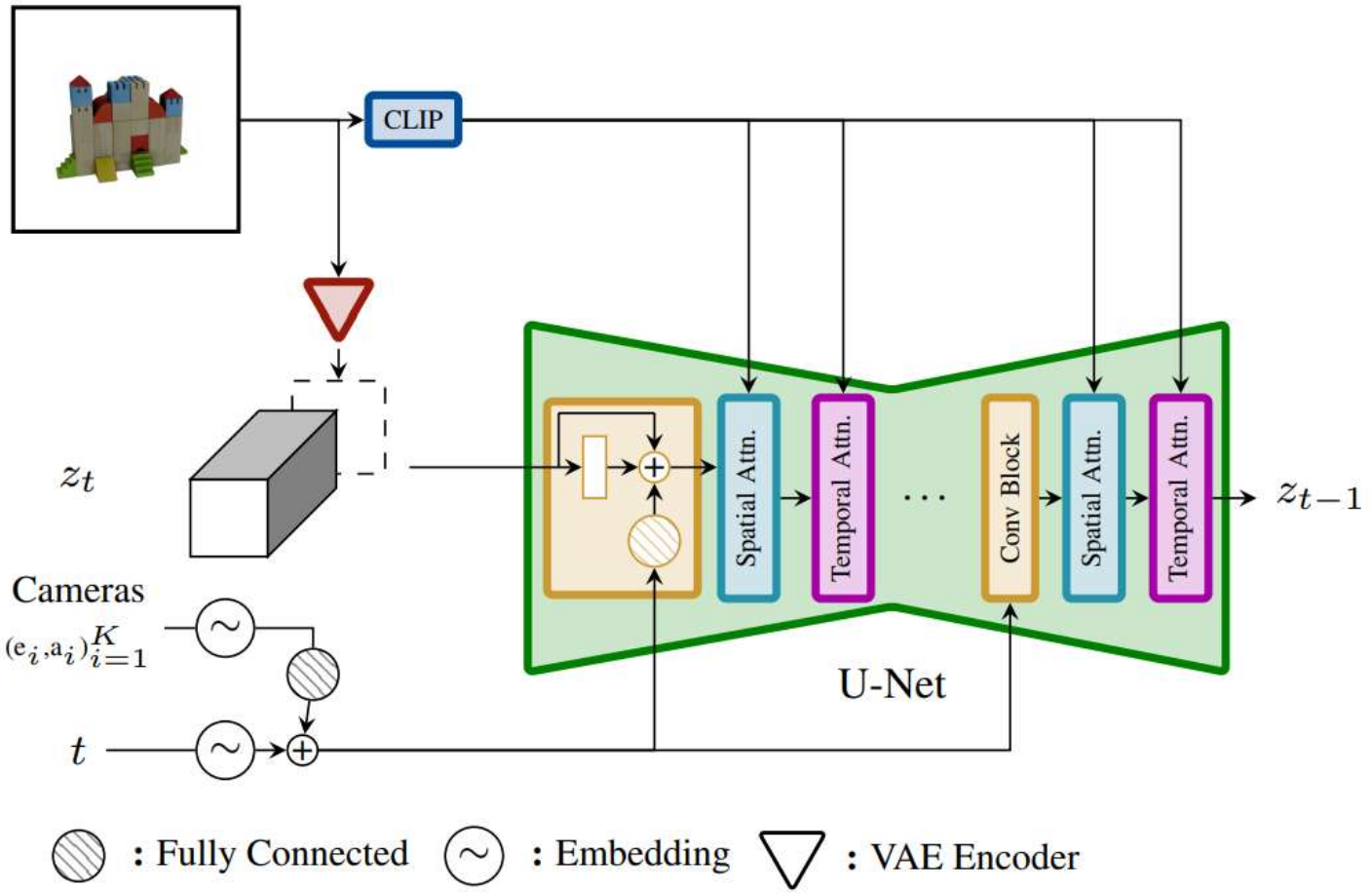
https://arxiv.org/pdf/2403.12008

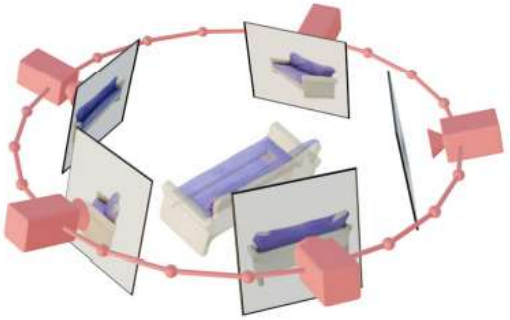
SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion



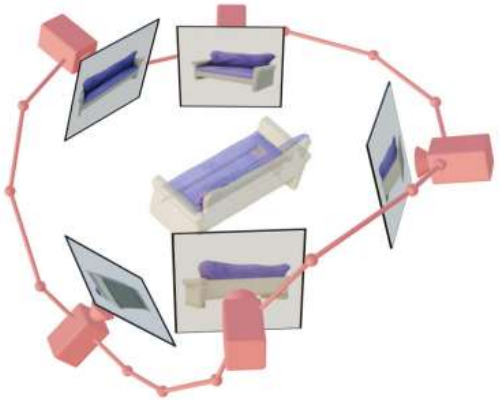
https://arxiv.org/pdf/2403.12008

SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion





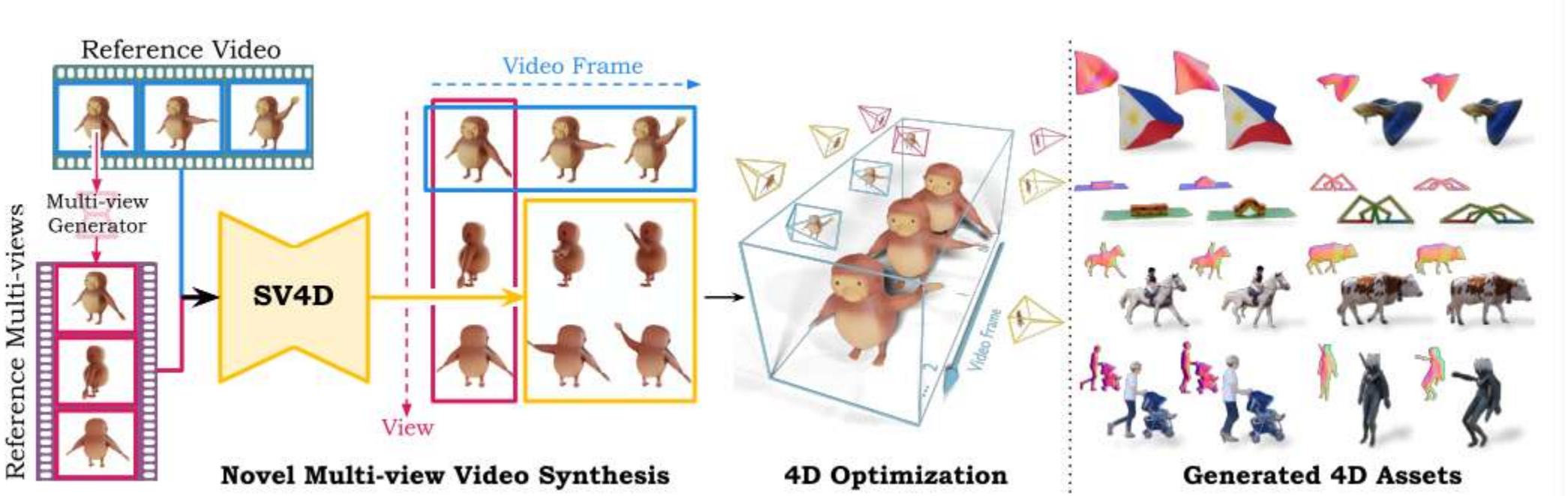
Static



Dynamic

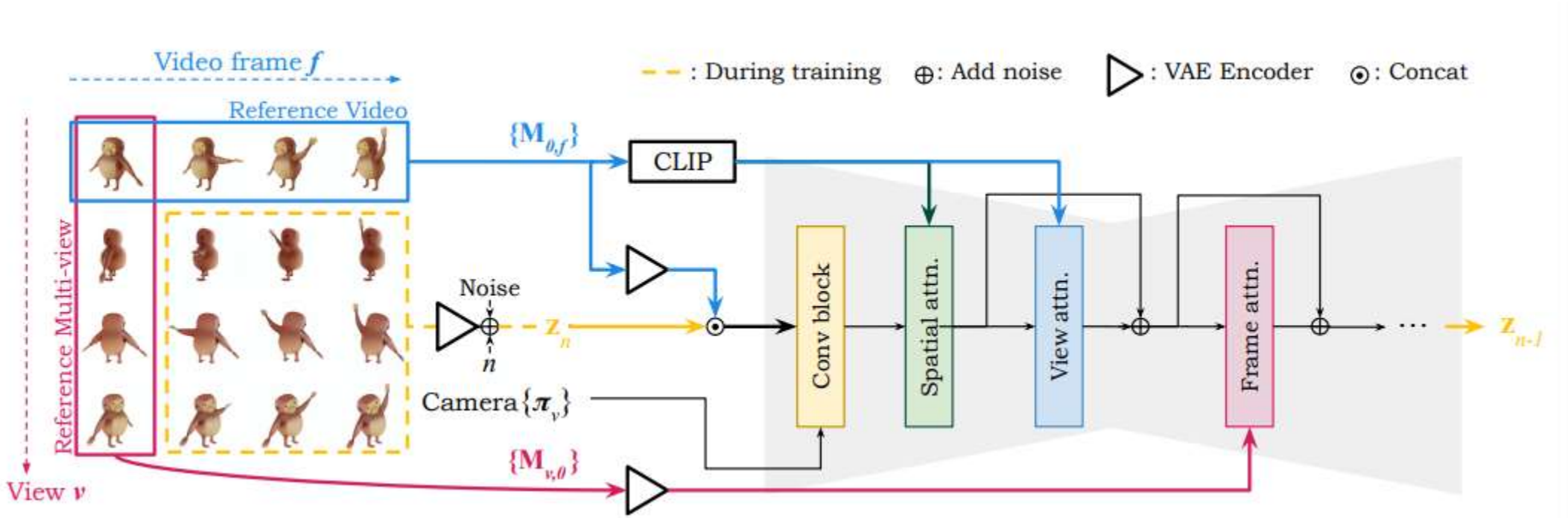
https://arxiv.org/pdf/2403.12008

SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency



https://arxiv.org/abs/2407.17470

SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency

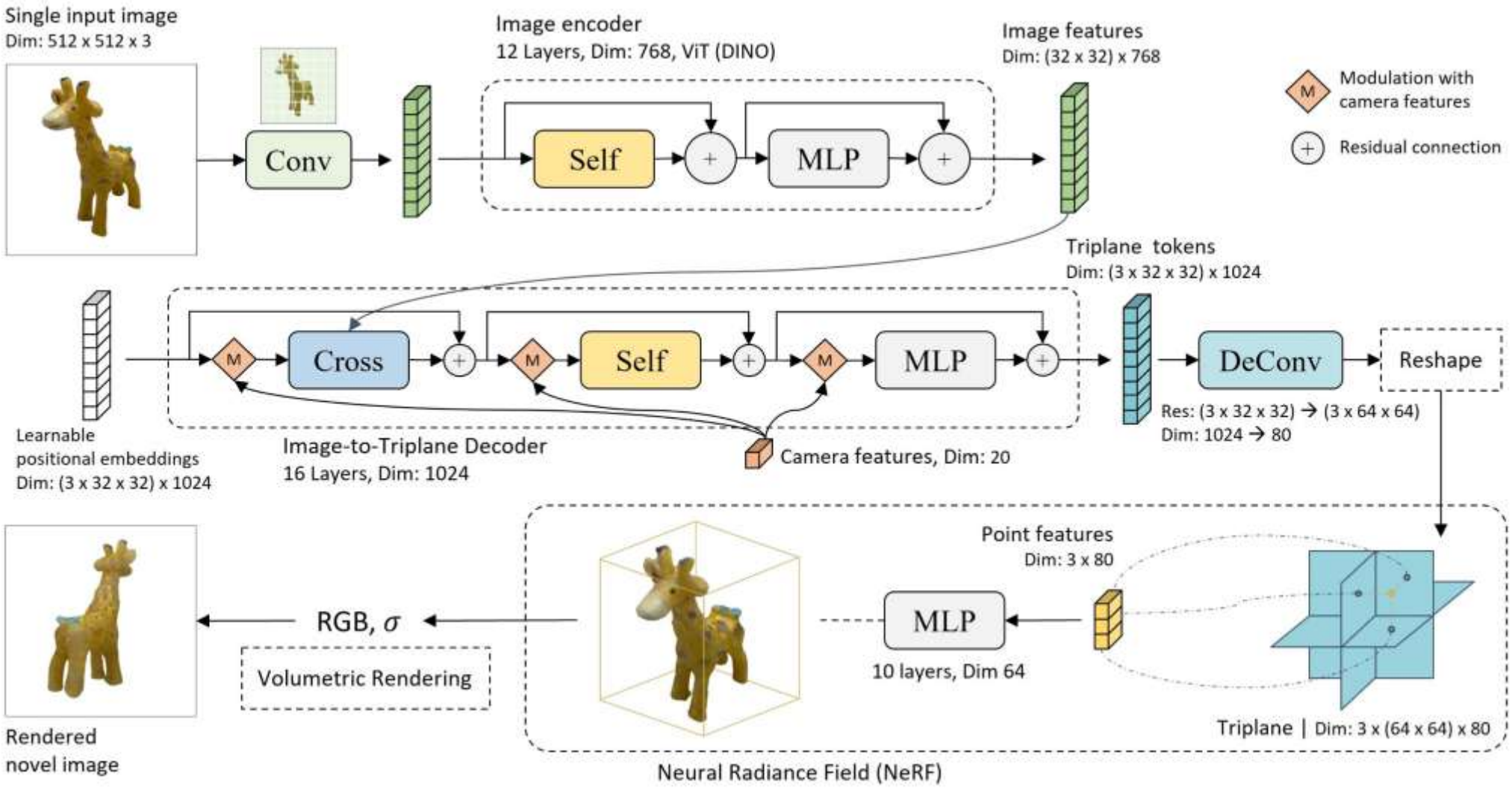


https://arxiv.org/abs/2407.17470

SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency

https://arxiv.org/abs/2407.17470

LRM: LARGE RECONSTRUCTION MODEL FOR SINGLE IMAGE TO 3D

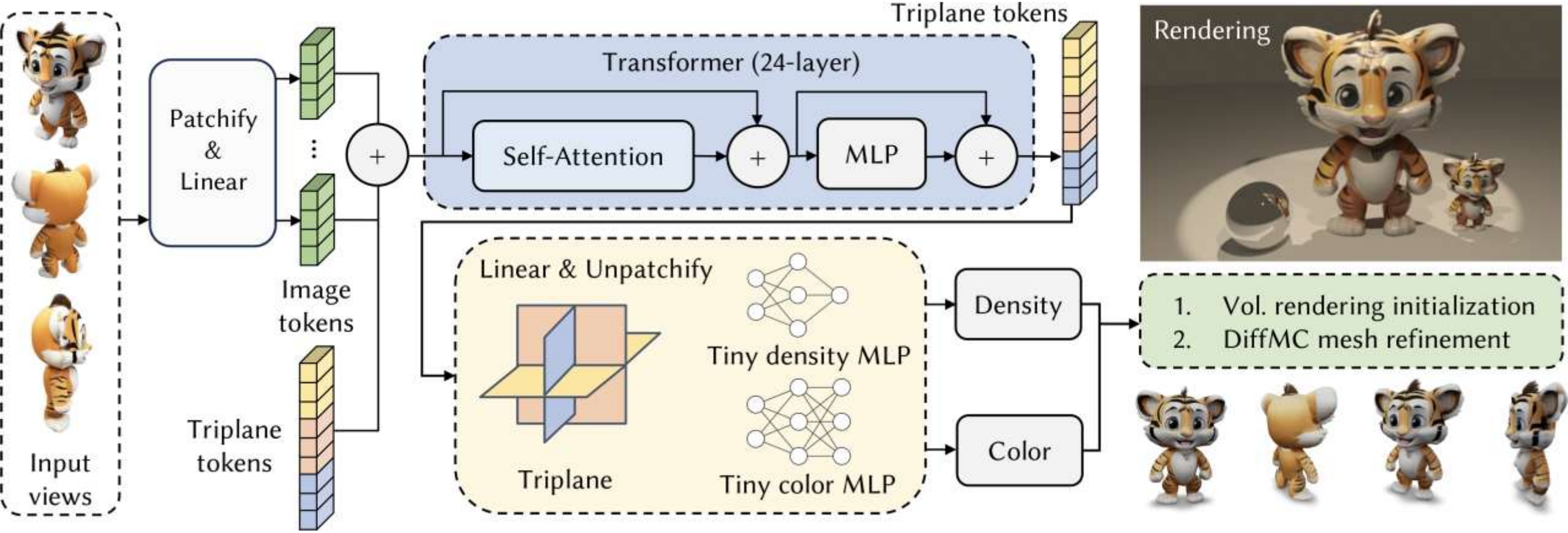


https://arxiv.org/pdf/2311.04400

LRM: LARGE RECONSTRUCTION MODEL FOR SINGLE IMAGE TO 3D

https://arxiv.org/pdf/2311.04400

MeshLRM

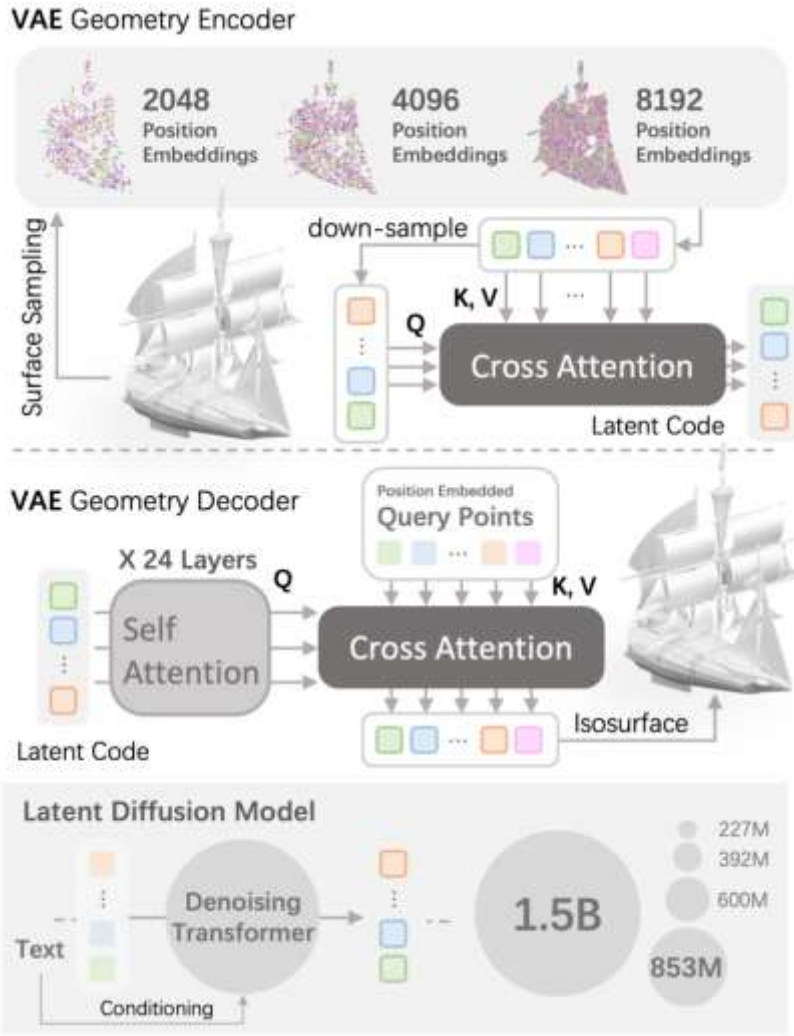


https://arxiv.org/abs/2404.12385

MeshLRM

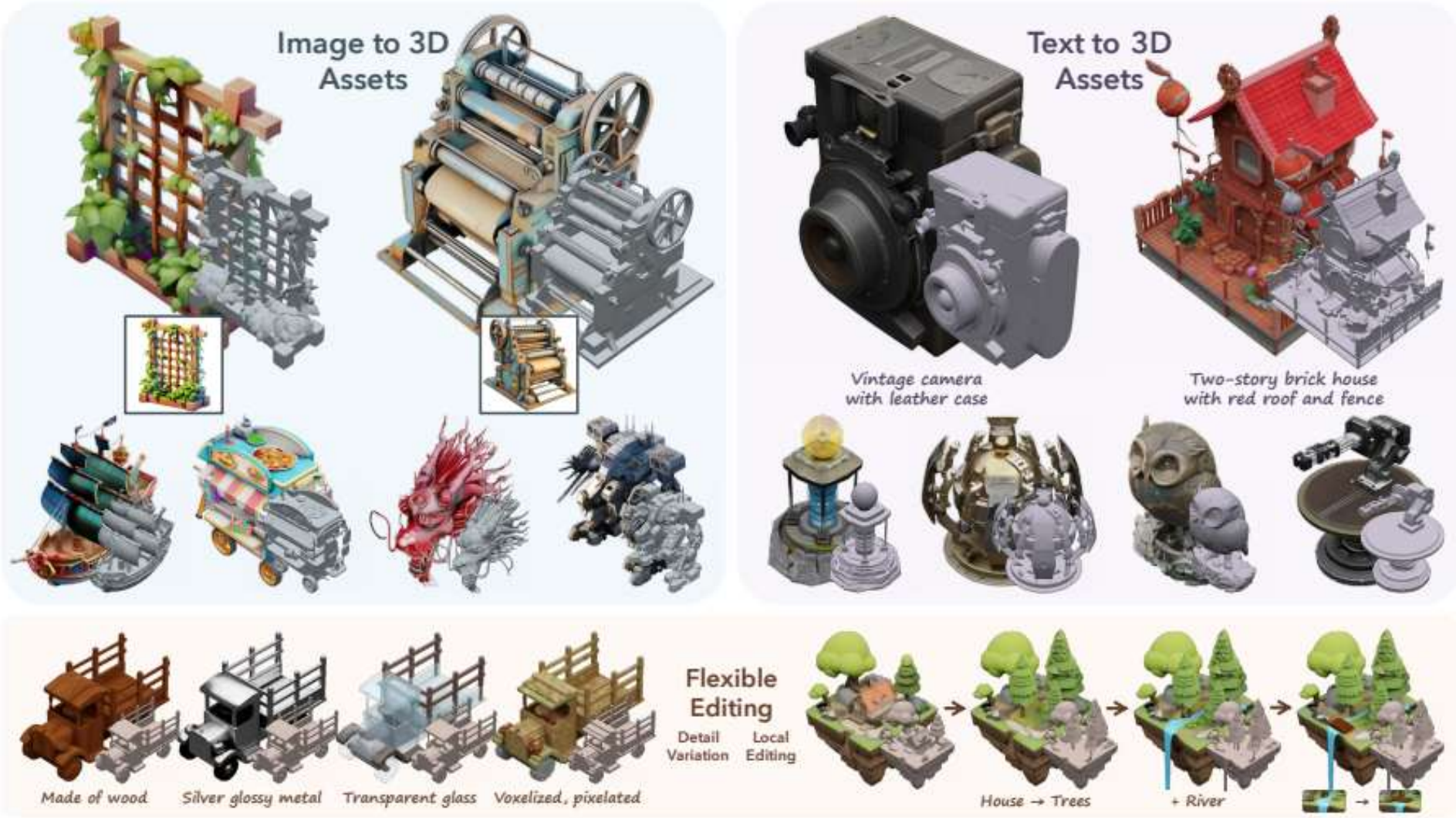
https://arxiv.org/abs/2404.12385

CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets



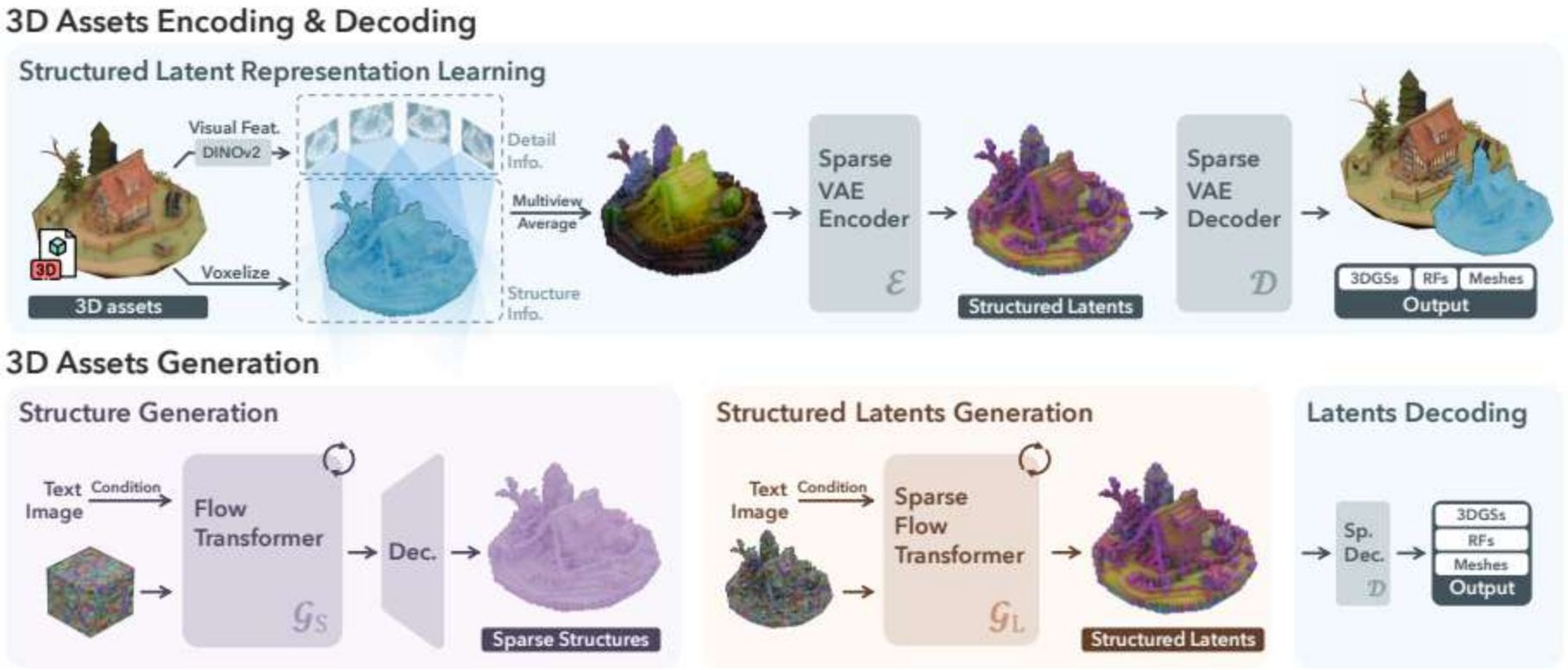
https://arxiv.org/pdf/2406.13897

TRELLIS



https://microsoft.github.io/TRELLIS/

TRELLIS

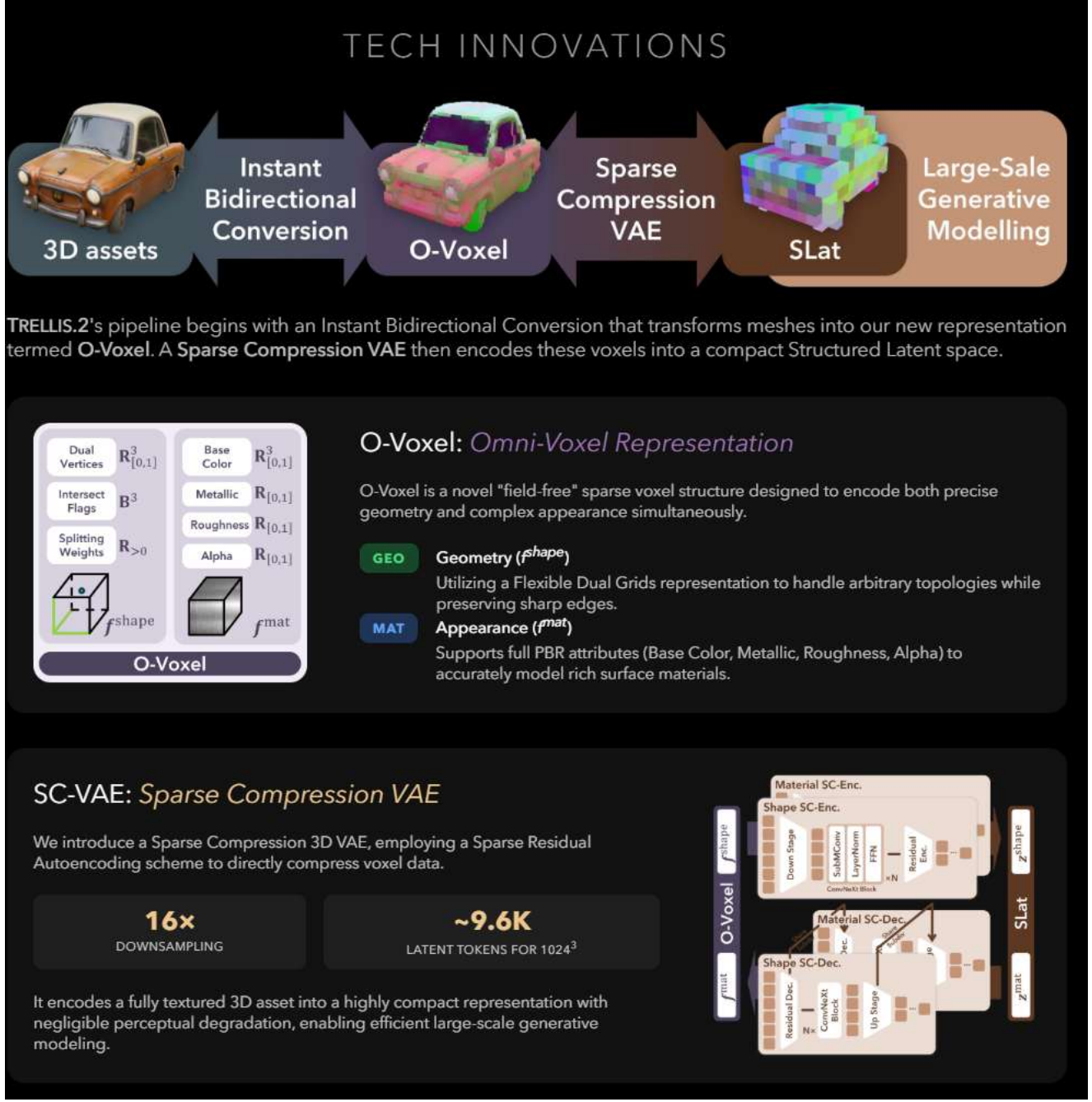


https://microsoft.github.io/TRELLIS/

TRELLIS.2

https://microsoft.github.io/TRELLIS.2/

TRELLIS.2

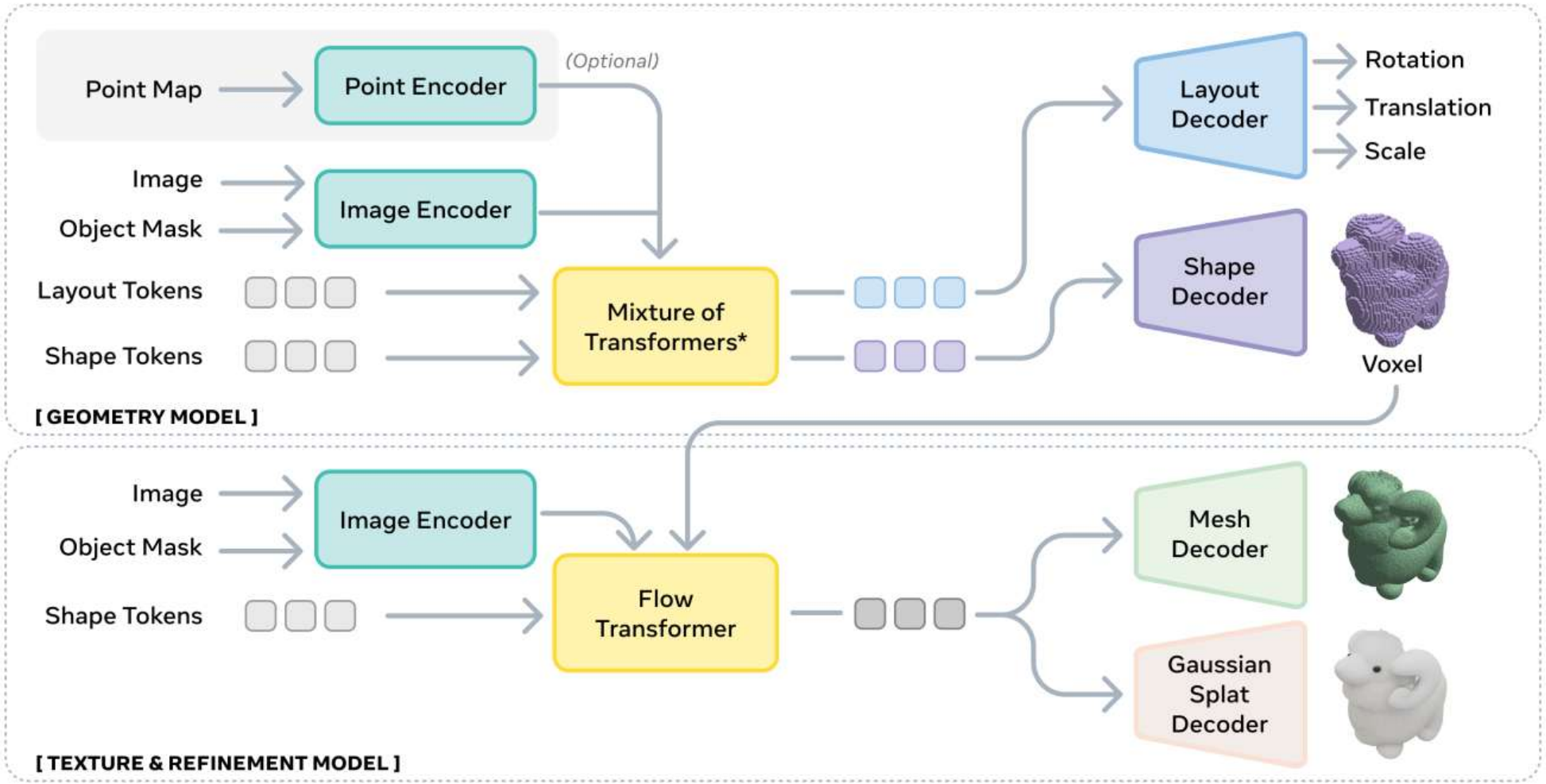


https://microsoft.github.io/TRELLIS.2/

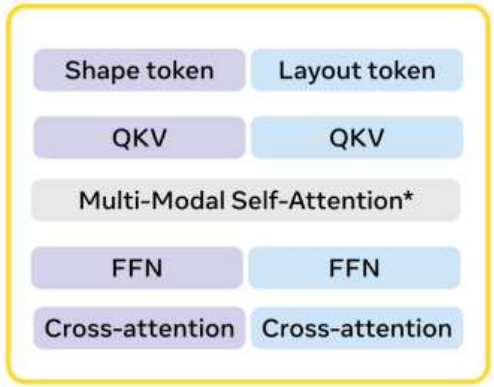
SAM 3D

https://arxiv.org/pdf/2511.16624

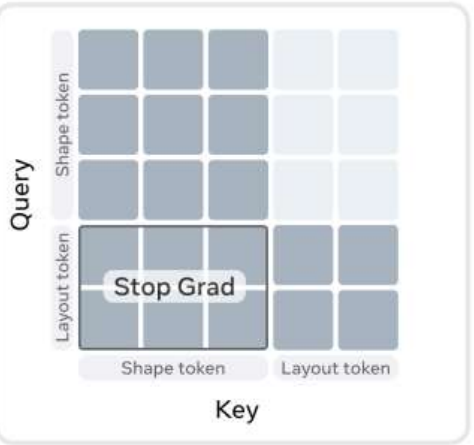
SAM 3D



*Mixture of Transformers

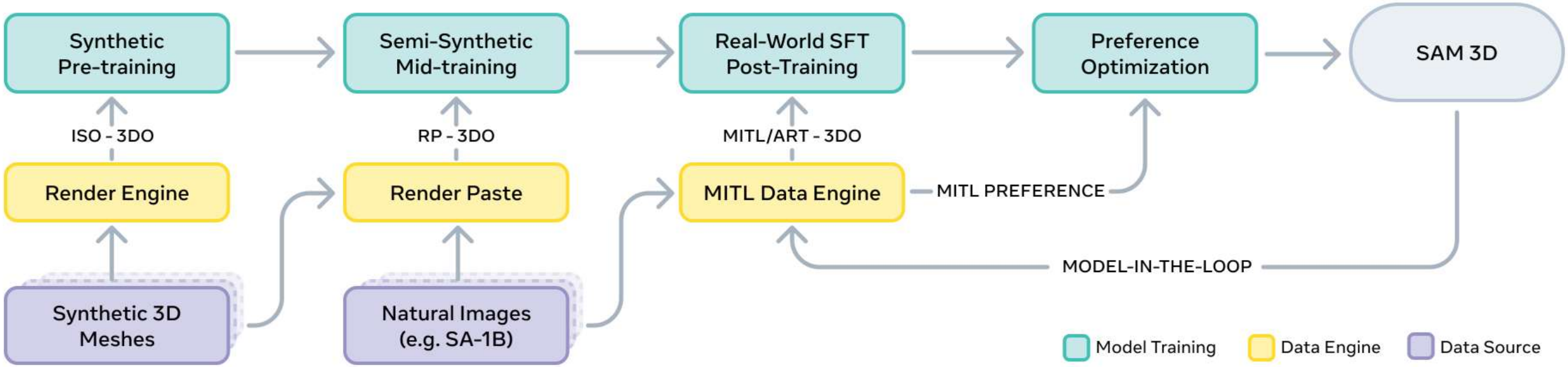


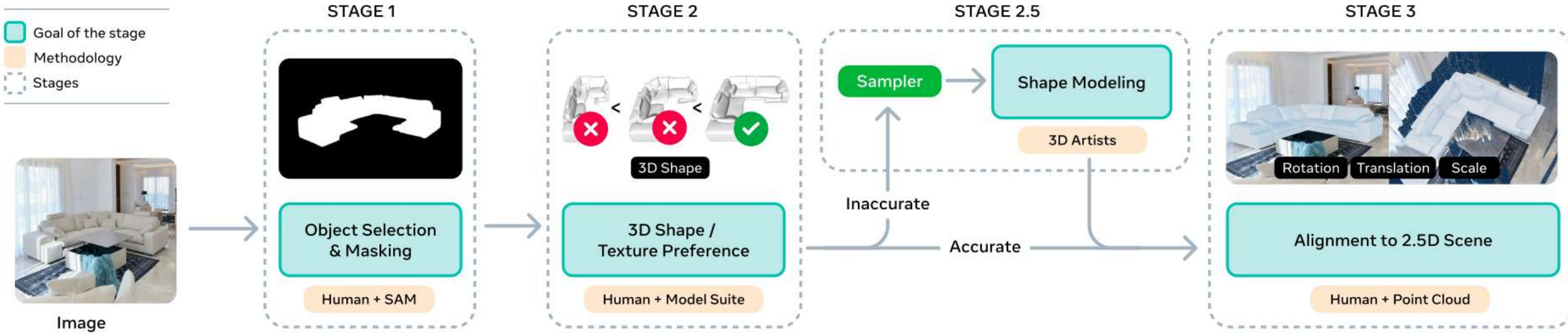
*Multi-Modal Self-Attention Mask



https://arxiv.org/pdf/2511.16624

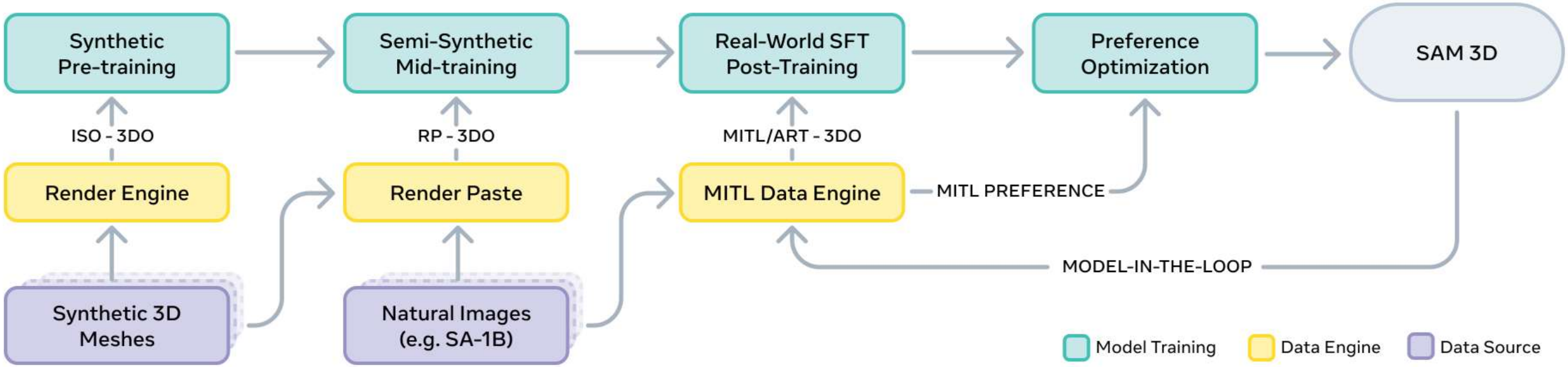
SAM 3D

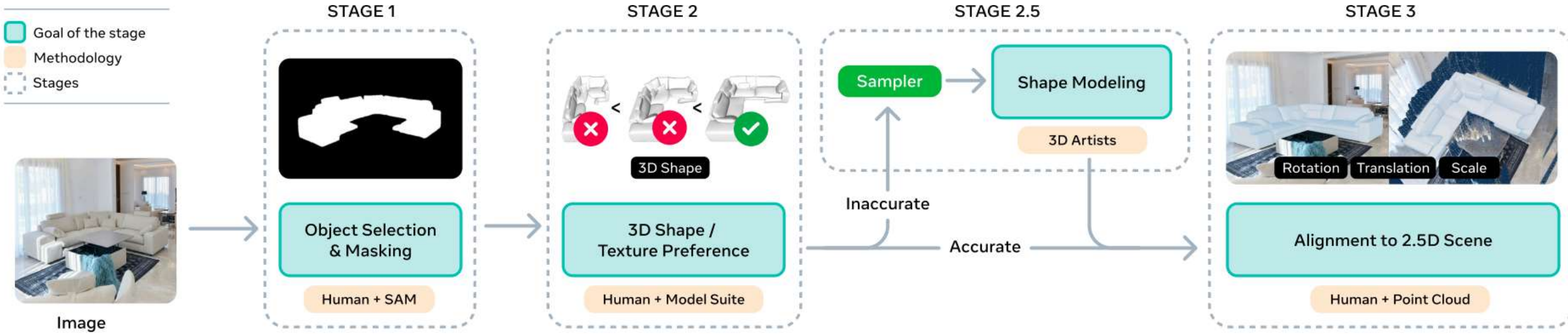




https://arxiv.org/pdf/2511.16624

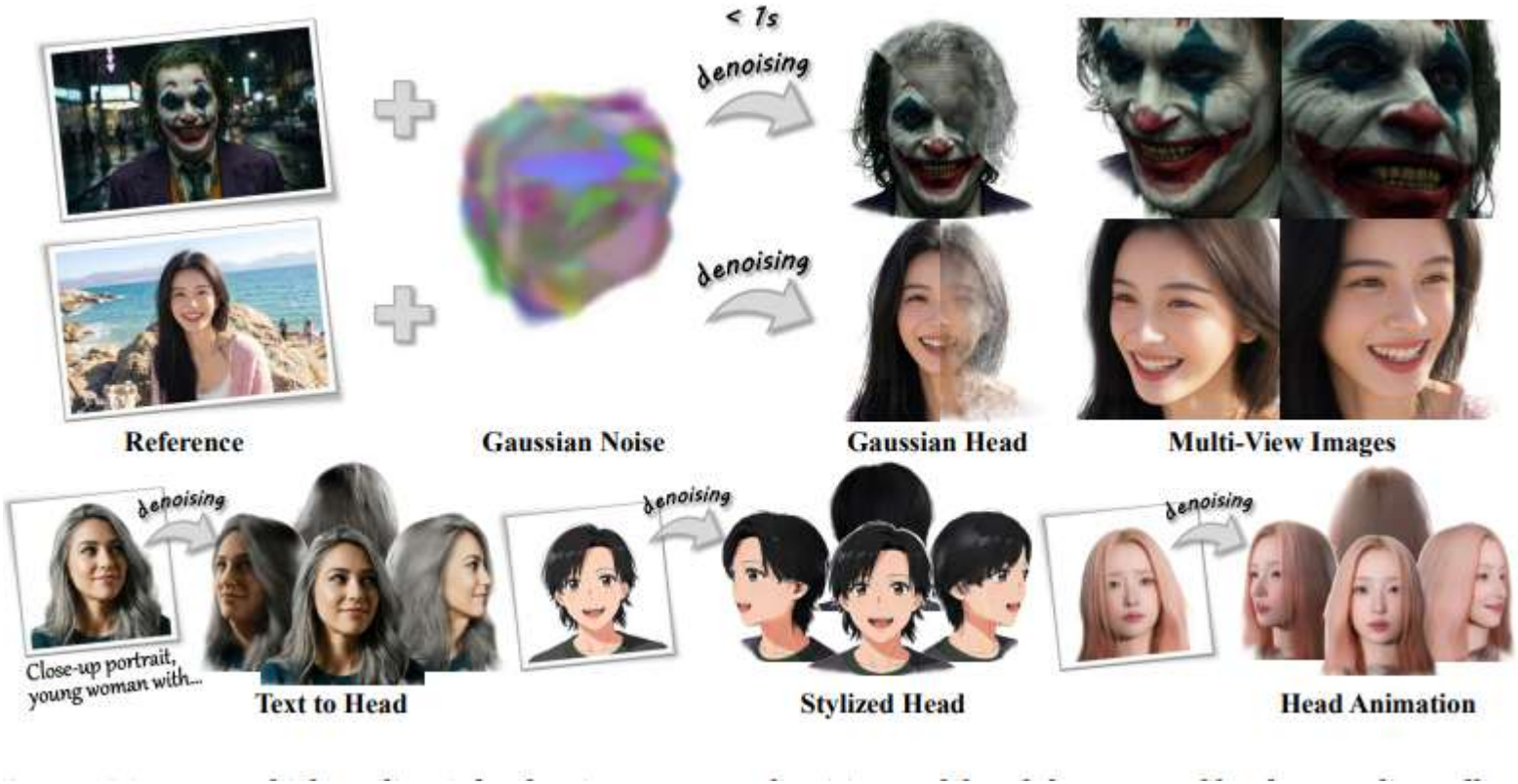
SAM 3D





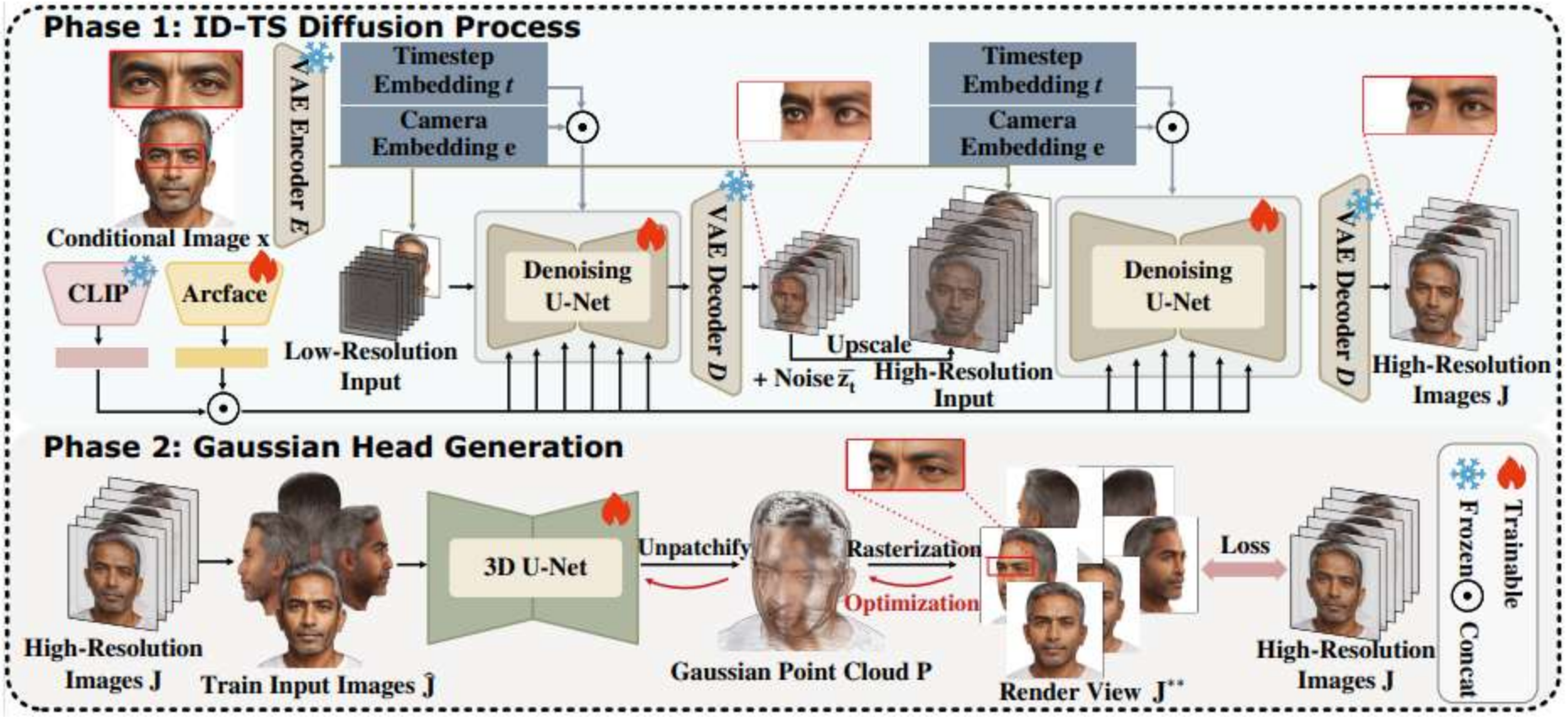
https://arxiv.org/pdf/2511.16624

AnyHead & AnyAvatar3D



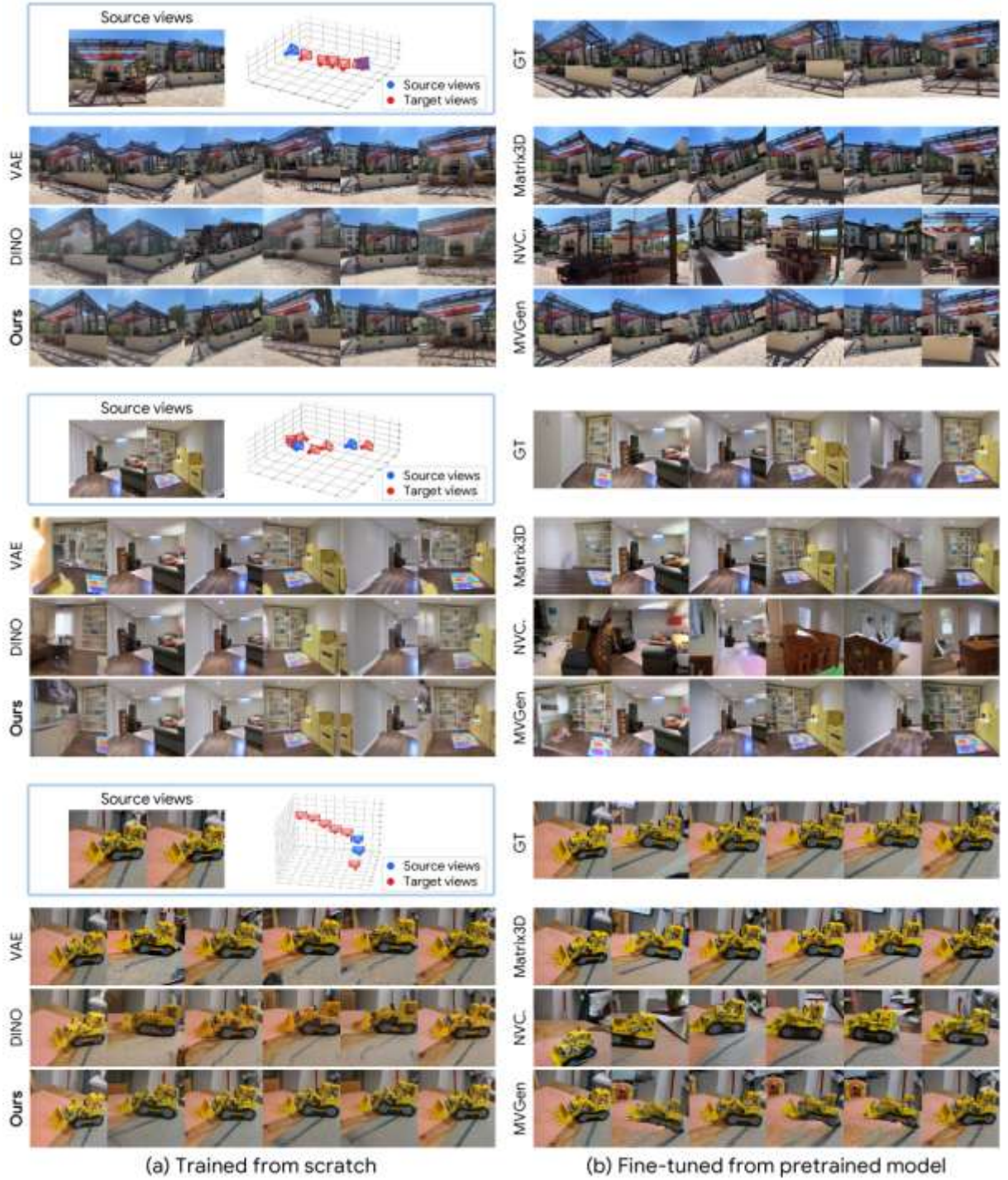
https://arxiv.org/pdf/2604.13856

AnyHead & AnyAvatar3D



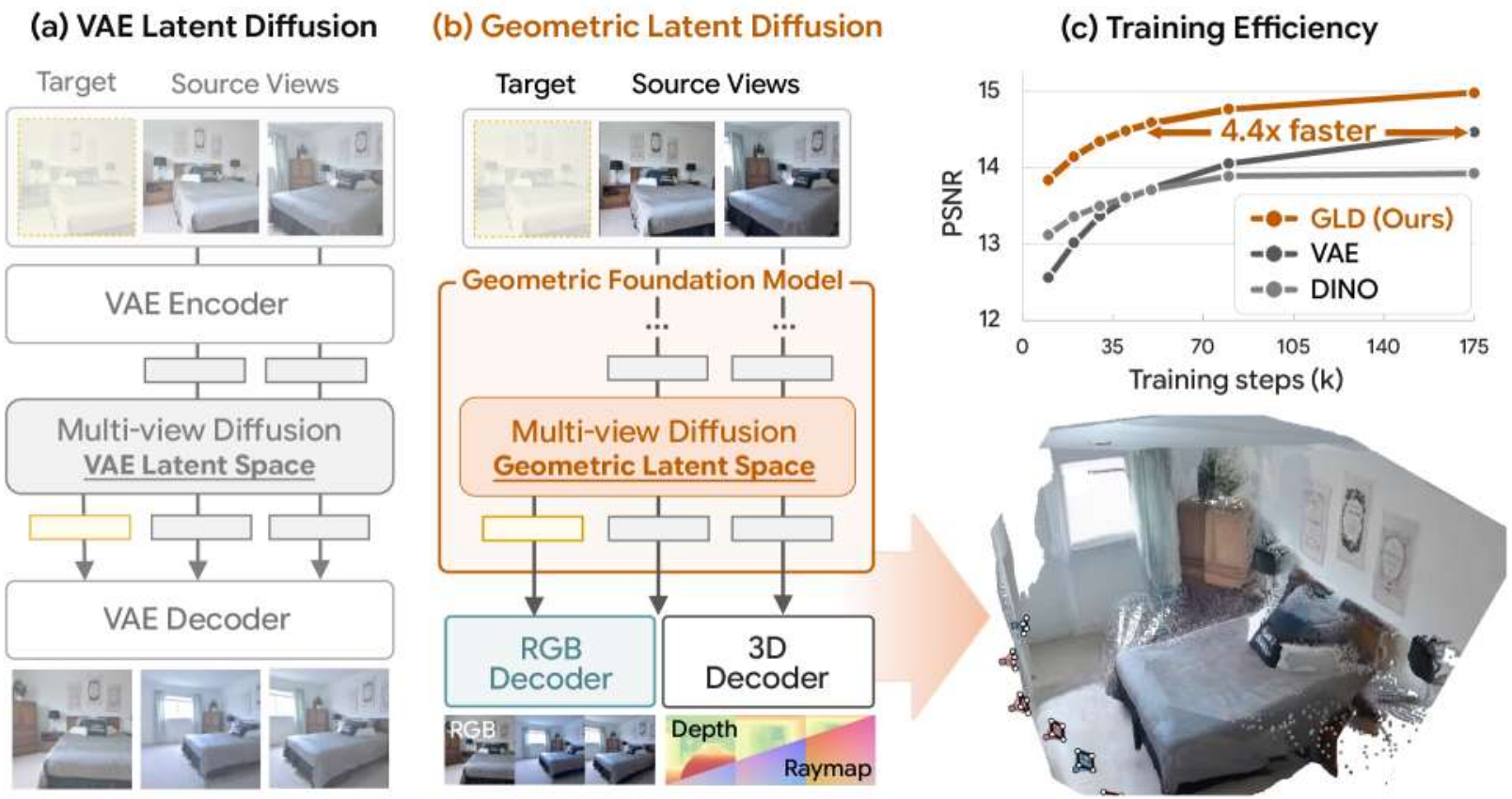
https://arxiv.org/pdf/2604.13856

GLD: Repurposing Geometric Foundation Models for Multi-view Diffusion



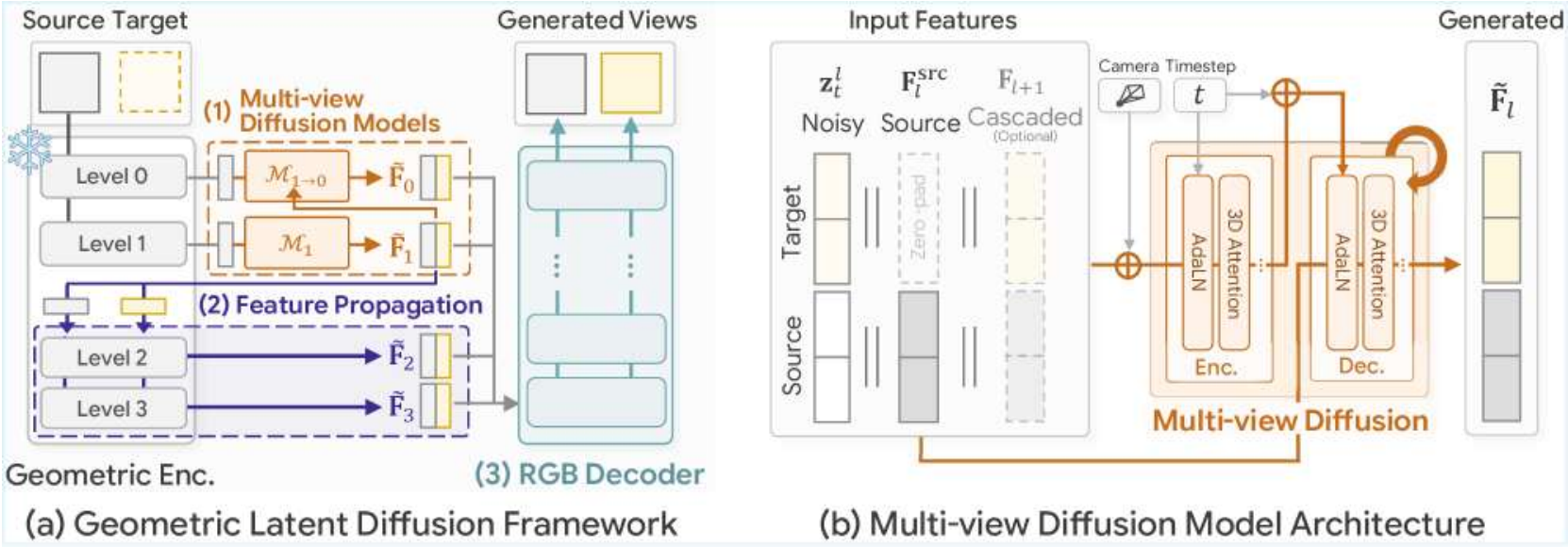
https://arxiv.org/pdf/2511.16624

GLD：Repurposing Geometric Foundation Models for Multi-view Diffusion



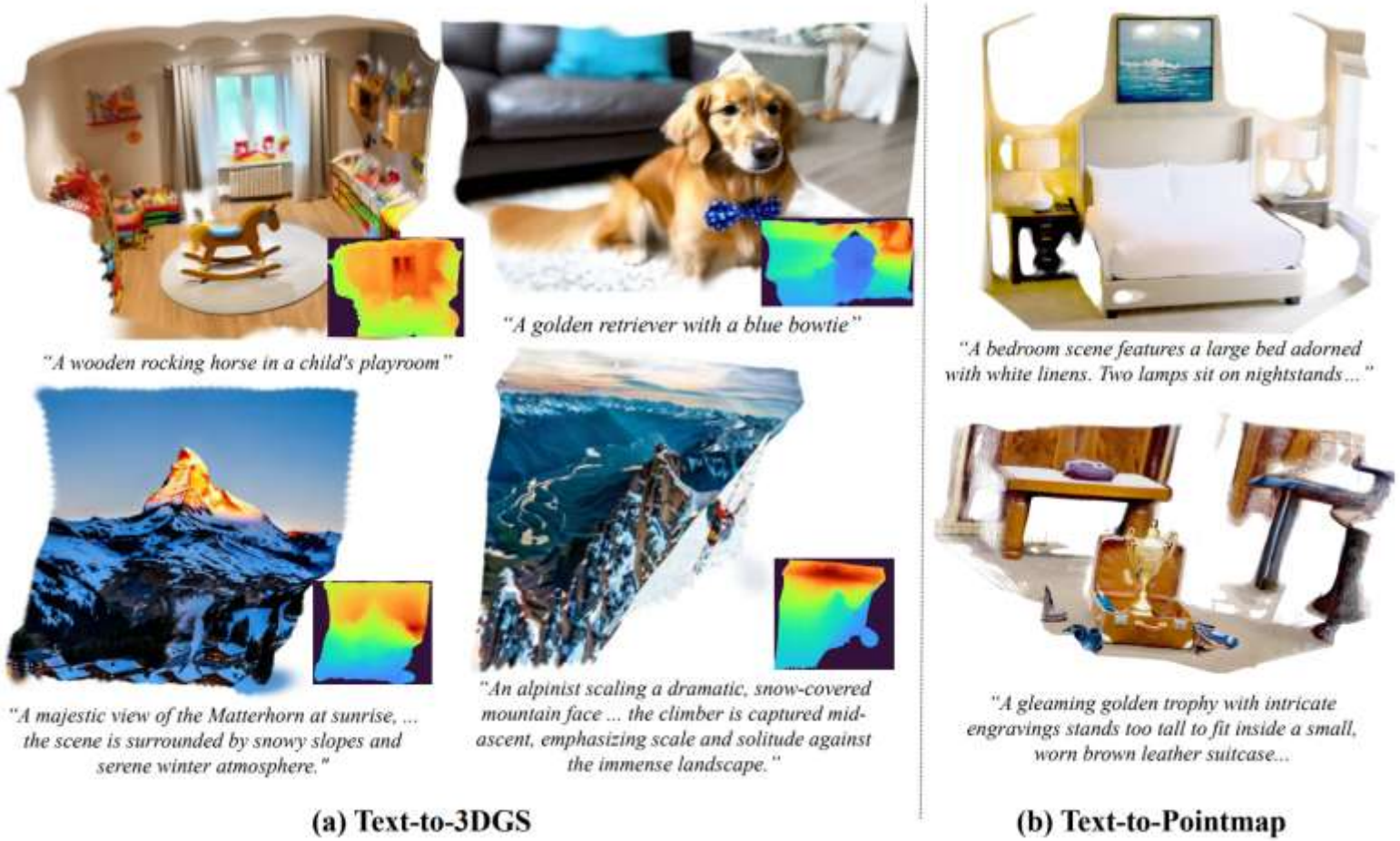
https://arxiv.org/pdf/2511.16624

GLD：Repurposing Geometric Foundation Models for Multi-view Diffusion



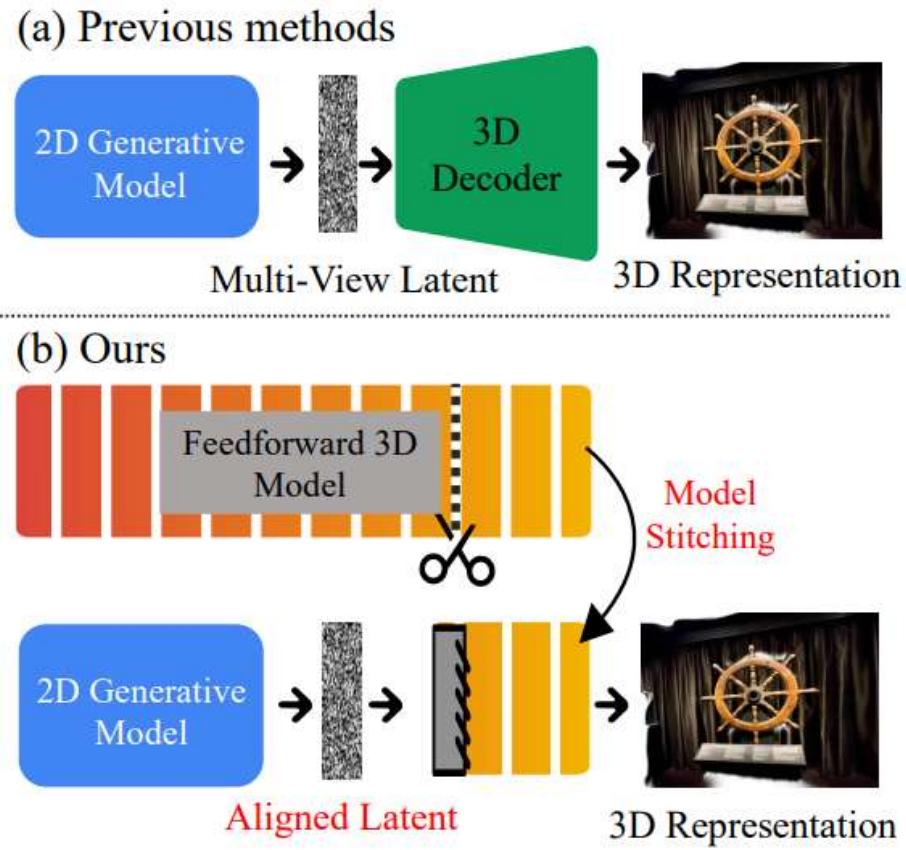
https://arxiv.org/pdf/2511.16624

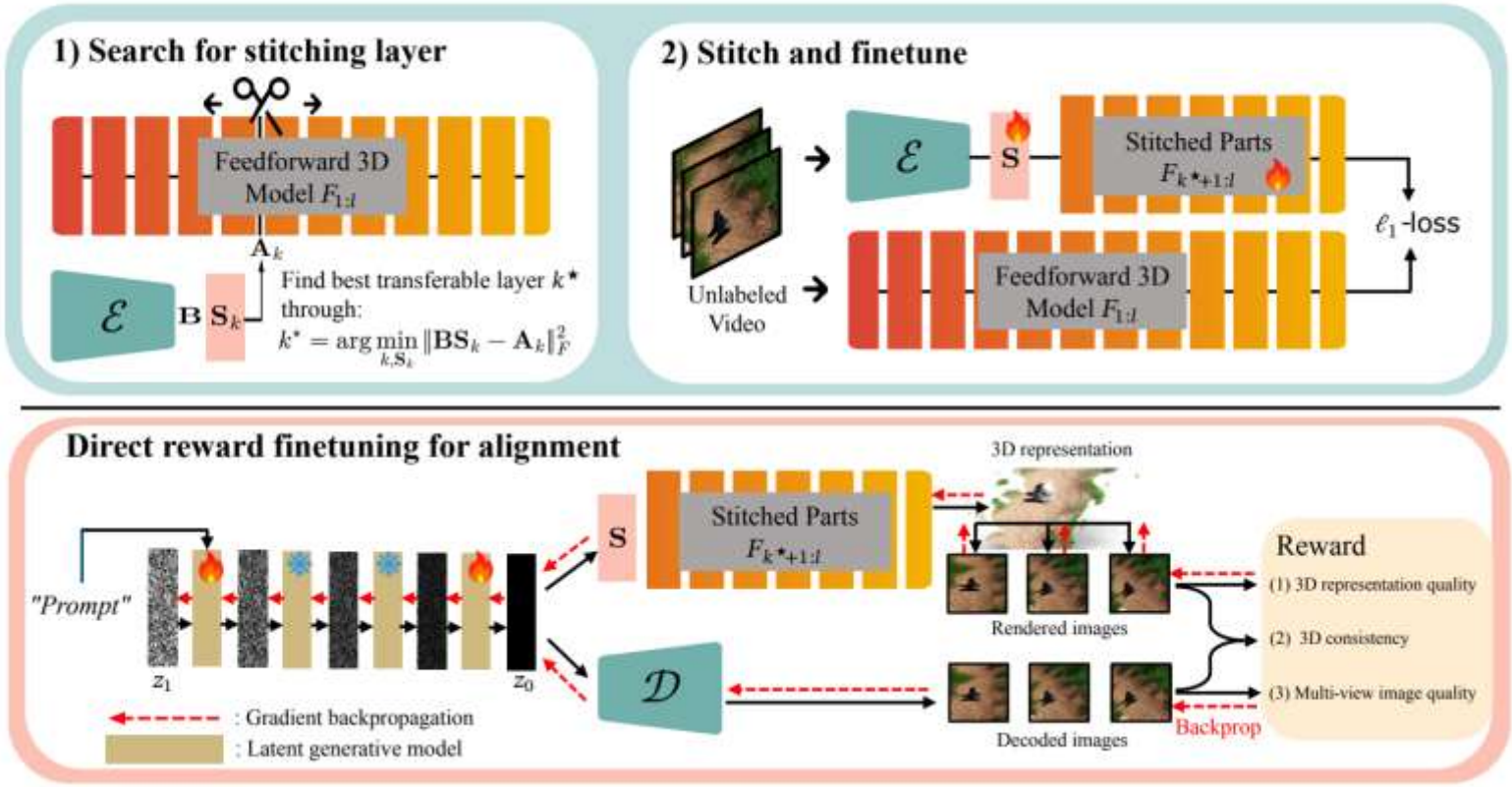
VIST3A：Text-to-3D by Stitching a Multi-view Reconstruction Network to a Video Generator



https://arxiv.org/abs/2510.13454

VIST3A：Text-to-3D by Stitching a Multi-view Reconstruction Network to a Video Generator

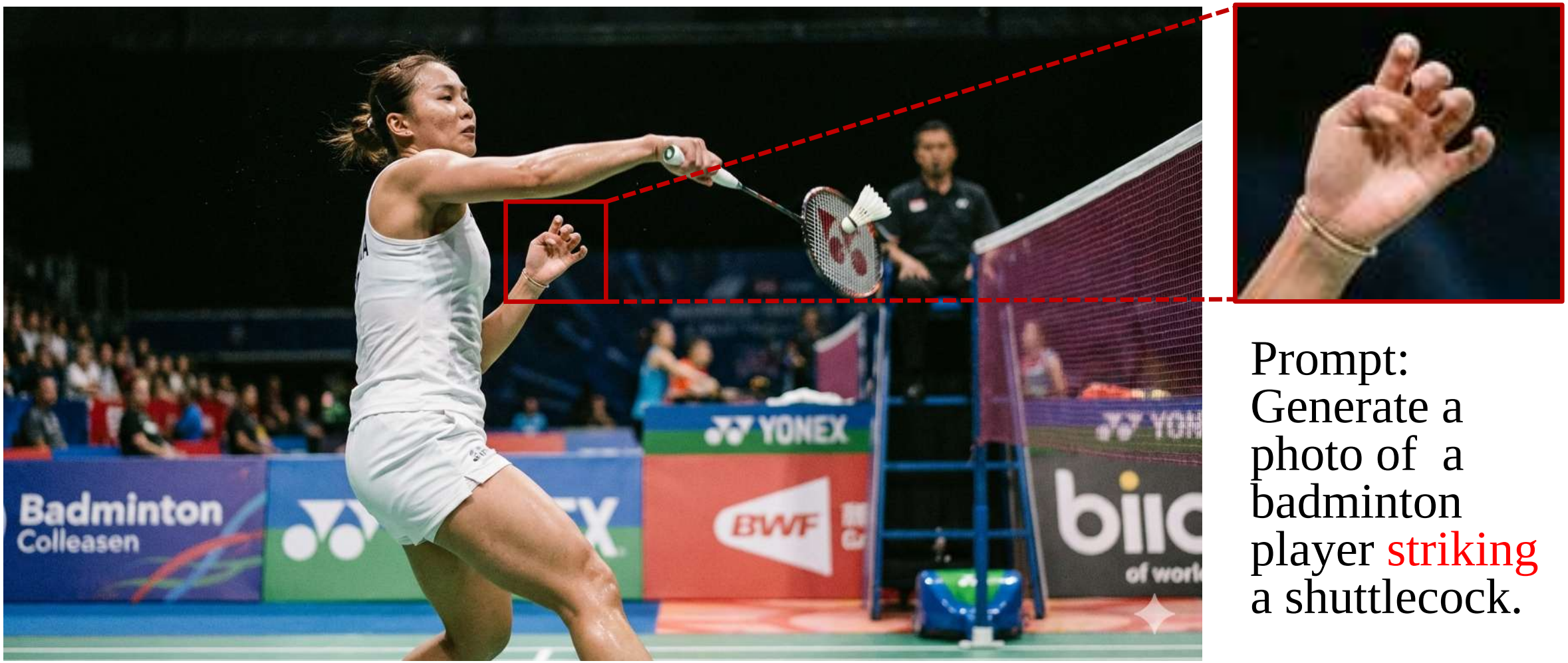




https://arxiv.org/abs/2510.13454

Detection of AI-Generated Visual Content

Issues of Generative Models: Controllability



Prompt: Generate a photo of a badminton player striking a shuttlecock.

Nano Banana Pro

Issues of Generative Models: Safety

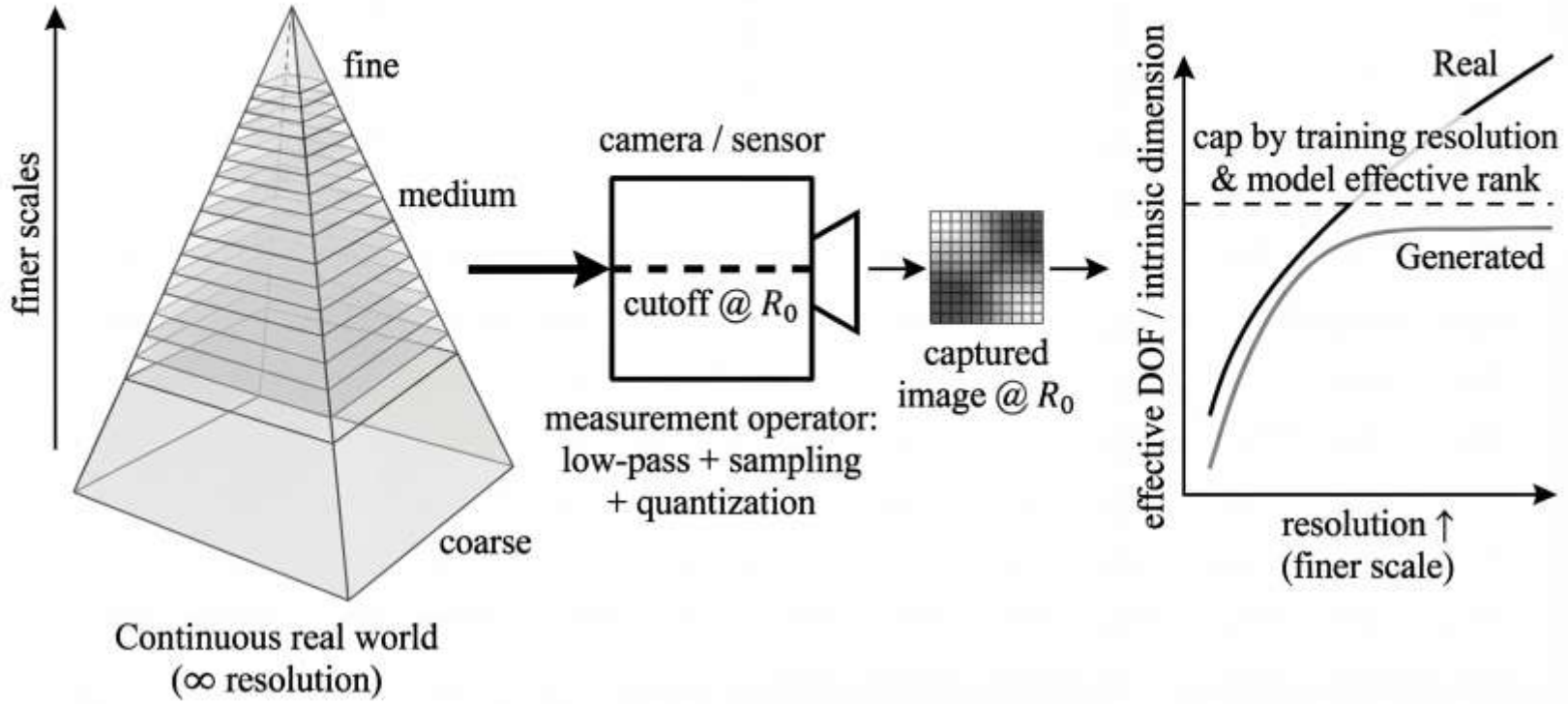
Real vs. Generated Images: A Multi-Scale Geometry Gap

- **Implication.**
  - The Real–Gen gap is not just pixel noise; it is a systematic shift in representation-space geometry, and it is scale-dependent.

- **Observation.**
  - Real images come from an effectively continuous world: as we probe finer scales, new coherent structure keeps emerging.
  - **Generated images are trained in a fixed-resolution regime**, so fine-scale statistics often **saturate**, showing either **compression** (over-smooth / repetitive textures) or **distortion** (hallucinated, incoherent micro-structure).

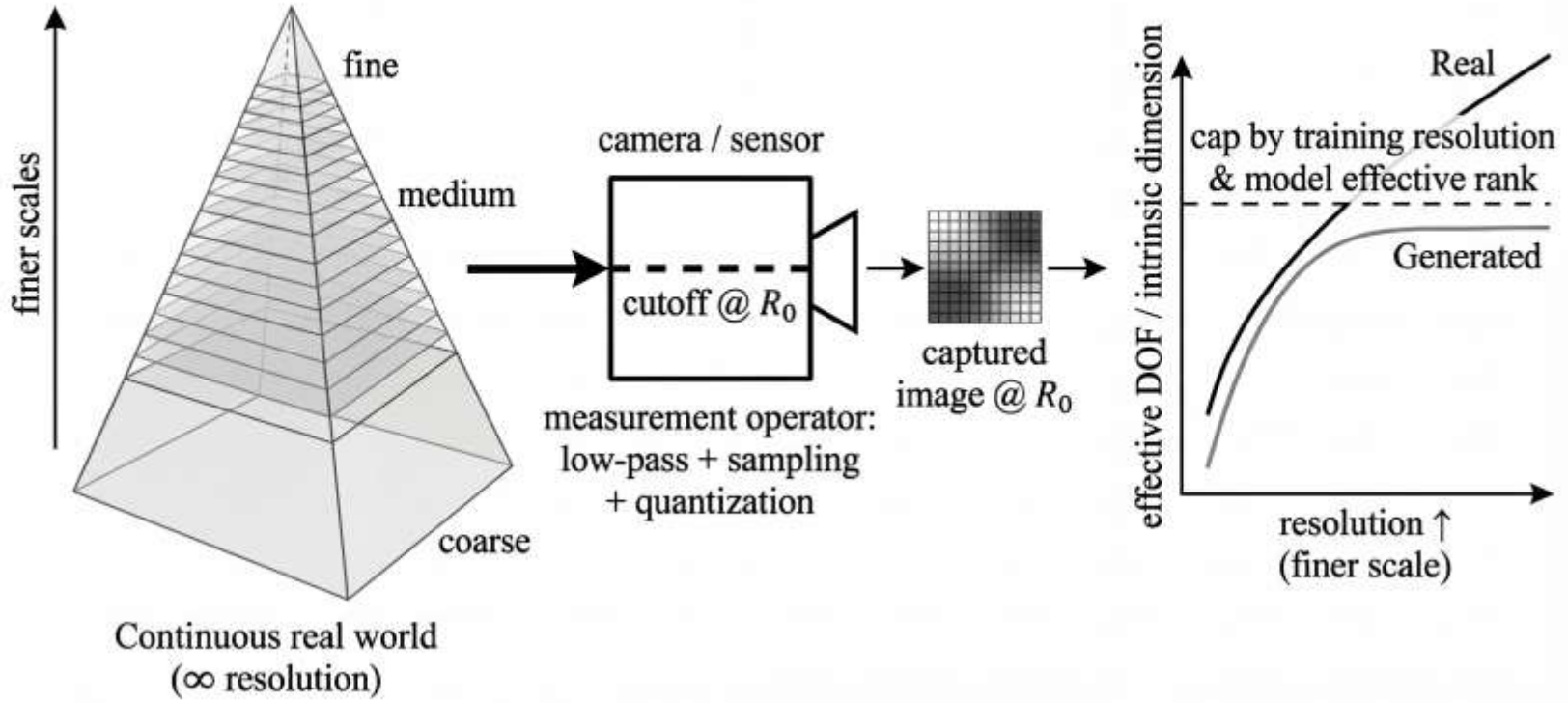
Real vs. Generated Images: A Multi-Scale Geometry Gap

- **Why this happens?**
  - **Measurement vs. learning mismatch:** the camera provides a finite-resolution *measurement* of a continuous scene (low-pass + sampling). Generators learn a **band-limited approximation** tied to the training resolution.

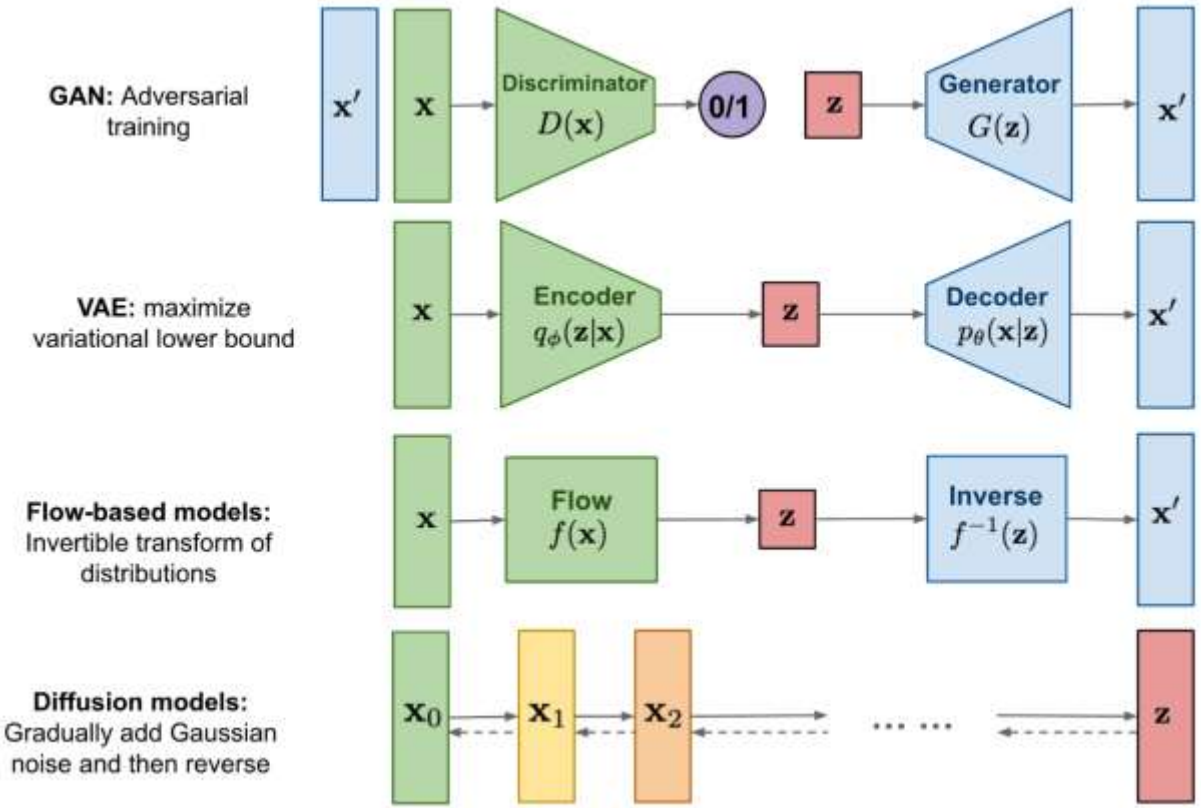


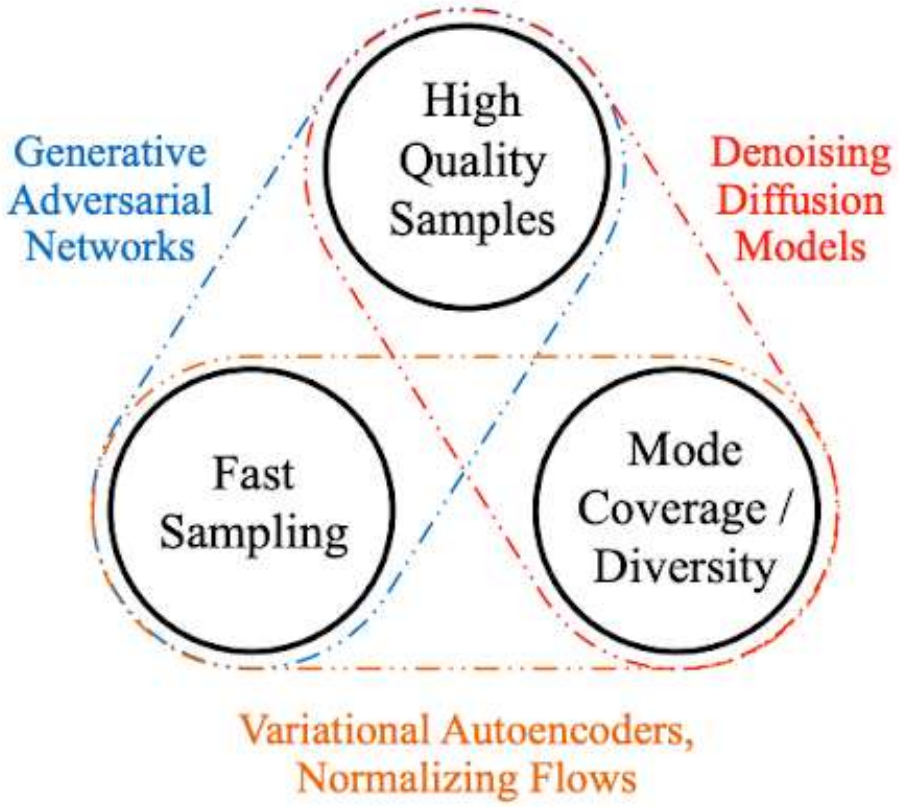
Real vs. Generated Images: A Multi-Scale Geometry Gap

- **Why this happens?**
  - **Effective DOF cap (model/optimization bias):** the generator parameterization and training objective impose a **bounded effective dimension / thinner manifold**, so extra fine-scale degrees of freedom are either **collapsed** (stable but repetitive) or **fabricated** (detailed but inconsistent).



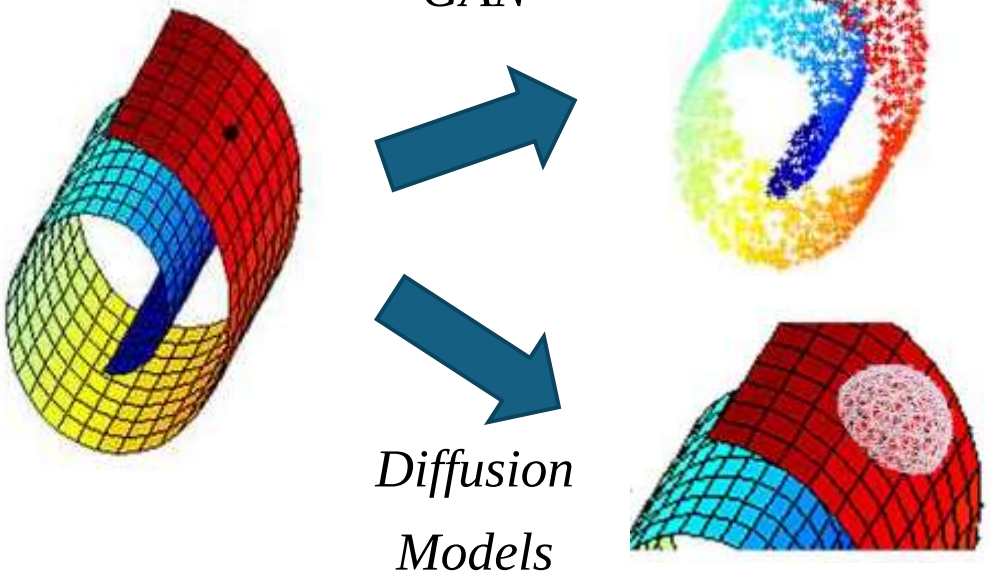
Impossible Triangles





Improving Diffusion Models as an Alternative To GAN, Part 1 | NVIDIA Technical Blog

Diffusion Models' Benefit



Distributional loss

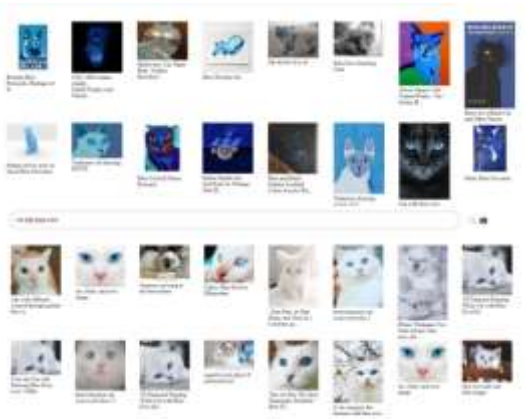
+ Clearer images

- Mode collapse, limited diversity

Point-wise loss

+ Maintain diversity

- Blur images (+ Multi-step)



Diffusion models can be trained on large datasets with

5 billion image-text pairs !

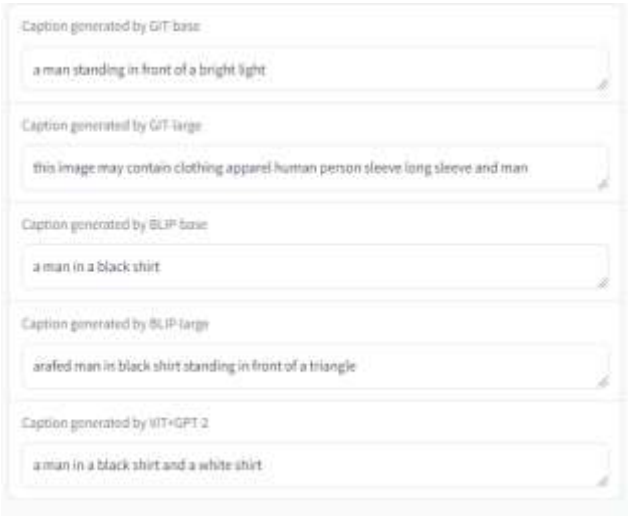
+ High quality and diversity, w/ speedup (ODE, LDM,…)

An unified model!

Rombach & Blattmann, et al. 2022

Diffusion Models' Benefit

SD: "A photo of Musk"



Caption: "A man in a black shirt"

SD: "A painting of Picasso"

Caption: "a painting of a man with a mustache"

+ Mode coverage

+ Generate images for different concepts (instances, persons, styles…)

Are Diffusion Models All You Need?

https://www.midjourney.com/showcase

+ Amazing fidelity

- Cherry picking

Prompt: pouring water from a teapot into a cup; Models: Stable Diffusion Series.

Are Diffusion Models All You Need?

+ High quality, w/ speedup (Flow-Matching, CM)

- Controlability

Are Diffusion Models All You Need?

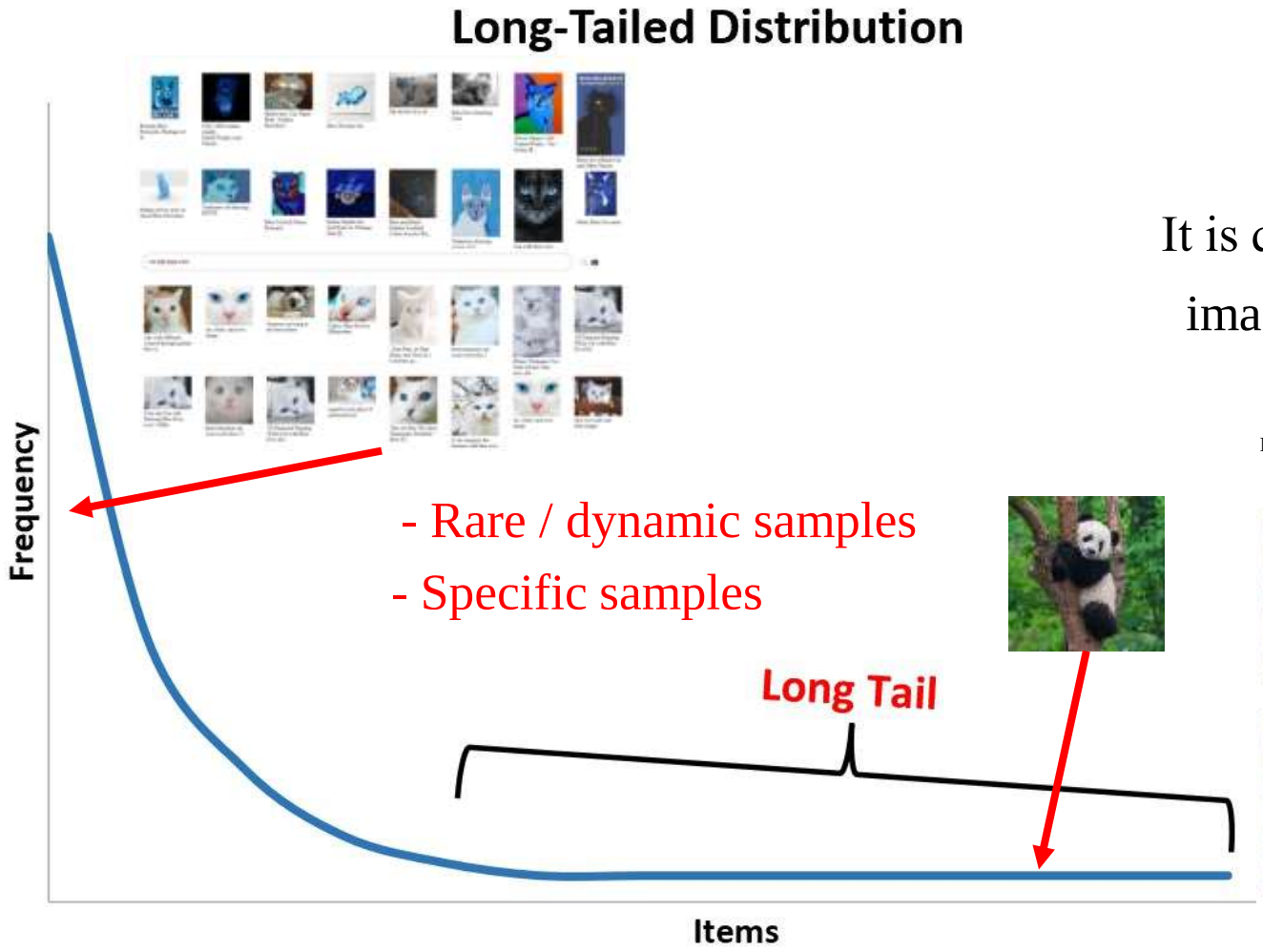
+ Amazing fidelity

https://openai.com/sora

- Texture

- Physic systems

Are Diffusion Models All You Need?

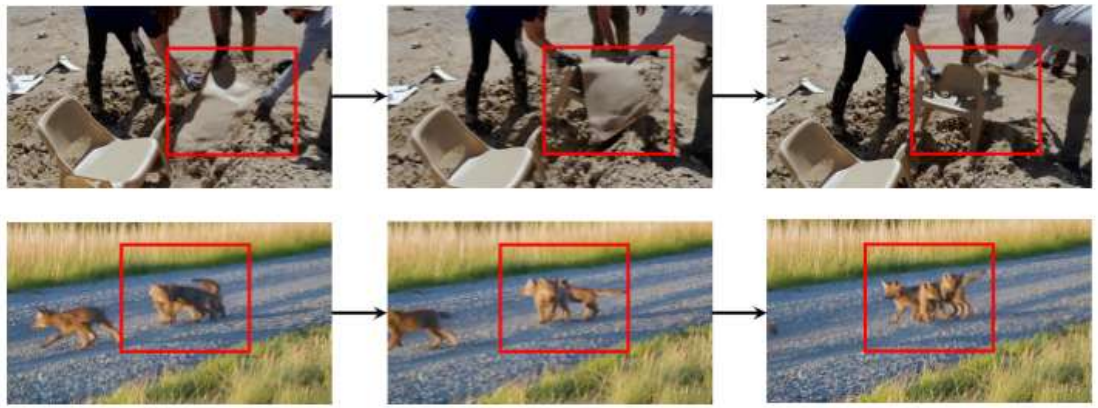


Training Subset	Testing Subset		
	DMs	GANs	Others
DMs	**99.7/99.9/99.9**	86.4/95.9/89.8	79.3/93.5/84.5
GANs	77.1/83.2/74.3	**98.1/99.0/99.6**	91.4/97.2/94.6
Others	76.4/77.2/70.1	82.5/96.0/91.2	**99.6/99.9/99.9**

It is quite easy to detect Diffusion Models' generated images. (But not that good to generalize to GANs)

WildFake: A Large-scale Challenging Dataset for AI-Generated Images Detection, AAAI 2025

DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark, SCIS 2026



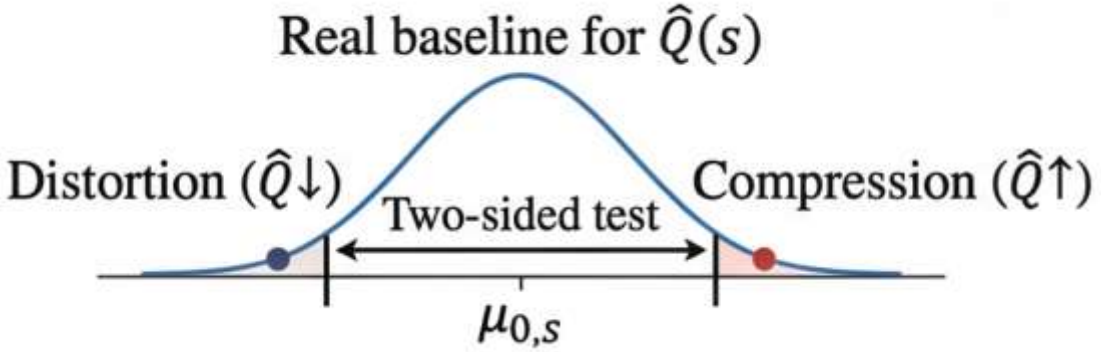
Are Diffusion Models All You Need?

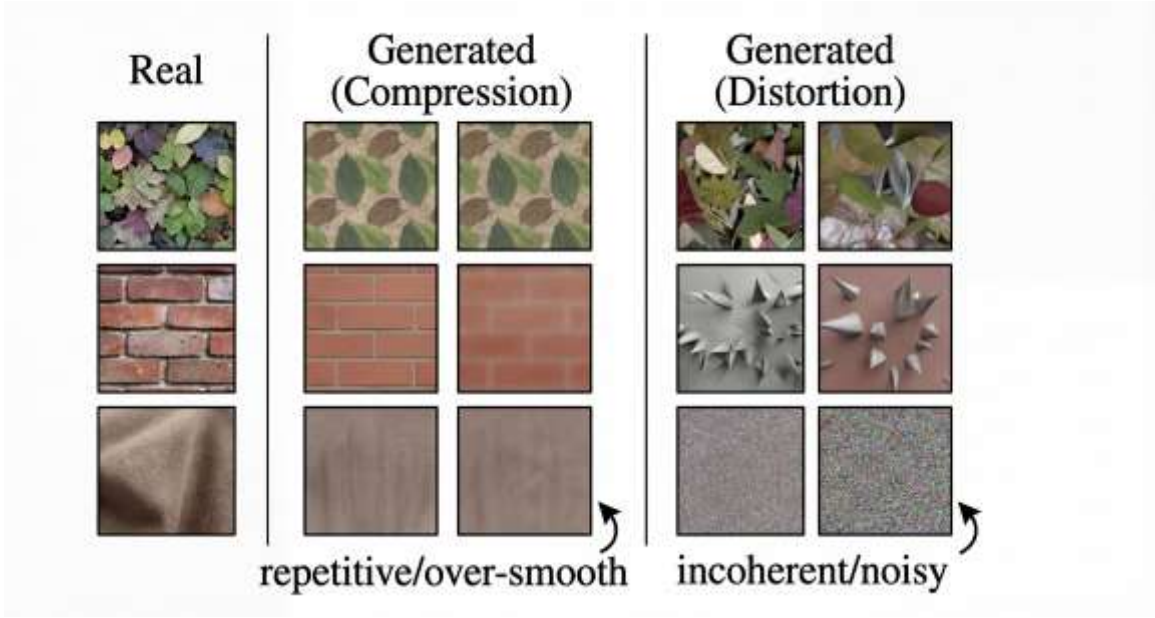
- Comprehensive
- Cross-Modality
- Activity

Detecting generated images with effective dimensions

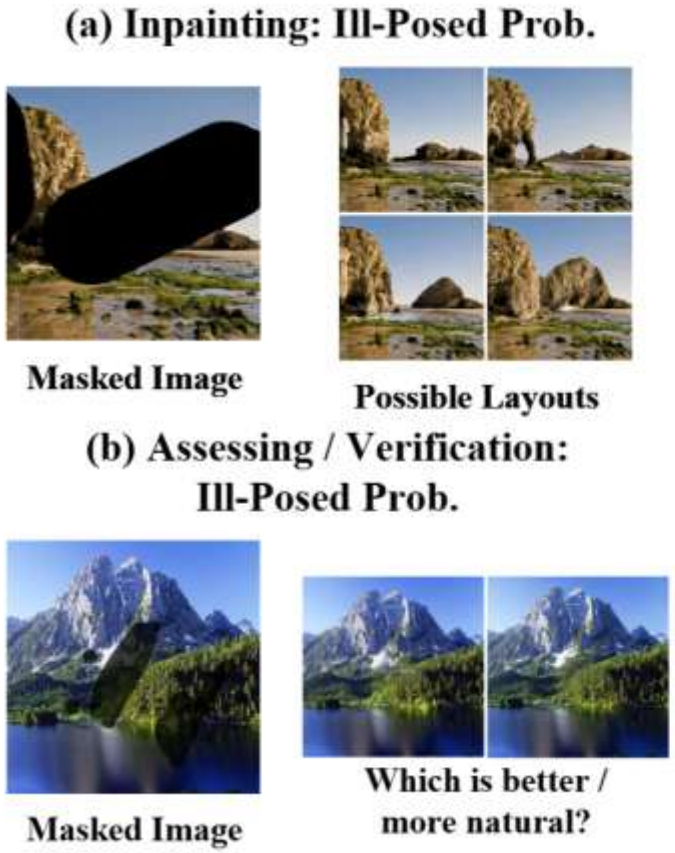
- At scale s, define the second moment: $\mathbf{M}_s = \mathbb{E}[\mathbf{z}(s)\mathbf{z}(s)^T]$, $\mathrm{tr}(\mathbf{M}_s)=1$
- Effective dimension proxy: $d_{\mathrm{eff}}(s) \approx 1/\hat{Q}_s$.

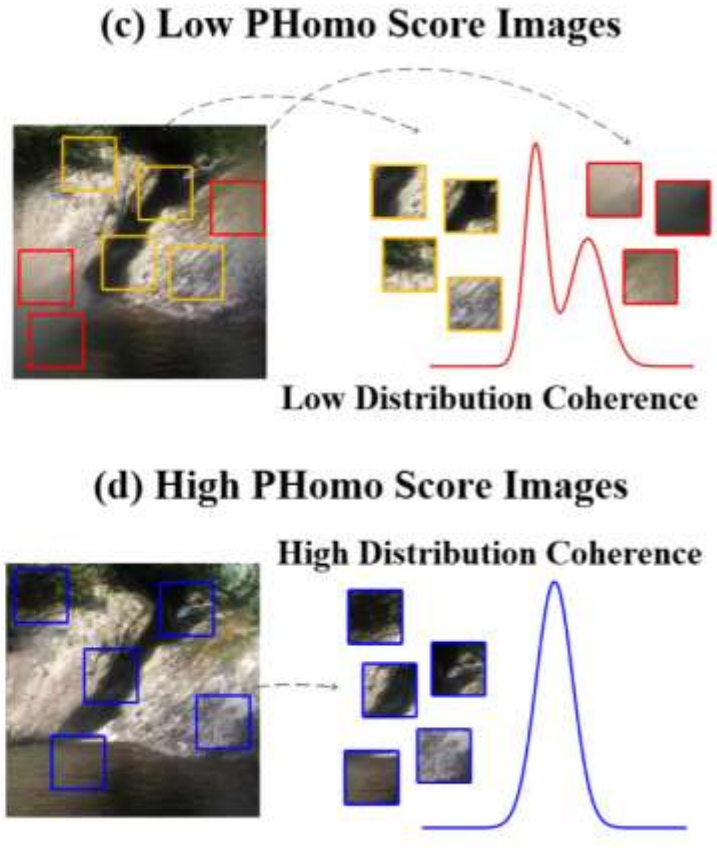
$$\hat{Q}_s = \frac{1}{N(N-1)} \sum_{i \neq j} (\mathbf{z}_i^T \mathbf{z}_j)^2 \approx \mathrm{tr}(\mathbf{M}_s^2) = \sum_k \lambda_k^2.$$

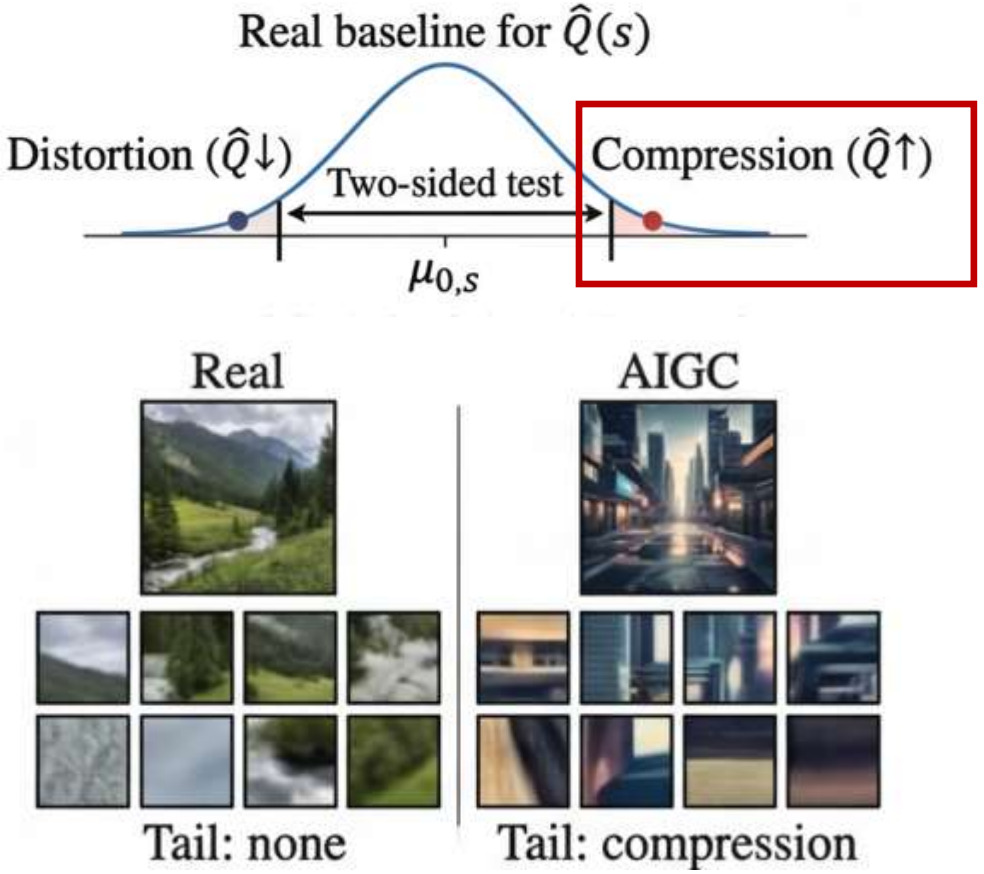




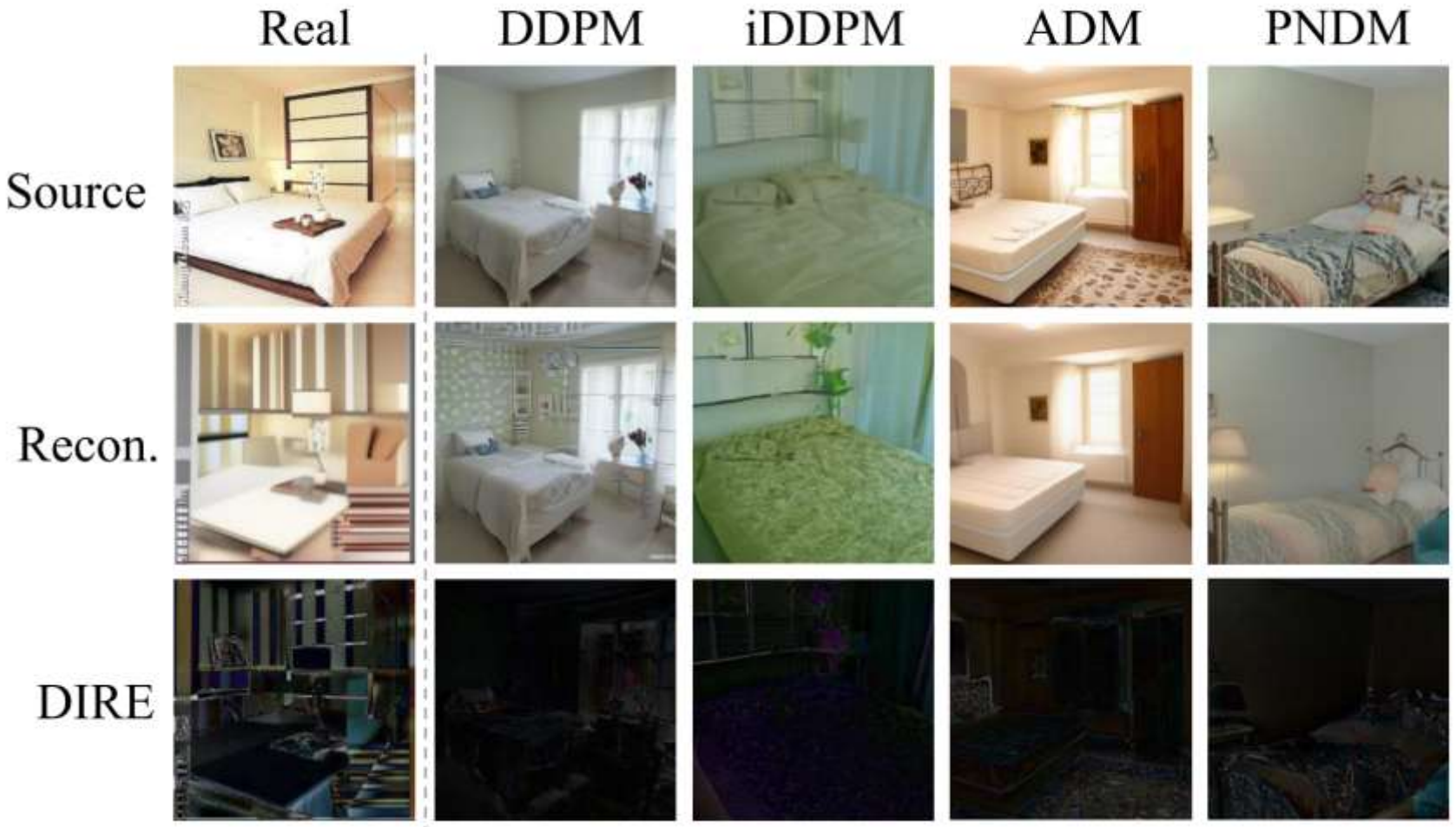
Detecting generated images with effective dimensions





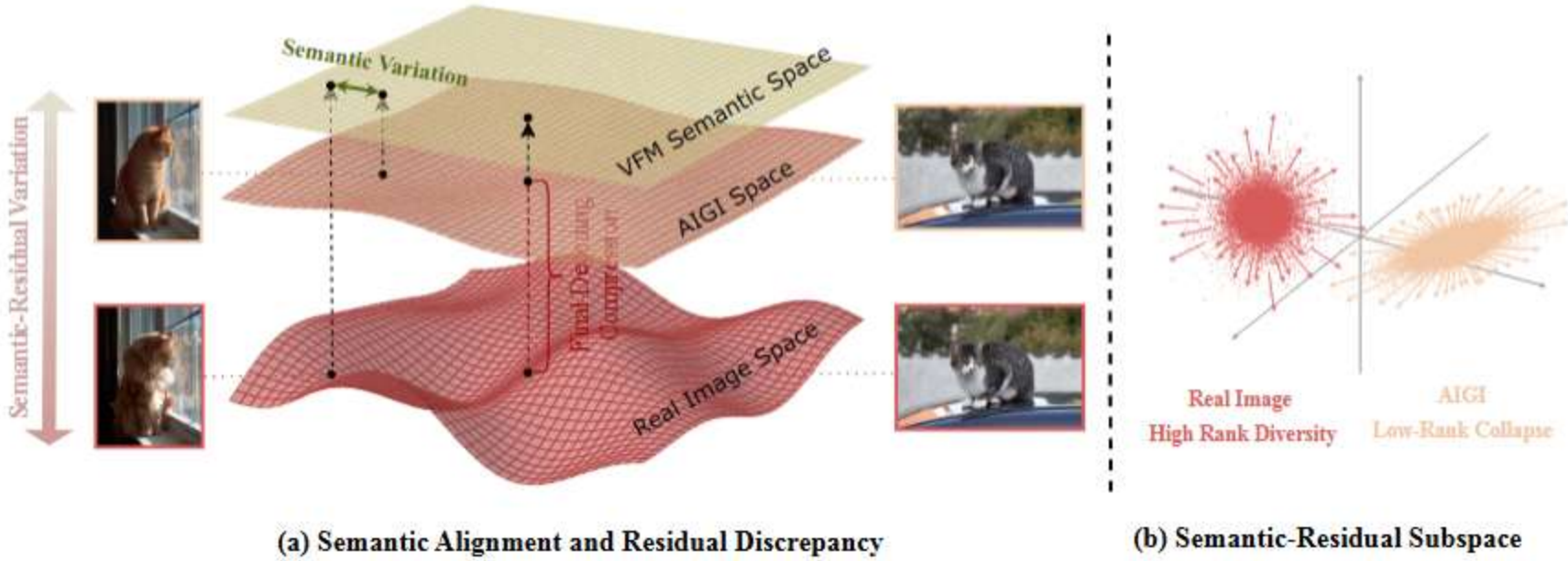


Detecting generated images with effective dimensions



https://arxiv.org/pdf/2303.09295

Detecting generated images with effective dimensions

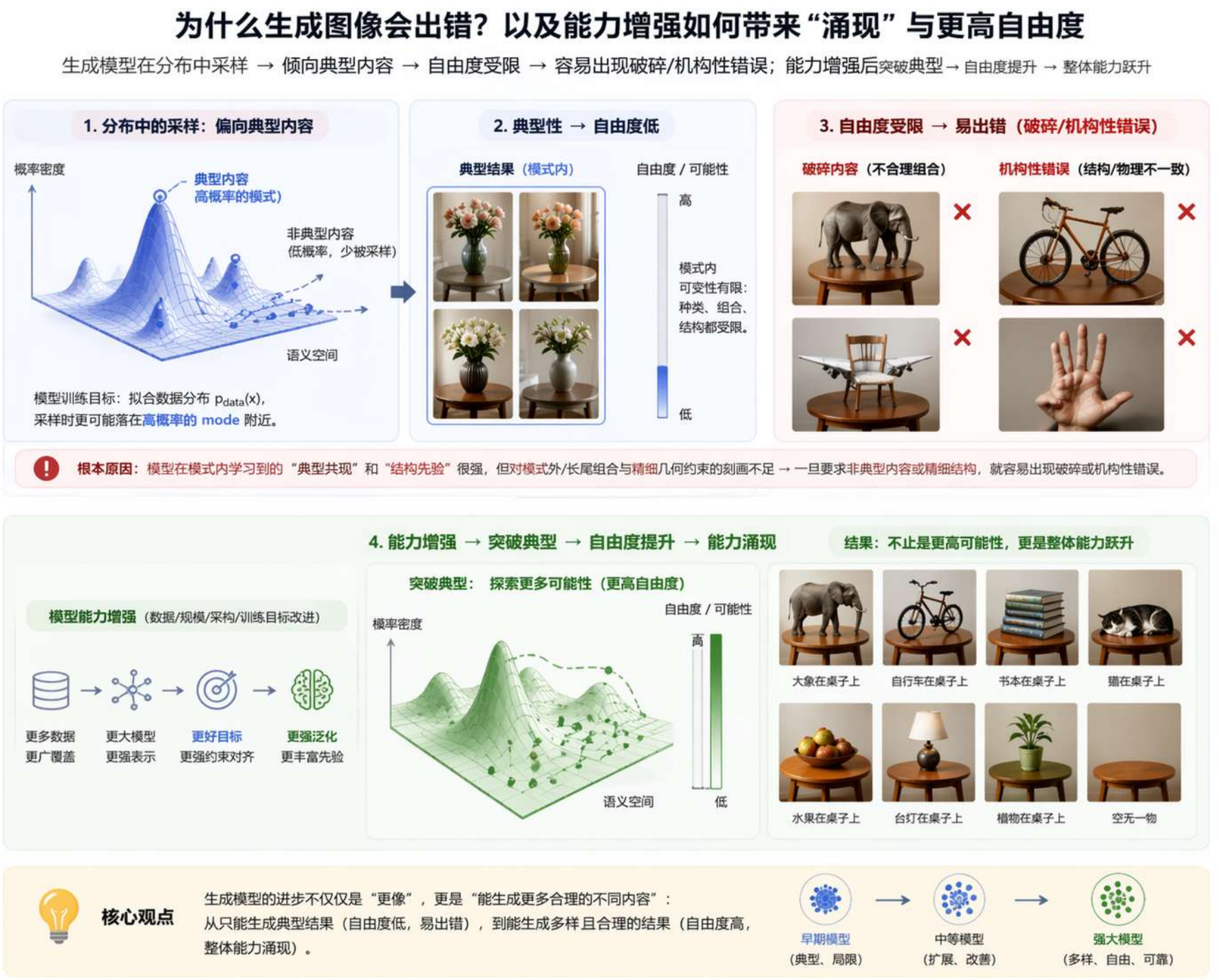


(a) Semantic Alignment and Residual Discrepancy

(b) Semantic-Residual Subspace

“Hallucination”
“Emergence”

“Hallucination”
“Emergence”



“Hallucination”
“Emergence”

“由 ChatGPT 内置图像生成工具根据文本提示生成。”


阿里千问全员会
阿里千问全员会

AIGI Detection with Reliable Explanations

<think>
经过仔细观察，可以发现：
<bbox>… 项圈左右两端断开</bbox>
<bbox>… 光泽异常</bbox>
整体来看，这张图像具有<tag>材质错误</tag>。
</think>
综上所述，这张图像是 AI 生成的。

Towards Explainable Fake Image Detection with Multi-Modal Large Language Models, ACM MM 25
FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning, ICLR 2026
Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images, CVPR 2026

AIGI Detection with Reliable Explanations

<think>
经过仔细观察，可以发现：
<bbox>… 轿车在火车铁轨上</bbox>
<bbox>… 车顶灯处文字模糊</bbox>
<bbox>… 播报屏幕上文字模糊</bbox>
整体来看，这张图像具有<tag>*属性错误*</tag>。
</think>
综上所述，这张图像是 AI 生成的。

Towards Explainable Fake Image Detection with Multi-Modal Large Language Models, ACM MM 25
FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning, ICLR 2026
Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images, CVPR 2026

AIGI Detection with Reliable Explanations

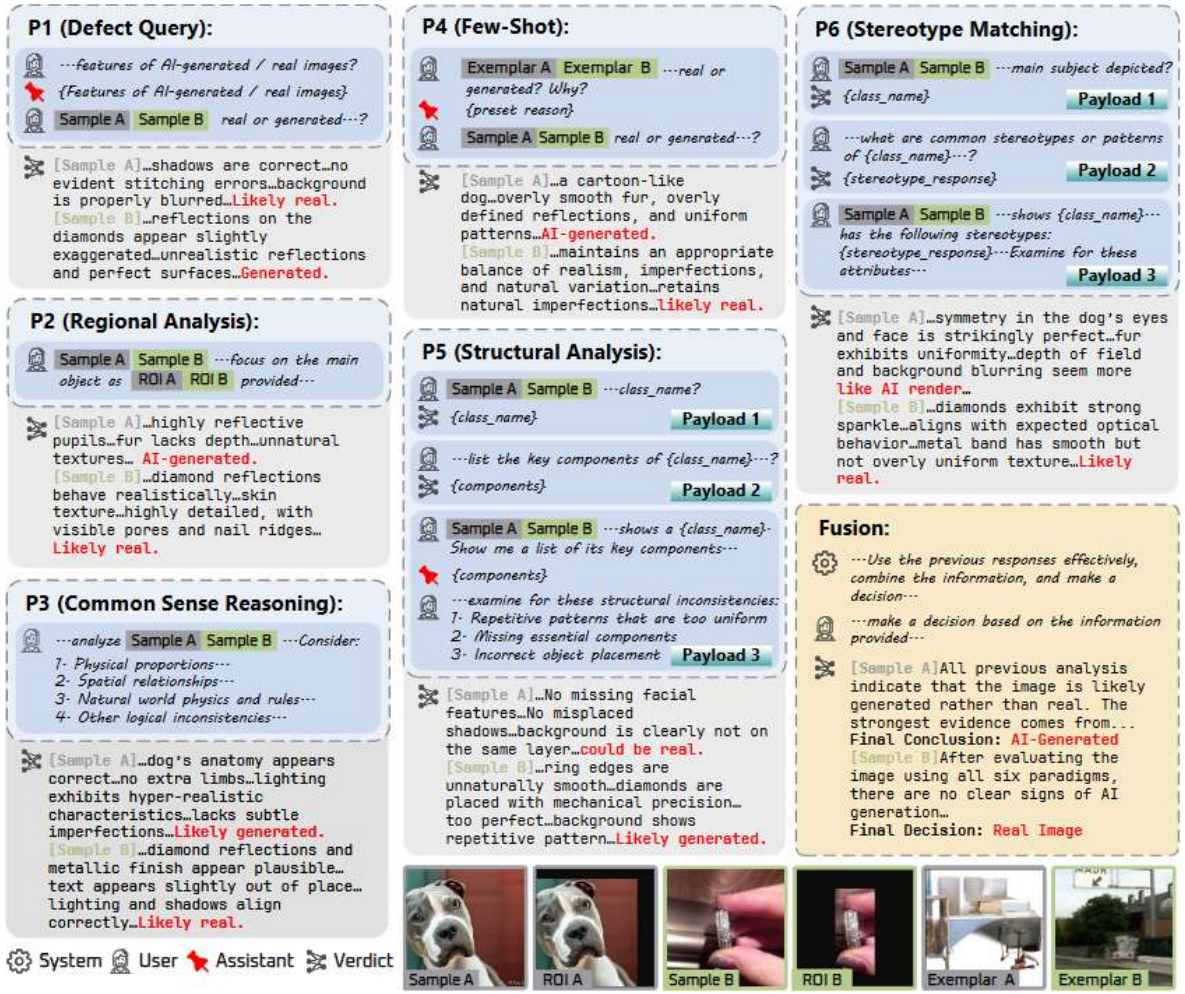
<think>
经过仔细观察，可以发现：
<bbox>… 铁塔下部火焰缺乏自然火焰混沌性，材质光滑</bbox>
整体来看，这张图像具有<tag>材质错误</tag>。
</think>
综上所述，这张图像是 AI 生成的。

Towards Explainable Fake Image Detection with Multi-Modal Large Language Models, ACM MM 25
FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning, ICLR 2026
Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images, CVPR 2026

AIGI Detection with Reliable Explanations

Towards Explainable Fake Image Detection with Multi-Modal Large Language Models, ACM MM 25
FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning, ICLR 2026
Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images, CVPR 2026

AIGI Detection with Reliable Explanations

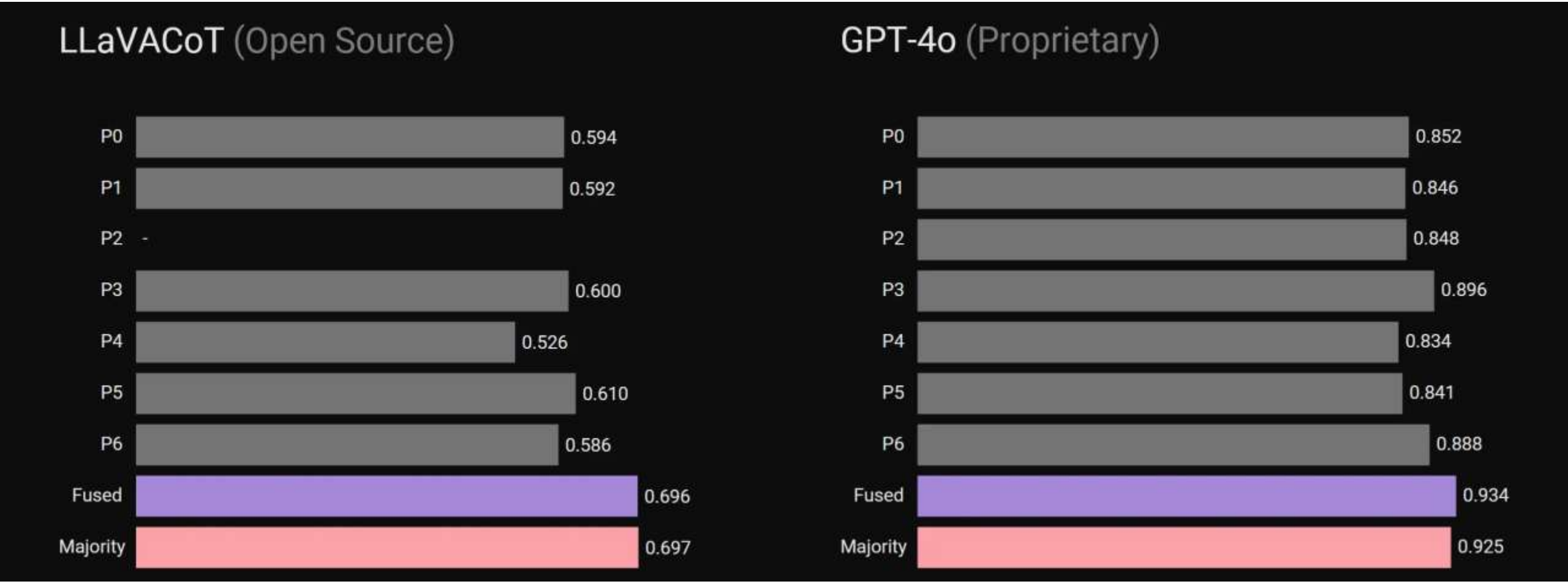


On the detection of AI-generated images, we notice that:
1.Using **one prompt is not enough** - MLLMs are not optimized for this specific type of problems.
2.The most decisive defects can give AI-generated images away easily. Humans can tell an image is AI-generated just by one or two aspects.

3.Individual prompts may hallucinate or miss subtle cues. To address this, we execute P1–P6 in parallel. A final fusion stage aggregates these diverse perspectives—balancing detailed inspection with logical reasoning—to produce a robust, explainable verdict.

Towards Explainable Fake Image Detection with Multi-Modal Large Language Models, ACM MM 25

AIGI Detection with Reliable Explanations



Towards Explainable Fake Image Detection with Multi-Modal Large Language Models, ACM MM 25

Dataset Construction **Fine-tuning MLLM**

FakeXplainer

Trained on the dataset with

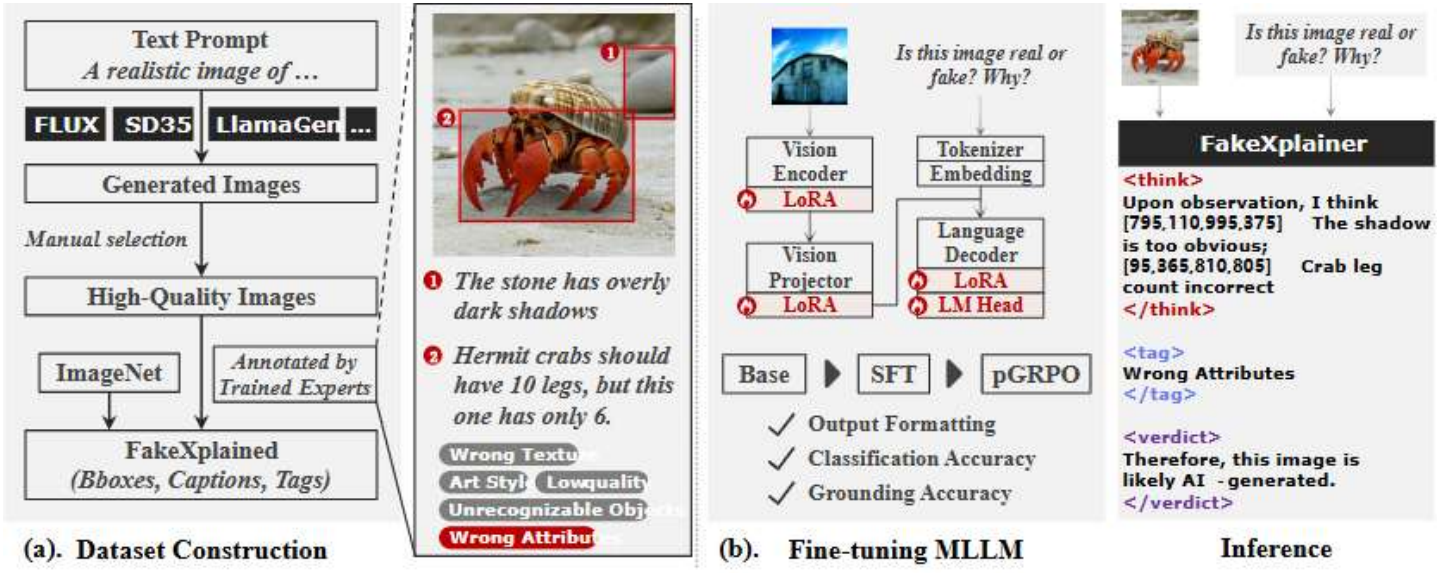
SFT + *Progressive* GRPO

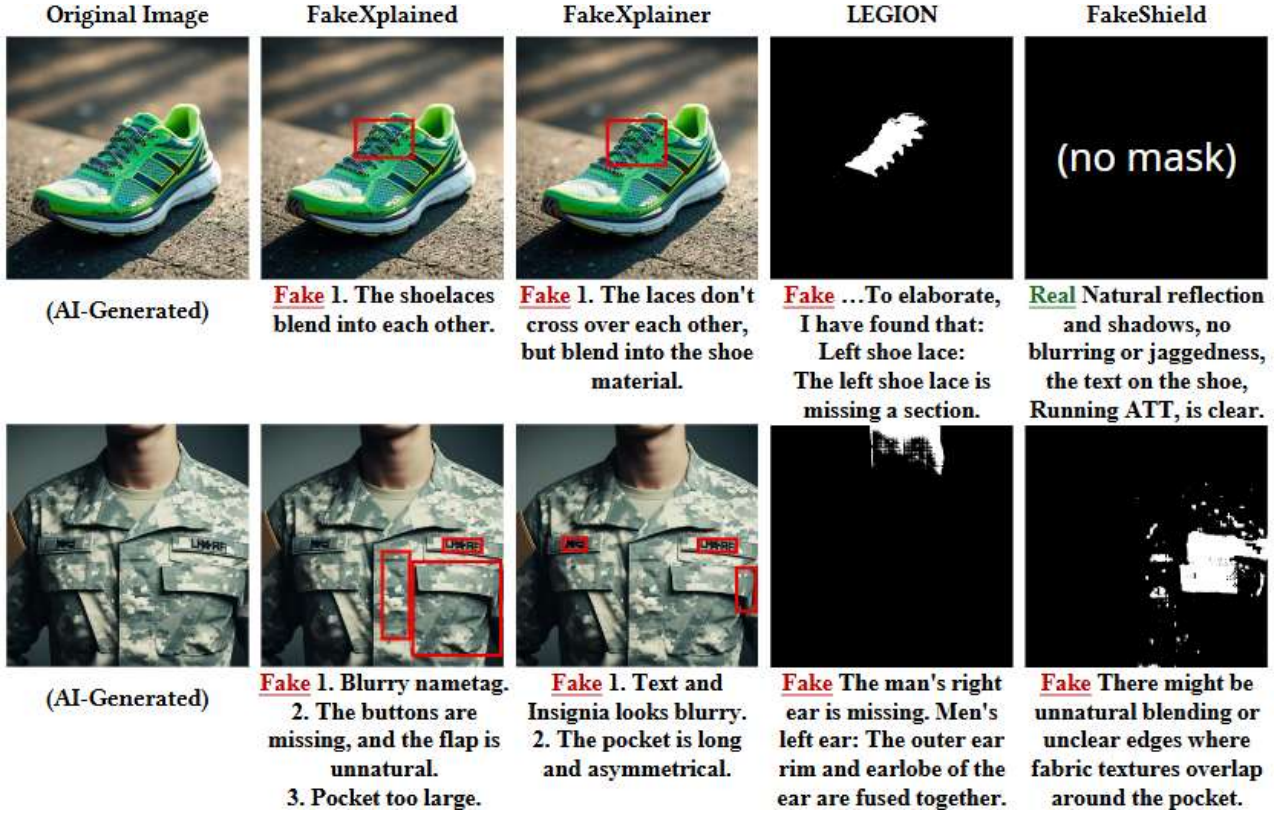
based on *Qwen-2.5-VL-32B-Instruct*

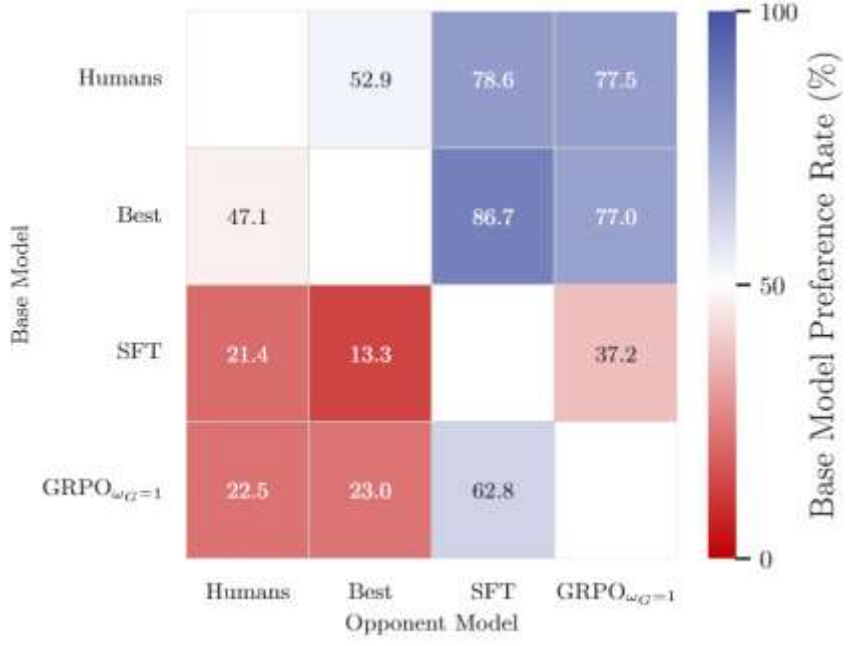
AIGI Detection with Reliable Explanations

FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning, ICLR 2026

AIGI Detection with Reliable Explanations







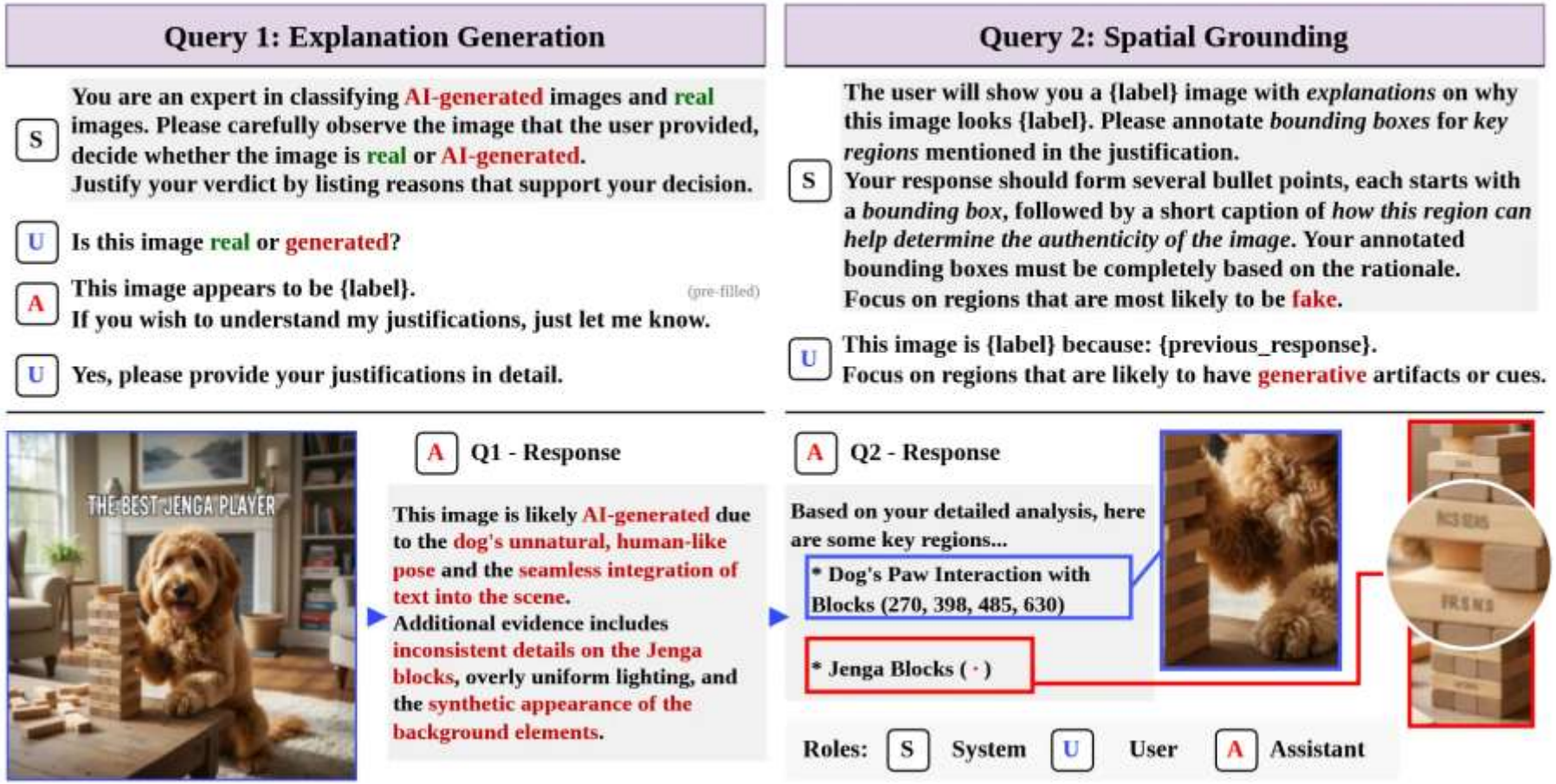
FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning, ICLR 2026

Generated by FLUX.1-Schnell

AIGI Detection with Reliable Explanations

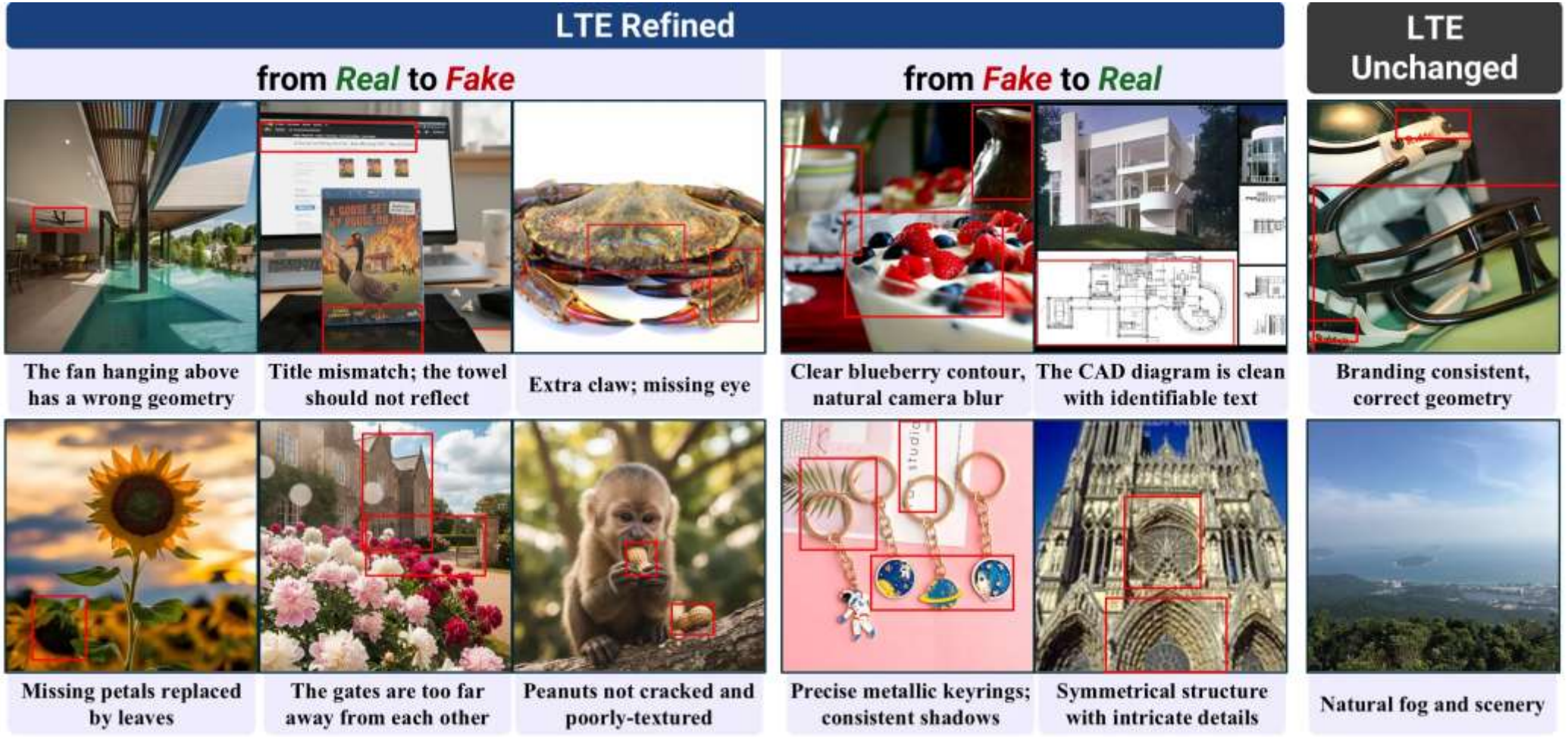
Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images, CVPR 2026

AIGI Detection with Reliable Explanations



Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images, CVPR 2026

AIGI Detection with Reliable Explanations

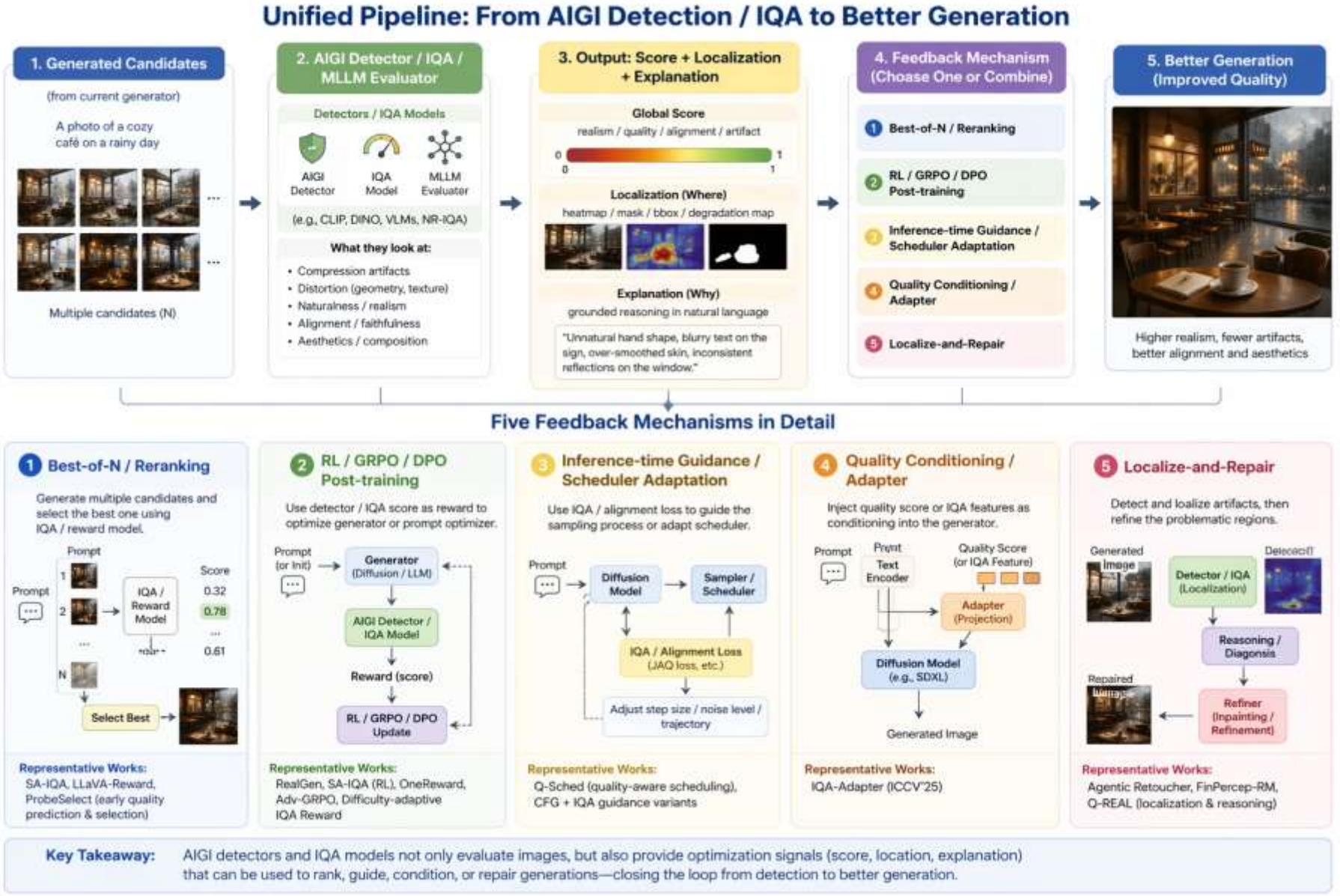


Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images, CVPR 2026

Re-Edit


图像质量的两种度量：真实性 vs. 真实度
我们可以从两个不同的维度来评价一张图像的质量，并分别用于优化生成模型。
1. 真实性（真/假）
图像是否真实存在，只有真假两类。
定义
• 如果图像由生成模型输出，则判定为“假”；
• 否则为“真”。
判断方式
生成模型输出
真实世界图像
二分类判别器 (Discriminator)
假 (Fake)
真 (Real)
优化方法
可通过 GAN 框架进行优化。
生成器 (Generator)
生成图像
判别器 (Discriminator)
真/假
对抗训练，不断提升生成器以欺骗判别器
存在的问题
• 定义边界模糊：对于后处理、编辑、超分、风格迁移、自编码器重构等结果，难以明确界定是否属于“假”。
• 容易被对抗样本欺骗：生成模型可能学会“骗过判别器”，但不等于更真实。
原始真实图像
后处理/调色
超分/重建
自编码器重构
这些图像算真还是假?
2. 真实度（真假的程度）
图像有多像真实世界，是一个连续的、相对的度量。
定义
衡量图像与真实世界分布的接近程度，通常是相对比较的结果，而不是绝对的“真/假”。
判断方式
通常由人类主观打分，或使用学习到的打分模型（如 CLIP-IQA、ManIQA 等）。
真实度得分（相对）
0
1
非常不真实 (0.1)
较不真实 (0.3)
中等真实 (0.5)
较真实 (0.7)
非常真实 (0.9)
优化方法
可通过强化学习（RL）或基于评分模型的优化来提升真实度。
生成器 (Generator)
生成图像
评分模型 (Reward Model)
真实度得分
强化学习优化 (PPO 等)
根据得分优化生成器，提升真实度
存在的问题
• 主观性强：不同人、不同文化、不同场景下，对“真实”的标准不同。
• 评分模型可能偏差：训练数据、模型偏好会影响评分结果。
• 相对比较的局限：只能告诉你“更真实多少”，无法绝对定义“真实”。
同一图像，不同人打分可能差异很大
0.85
0.60
0.90
对比总结
真实性（真/假）
• 优点：判断明确，易于建模和训练。
• 缺点：定义边界模糊，容易被对抗攻击。
真实度（真假的程度）
• 优点：更细粒度，能反映质量差异，指导生成优化。
• 缺点：主观性强，评分不一致，难以绝对定义。
结论
真实性解决“是不是假的”问题，
真实度解决“有多像真的”问题。
二者互补，共同推动生成模型的进步。

Re-Edit


Unified Pipeline: From AIGI Detection / IQA to Better Generation
1. Generated Candidates
(from current generator)
A photo of a cozy café on a rainy day
Multiple candidates (N)
2. AIGI Detector / IQA / MLLM Evaluator
Detectors / IQA Models
AIGI Detector
IQA Model
MLLM Evaluater
(e.g., CLIP, DINO, VLMs, NR-IQA)
What they look at:
• Compression artifacts
• Distortion (geometry, texture)
• Naturalness / realism
• Alignment / faithfulness
• Aesthetics / composition
3. Output: Score + Localization + Explanation
Global Score
realism / quality / alignment / artifact
Localization (Where)
heatmap / mask / bbox / degradation map
Explanation (Why)
grounded reasoning in natural language
"Unnatural hand shape, blurry text on the sign, over-smoothed skin, inconsistent reflections on the window."
4. Feedback Mechanism (Choase One or Combine)
1 Best-of-N / Reranking
2 RL / GRPO / DPO Post-training
3 Inference-time Guidance / Scheduler Adaptation
4 Quality Conditioning / Adapter
5 Localize-and-Repair
5. Better Generation (Improved Quality)
Higher realism, fewer artifacts, better alignment and aesthetics
Five Feedback Mechanisms in Detail
1 Best-of-N / Reranking
Generate multiple candidates and select the best one using IQA / reward model.
Prompt
IQA / Reward Model
Score
0.32
0.78
0.61
Select Best
Representative Works:
SA-IQA, LLaVA-Reward, ProbeSelect (early quality prediction & selection)
2 RL / GRPO / DPO Post-training
Use detector / IQA score as reward to optimize generator or prompt optimizer.
Prompt (or Init)
Generator (Diffusion / LLM)
AIGI Detector / IQA Model
Reward (score)
RL / GRPO / DPO Update
Representative Works:
RealGen, SA-IQA (RL), OneReward, Adv-GRPO, Difficulty-adaptive IQA Reward
3 Inference-time Guidance / Scheduler Adaptation
Use IQA / alignment loss to guide the sampling process or adapt scheduler.
Prompt
Diffusion Model
Sampler / Scheduler
IQA / Alignment Loss (JAQ loss, etc.)
Adjust step size / noise level / trajectory
Representative Works:
Q-Sched (quality-aware scheduling), CFG + IQA guidance variants
4 Quality Conditioning / Adapter
Inject quality score or IQA features as conditioning into the generator.
Prompt
Text Encoder
Quality Score (or IQA Feature)
Adapter (Projection)
Diffusion Model (e.g., SDXL)
Generated Image
Representative Works:
IQA-Adapter (ICCV'25)
5 Localize-and-Repair
Detect and localize artifacts, then refine the problematic regions.
Generated Image
Detector / IQA (Localization)
Reasoning / Diagonsis
Refiner (Inpainting / Refinement)
Repaired Image
Representative Works:
Agentic Retoucher, FinPercep-RM, Q-REAL (localization & reasoning)
Key Takeaway:
AIGI detectors and IQA models not only evaluate images, but also provide optimization signals (score, location, explanation) that can be used to rank, guide, condition, or repair generations—closing the loop from detection to better generation.

Video & Music

为什么视频生成更困难？ —— 需要更高能力，自由度更低，协调更复杂

1. 需要建模时空连续性（能力要求更高）

不仅要生成“是什么”，还要准确记录：

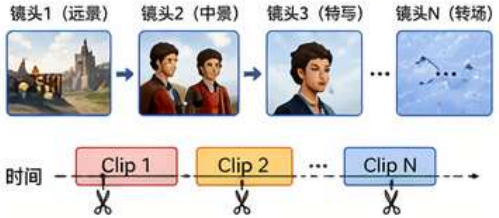
- ✓ 物体的运动轨迹
- ✓ 速度与加速度变化
- ✓ 形态与状态的演化
- ✓ 物理交互与因果关系

理论上需要建模的变量和约束远多于图像，
所以自由度更低，出错更易累积放大。

2. 需要分镜与叙事结构（协调更复杂）

一个长视频由多个 clip 组成，涉及：

- ✓ 镜头语言设计（景别、角度、运动）
- ✓ 时间线与节奏安排
- ✓ 场景/角色的一致性保持
- ✓ 剪辑点的合理衔接



每个clip之间如何协调一致非常困难，
一旦衔接不当，观感会明显割裂。

3. 受硬件与计算限制（实现更困难）

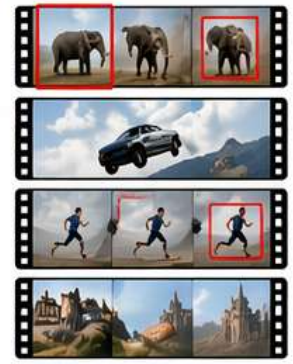
视频 = 大量图像帧的时序生成与存储

高分辨率、长时长、高帧率的视频生成，
对硬件资源要求极高，成本昂贵。

4. 综合结果：自由度更低，错误更难控制

⚠ 视频生成的典型问题

- ✕ 物体凭空出现/消失
- ✕ 运动不合理（漂浮、穿模）
- ✕ 速度突变或不连续
- ✕ 长时一致性崩溃
- ✕ 镜头切换生硬
- ✕ 因果关系错误

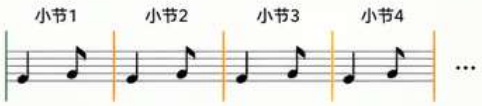


因此，视频生成在当前阶段是：
高难度 + 低自由度 + 高成本 + 高风险

为什么音乐更适合作为生成模型的创作载体？ —— 结构天然友好，可叠加，更高自由度

1. 有小节/段落结构，天然边界清晰

音乐通常以小节（bar）为基本单位组织。



小节之间相对独立，局部生成错误
不会破坏整体结构，修改和替换更容易。

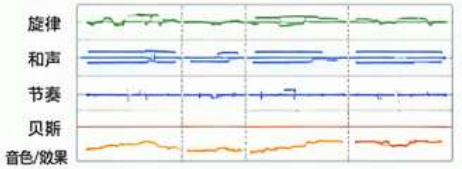
2. 小节之间相互影响小，易于拼接与扩展

只要和声/节奏/风格匹配，就可以自然衔接。

可按需生成、替换、插入、循环，
组合出更长的作品，协调用由度低。

3. 音乐可多轨叠加，信息密度高

旋律、和声、节奏、音色可在不同轨道并行叠加。



不同轨道解耦，模型更容易学习和生成，
整体自由度更高，表现力更丰富。

4. 评估更灵活，主观空间更大

音乐的“好听”有更多风格化空间，
不像视频那样强依赖物理真实与一致性。

因此，音乐生成更容易获得多样且
合理的结果，模型更容易发挥创造力。

视频生成（当前阶段）

- ✕ 时空建模复杂，约束多
- ✕ 镜头与叙事协调困难
- ✕ 硬件成本高，训练推理都贵
- ✕ 自由度低，错误难避免

VS.

音乐生成（当前阶段）

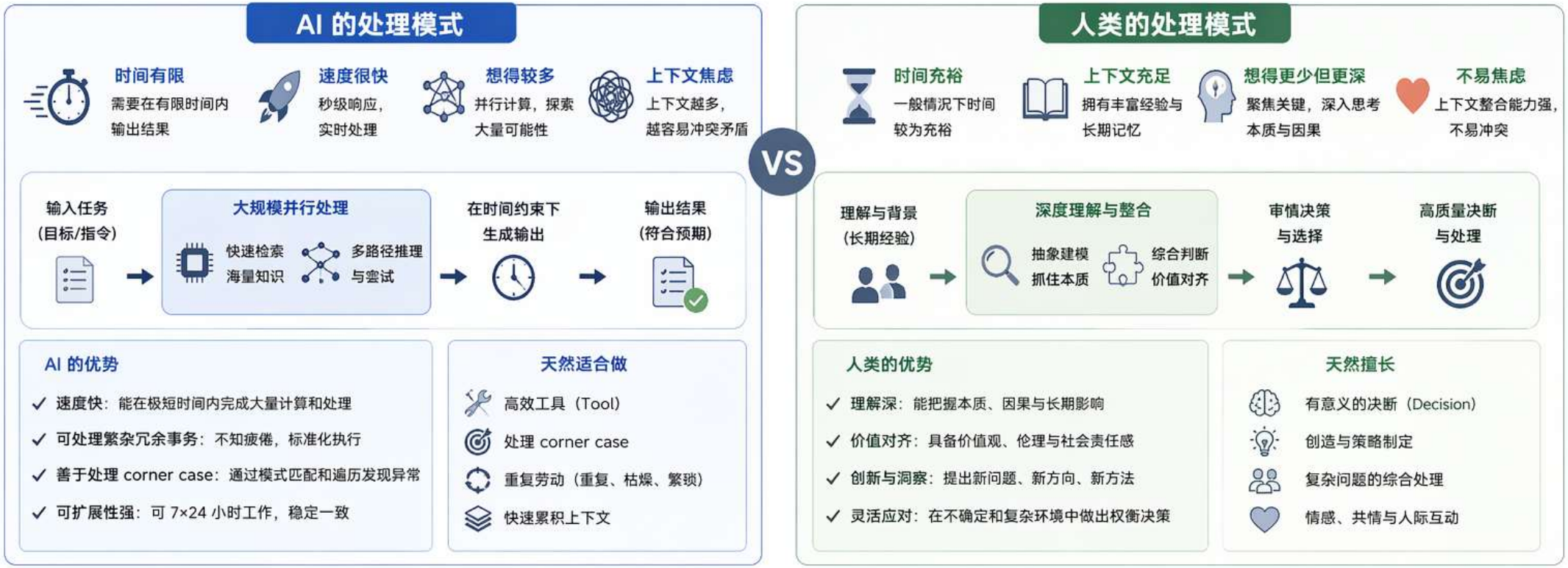
- ✓ 结构清晰（小节/段落）
- ✓ 可拼接、可扩展、易编辑
- ✓ 多轨叠加，信息表达丰富
- ✓ 自由度高，创作空间大

在当前技术与硬件条件下，
音乐是更适合作为生成模型
进行创作与表达的载体；
视频生成仍是更高难度挑战，
需要更强模型与更大算力的支持。

Fin.

AI 与人类：不同的处理模式，互补的最佳分工

AI 和人类处理事情的底层模式不同，但可以形成完美互补：AI 作为高效工具，为人类积累上下文、处理繁杂事务，让人类专注于更有意义的决断。



互补协同：AI 为人类赋能，人类引导 AI

AI 天生适合做“工具”，人类天生适合做“决断”。

让 AI 处理繁杂、重复、边缘的事务，让人类专注于创造、判断与责任，共同放大彼此的能力，创造更大的价值。

用 AI 的速度 + 人类的智慧 = 更高效、更可靠、更有意义的未来

AI 扩展能力（效率） + 人类赋予方向（智慧） = 价值创造（意义）

Thanks!

- Takeaways
  - Real images exhibit a multi-scale complexity profile (not a single fixed-resolution statistic).
  - Generation errors are typically two-sided: compression (over-smooth/repetitive) or distortion (artifact/incoherence).
    - MLLMs are good at detecting compression, traditional models are good at detecting distortion.
  - We need a single, representation-space metric that captures this geometry gap across scales.
    - We can improve generative models by capturing the geometry gap across scales.