

Recap.

The design space of text-to-image generation

Training

- Training paradigm
 - DDPM
 - Flow matching

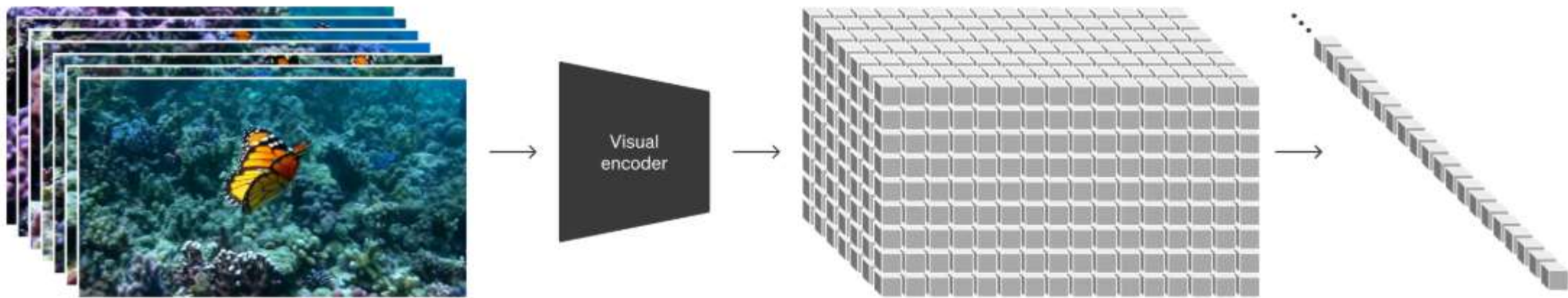
Model

- Latent Space
 - VQ-VAE/VQGAN
 - Advanced VAE
- Model architecture
 - U-Net
 - DiT

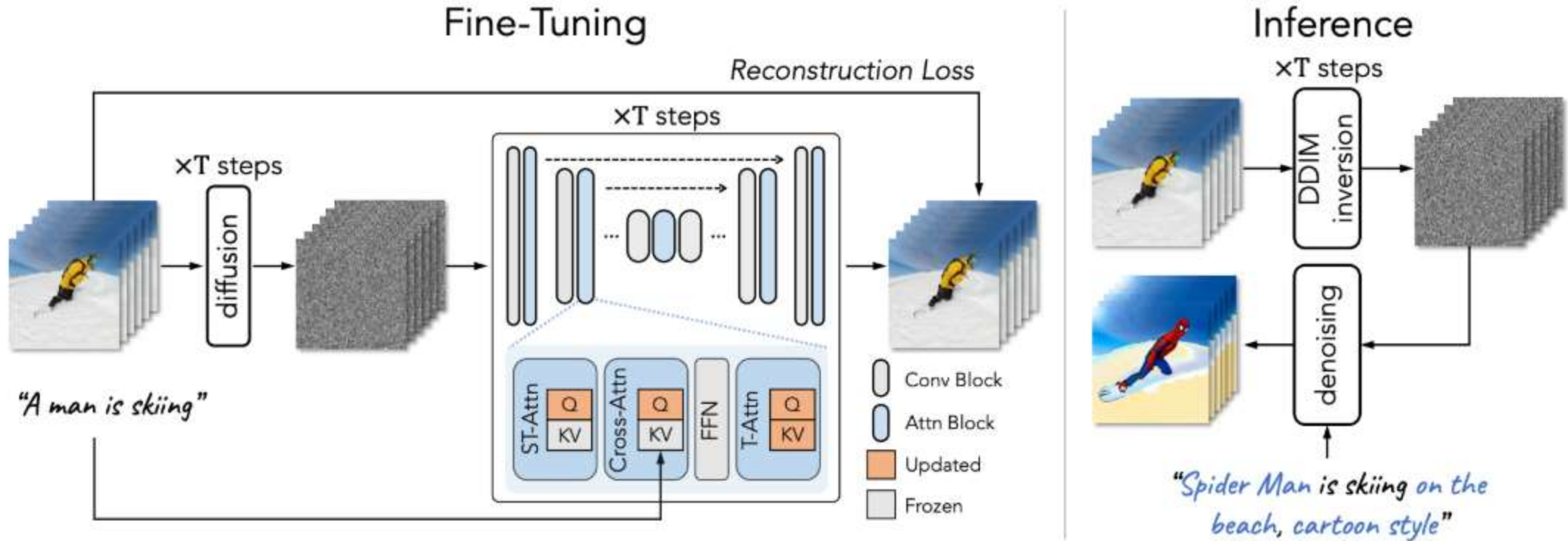
Text Encoding

- Text Encoder
 - CLIP
 - CLIP + T5
 - LLM/VLM/MLLM
- Text Conditioning
 - Cross Attention
 - MM-DiT
 - Native MM

Sora: DiT for Video Generation



Adapting Image Models to Generate Videos



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

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

Adapting Image Models to Generate Videos

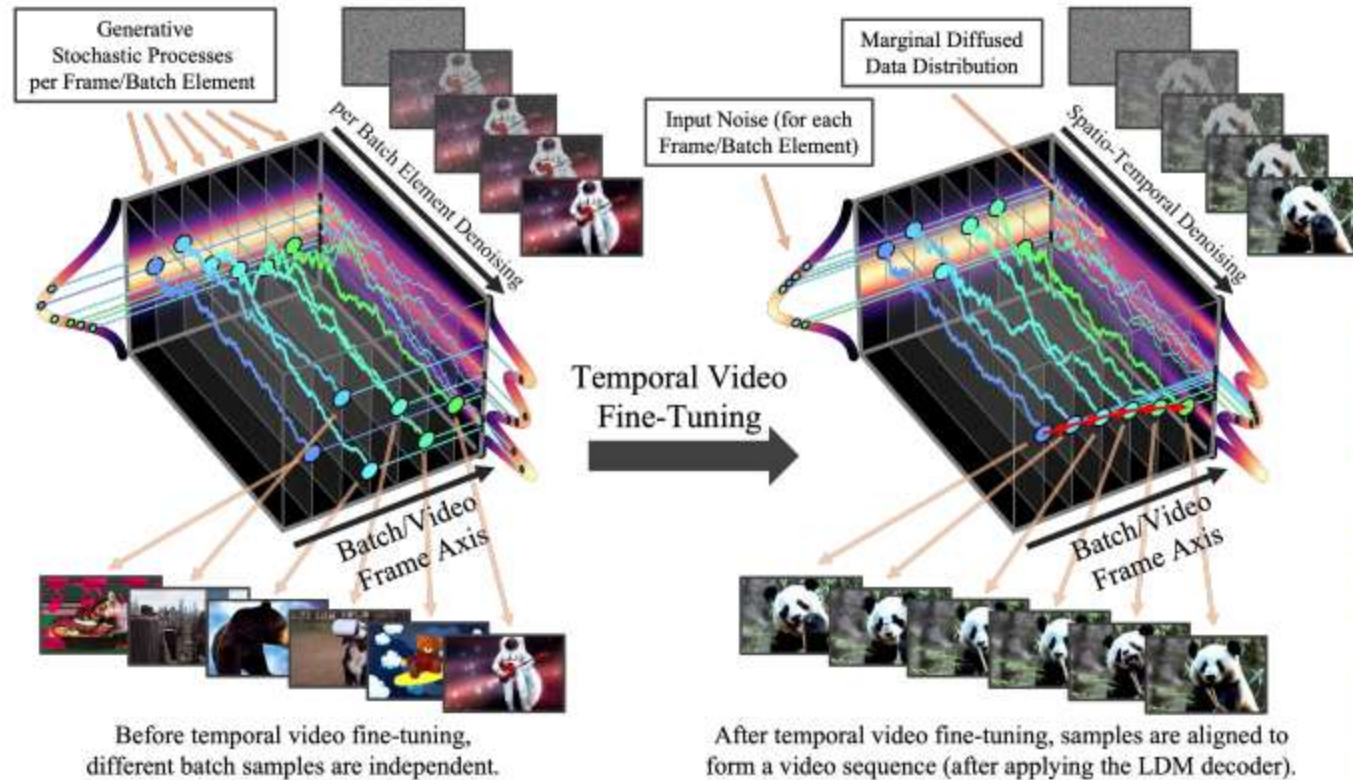


Figure 2. **Temporal Video Fine-Tuning.**

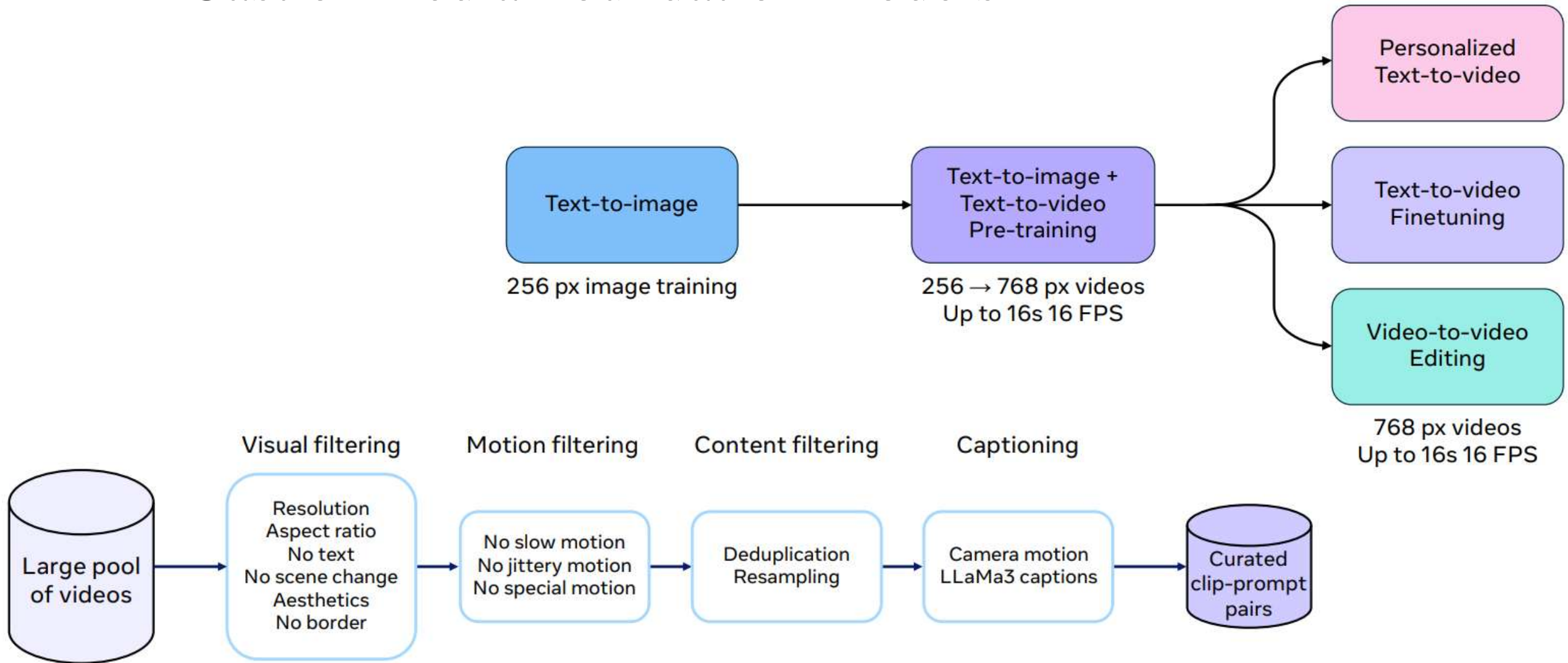
We turn pre-trained image diffusion models into temporally consistent video generators. Initially, different samples of a batch synthesized by the model are independent. After temporal video fine-tuning, the samples are temporally aligned and form coherent videos. The stochastic generation process before 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 up-samplers in pixel or latent space (Sec. 3.4).

Genie: Generative Interactive Environments

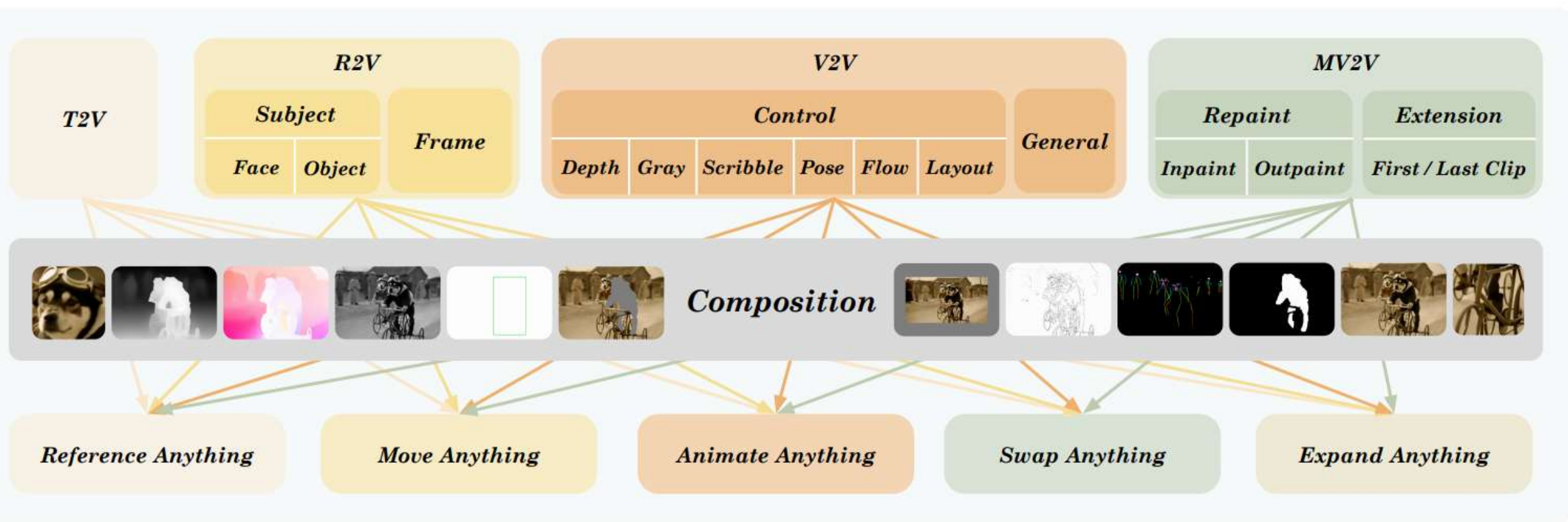


“Remarkably, Genie learns not only which parts of an observation are generally controllable, but also infers diverse latent actions that are consistent across the generated environments. Note here how the same latent actions yield similar behaviors across different prompt images. ”

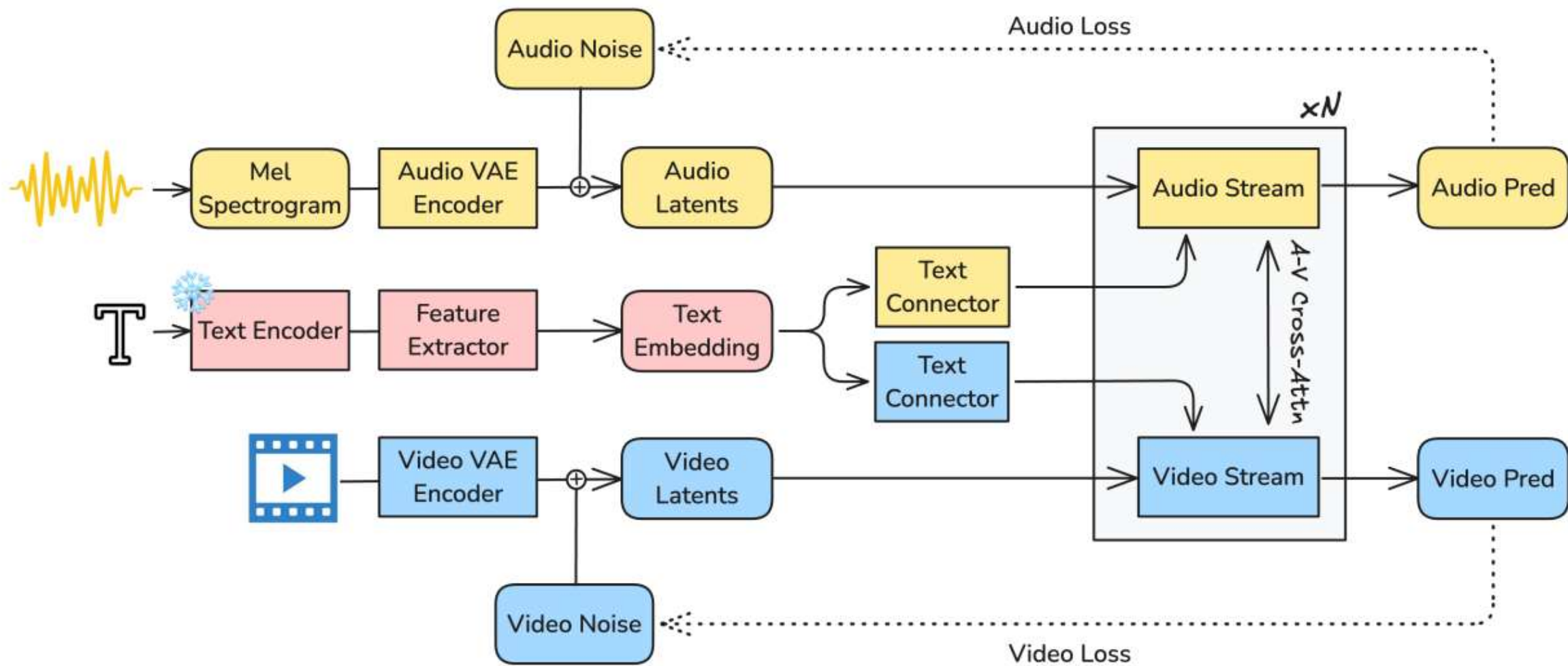
Movie Gen: A Cast of Media Foundation Models



VACE



LTX

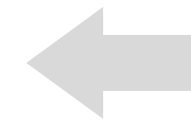


3D Generation

Text-to-3D structure generation



Image-to-3D structure generation



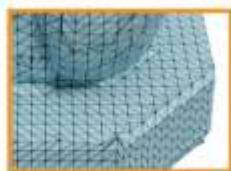
x : generated 3D structures



y : image prompt

Surface-Based Rendering: Mesh

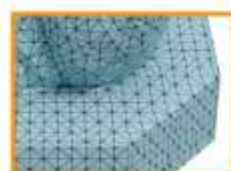
Isosurfacing
w/
gradients



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

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



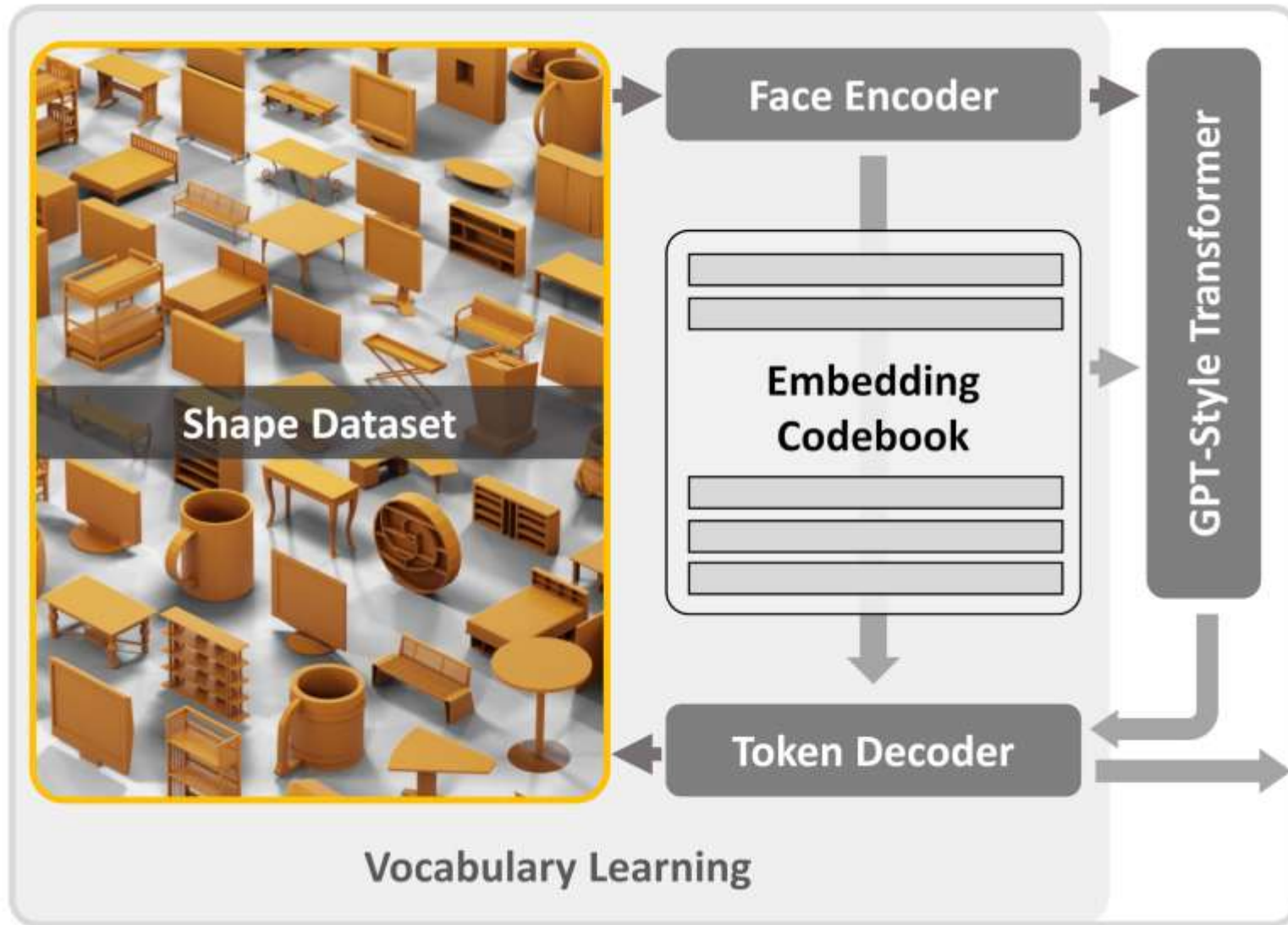
Reference



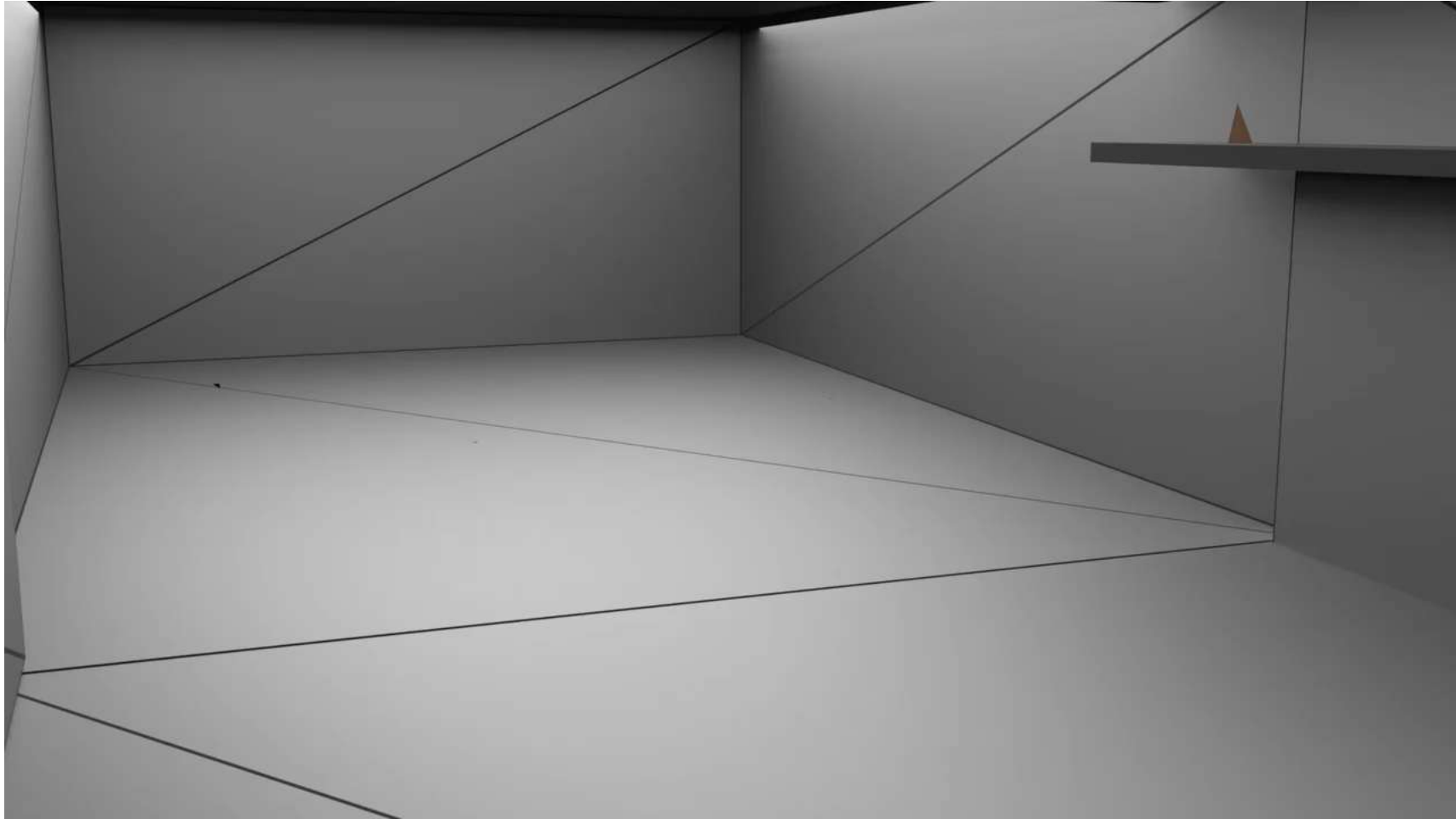
T-pose optimization

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

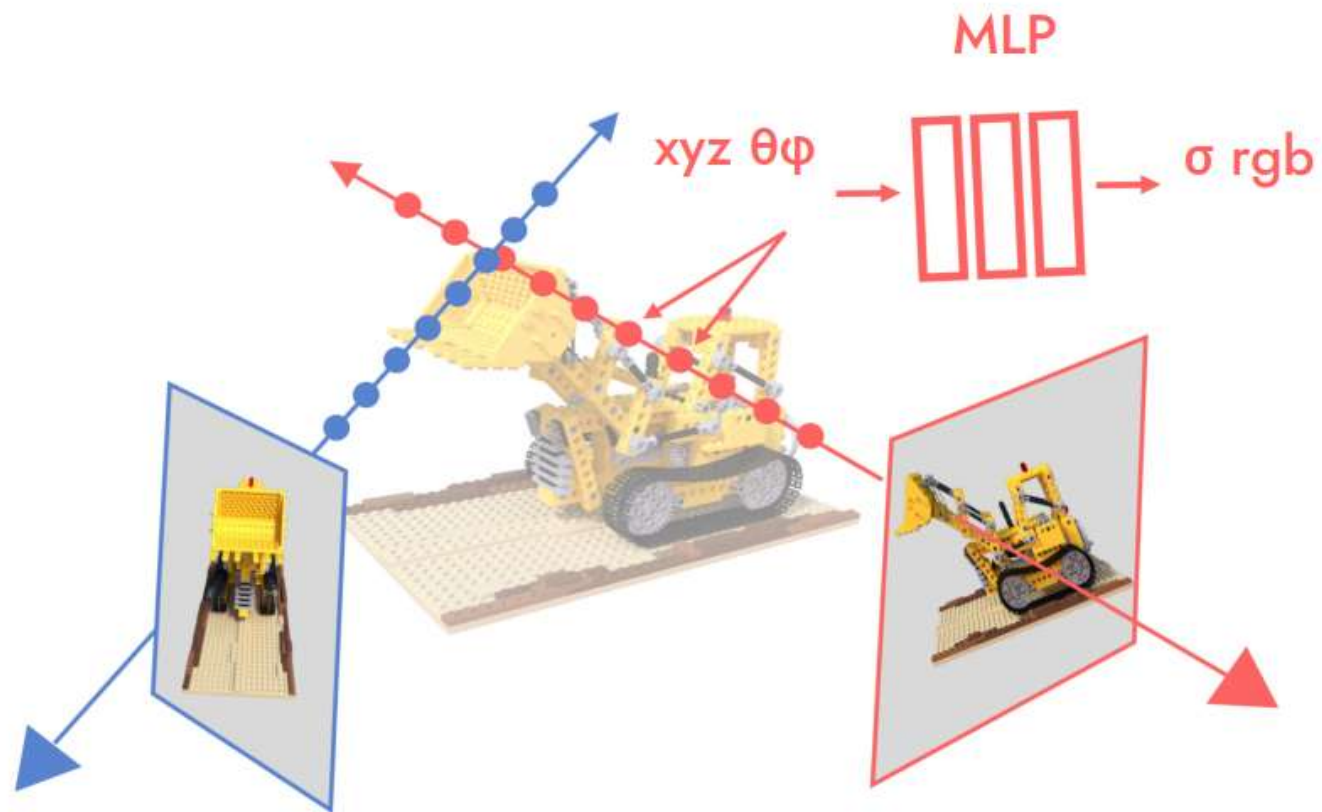
Surface-Based Rendering: Mesh



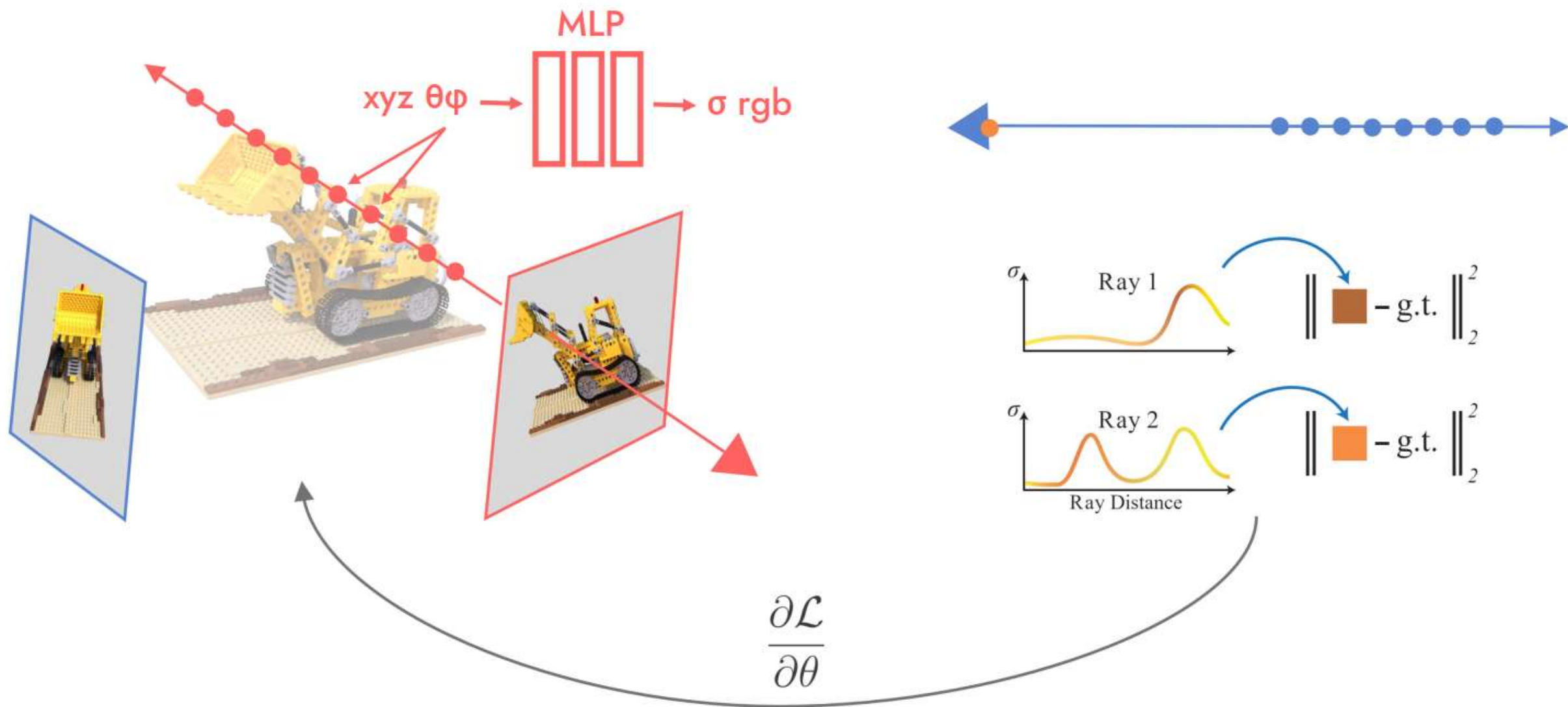
Surface-Based Rendering: Mesh



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

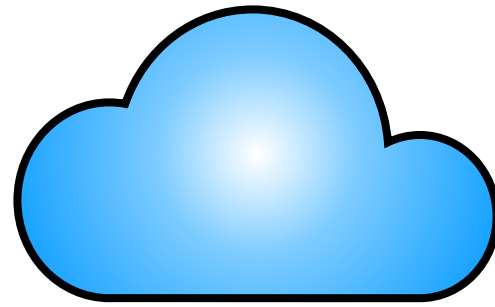


<https://dynamic3dgaussians.github.io/>

Surface-Based vs. Volume-Based

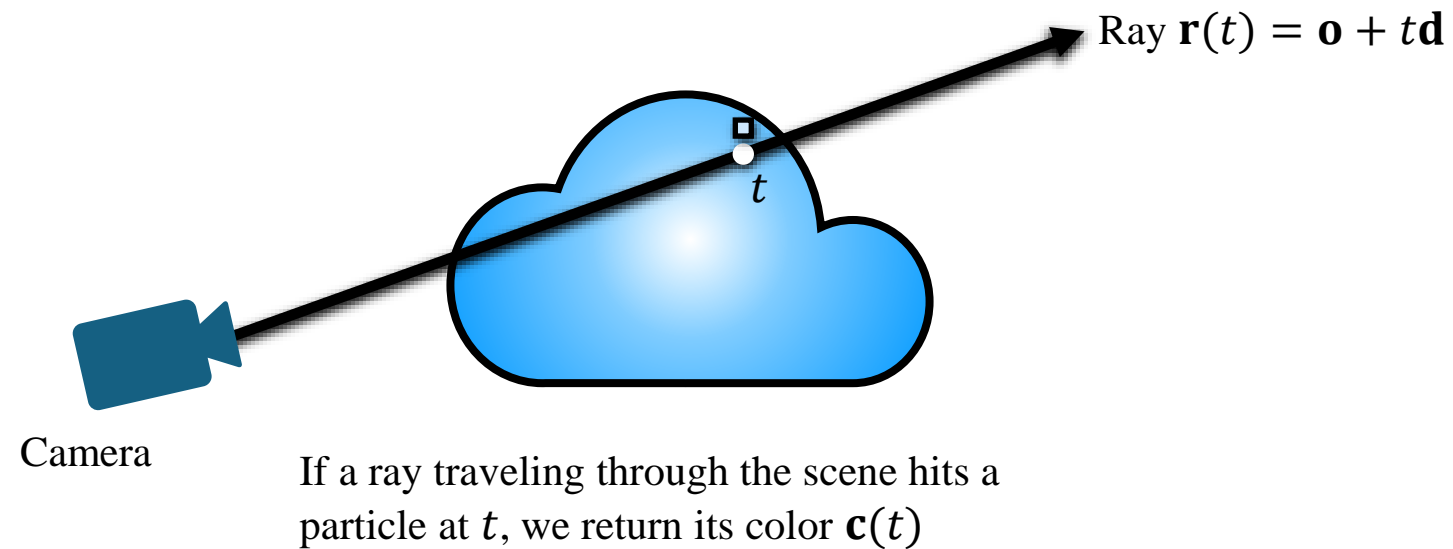
Aspect	NeRF / 3DGS (Volume-based)	Mesh (Surface-based)
Modeling Paradigm	Perception-driven learning of unknown 3D structures	Explicit modeling of known structures
Input Requirement	Real-world images/videos (multi-view), camera poses	Explicit meshes, CAD models, or known geometry
Topology Prior	Not required; 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

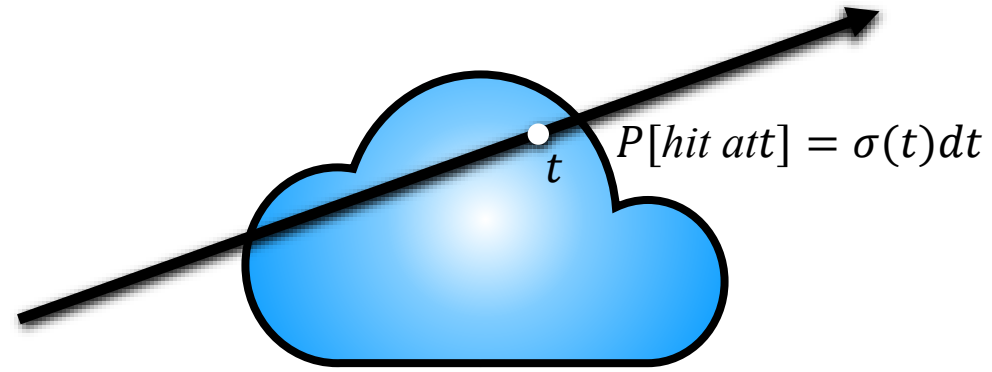


Scene is a cloud of tiny colored particles

Volume Rendering

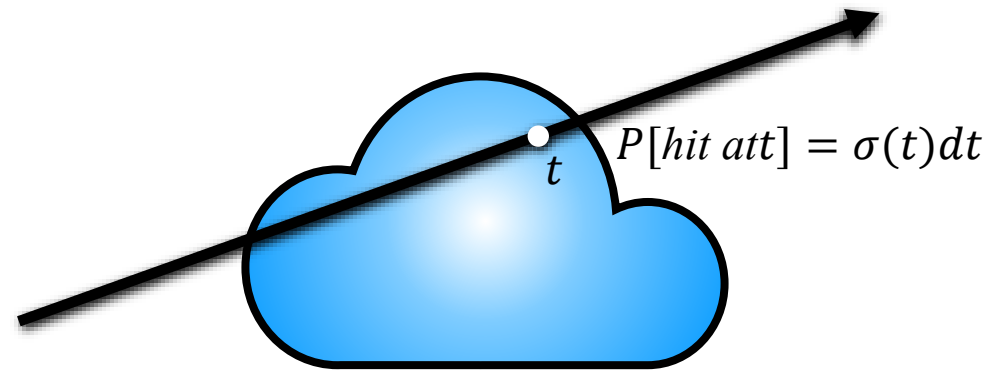


Volume Rendering



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

Volume Rendering

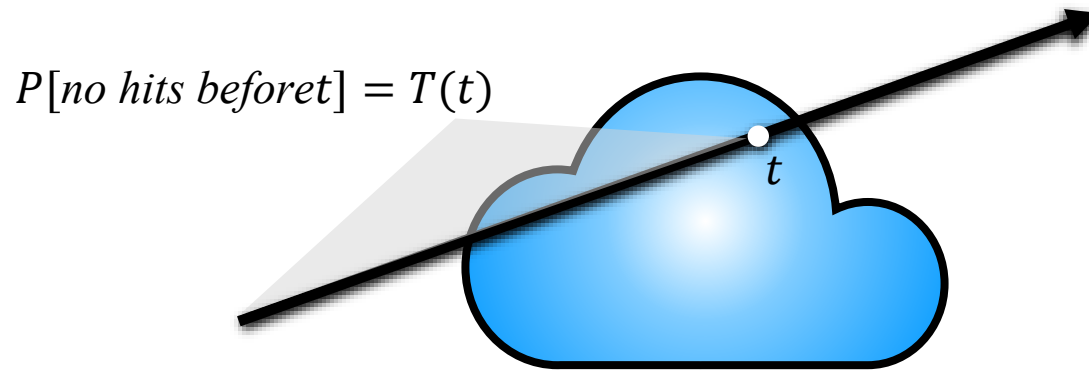


Probability that ray stops in a small interval around t is $\sigma(t)dt$.
 σ is known as the Volume Density.

Our field thus has the function signature:

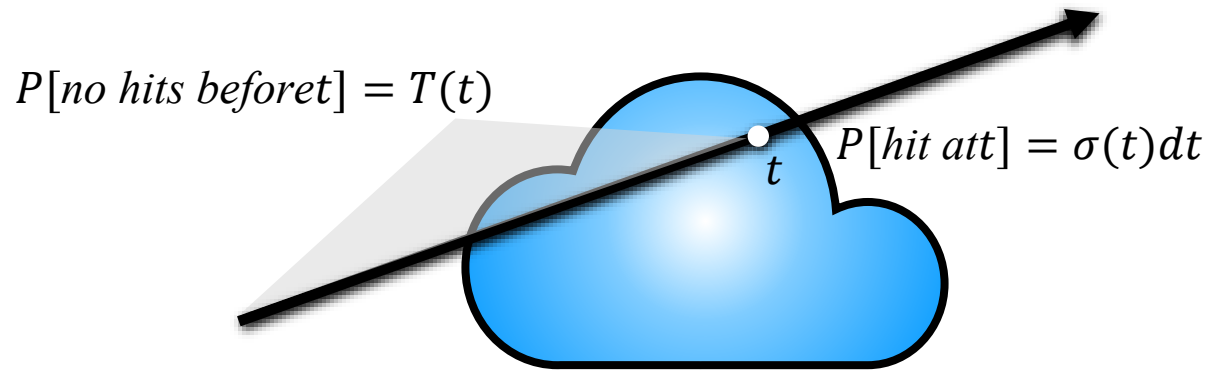
$$\Phi: \mathbb{R}^3 \rightarrow \mathbb{R}^3 \times \mathbb{R}^+; \Phi(\mathbf{x}) = (\sigma, \mathbf{c})$$

Volume Rendering



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.

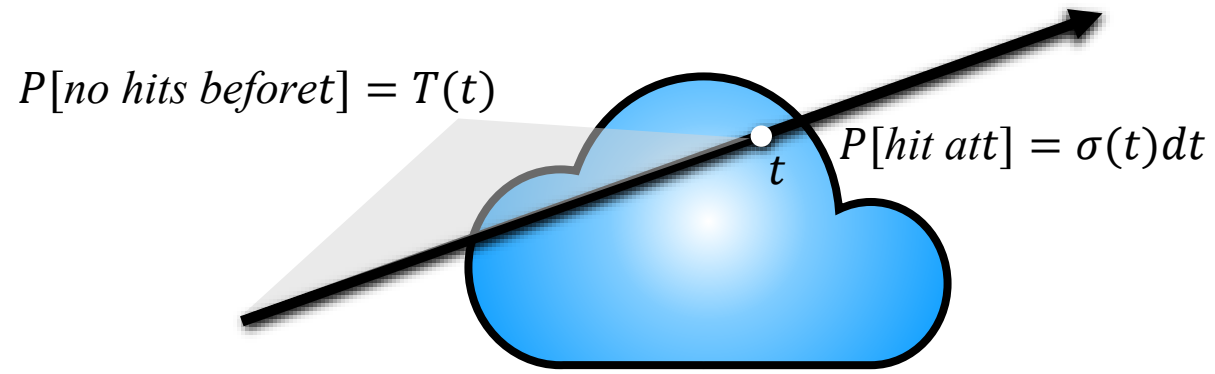
Volume Rendering



σ and T are related via

$$P[\text{no hit before} + dt] = P[\text{no hit before}] \times P[\text{no hit att}]$$

Volume Rendering



σ and T are related via

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

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 .

Volume Rendering

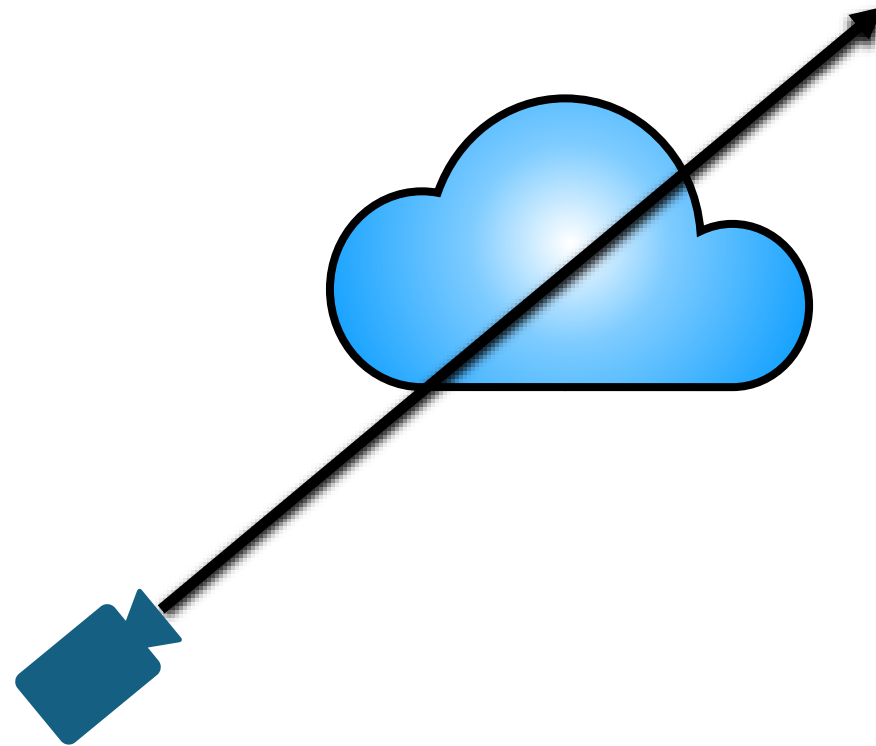
Then $T'(t) = T(t)\sigma(t)$ is probability that ray stops *exactly* at t .

So the expected color returned by the ray will be

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

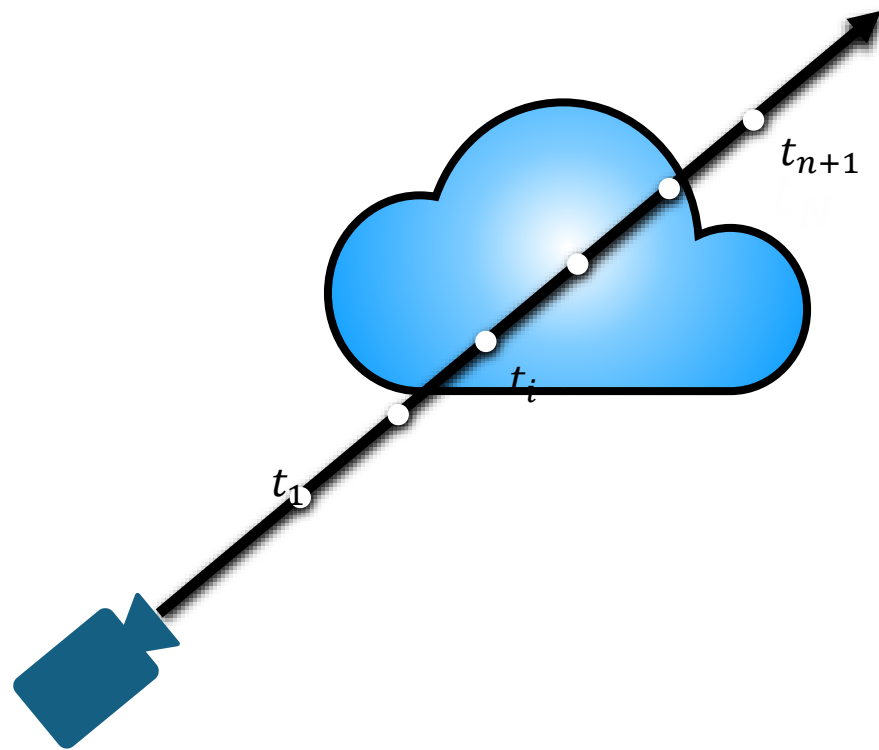
Note the nested integral!

Approximating the nested integral



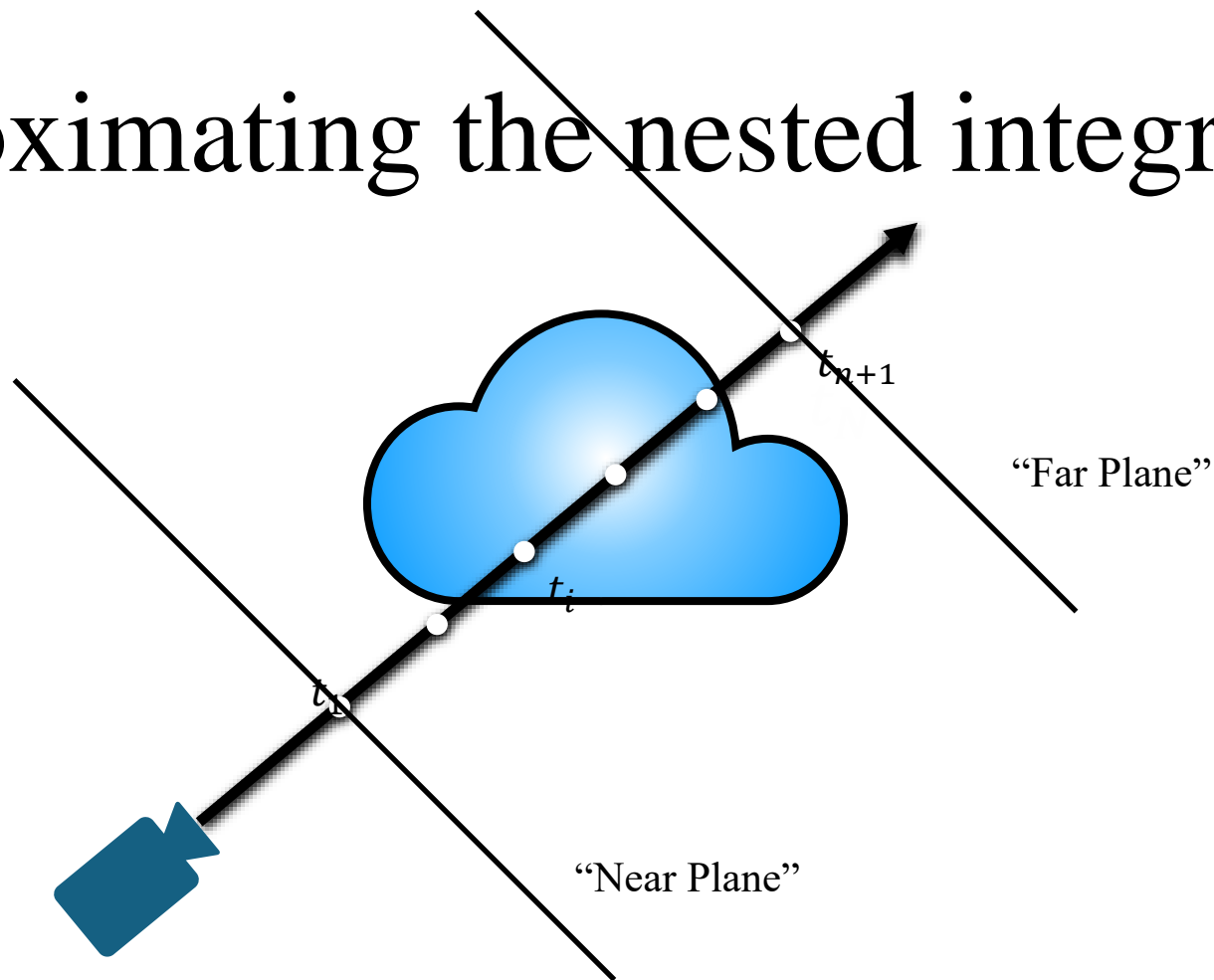
We use quadrature to approximate the nested integral,

Approximating the nested integral



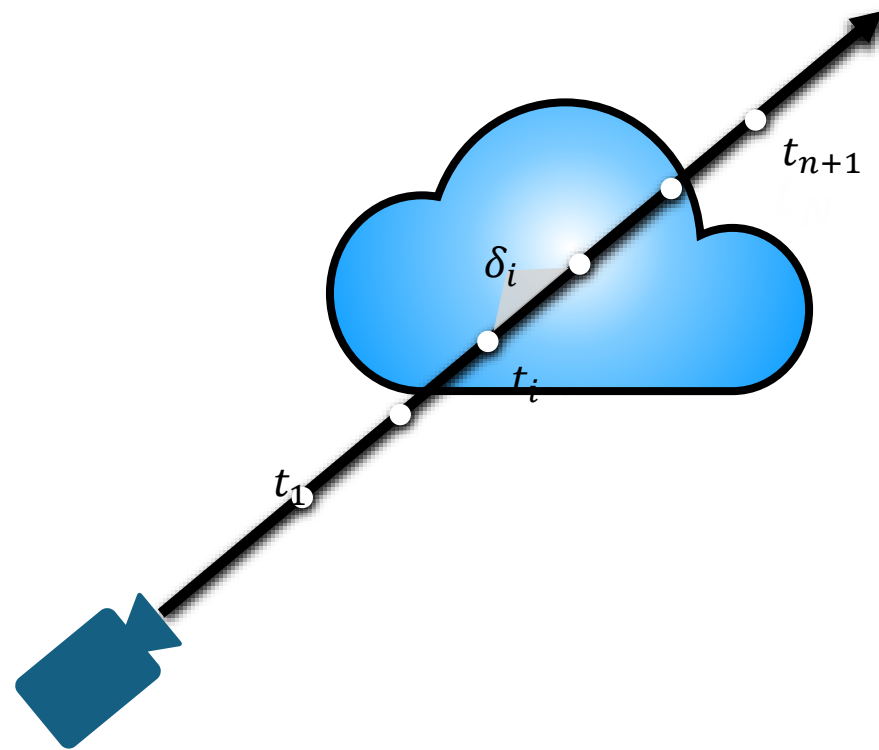
We use quadrature to approximate the nested integral, splitting the ray up into n segments with endpoints $\{t_1, t_2, \dots, t_{n+1}\}$

Approximating the nested integral



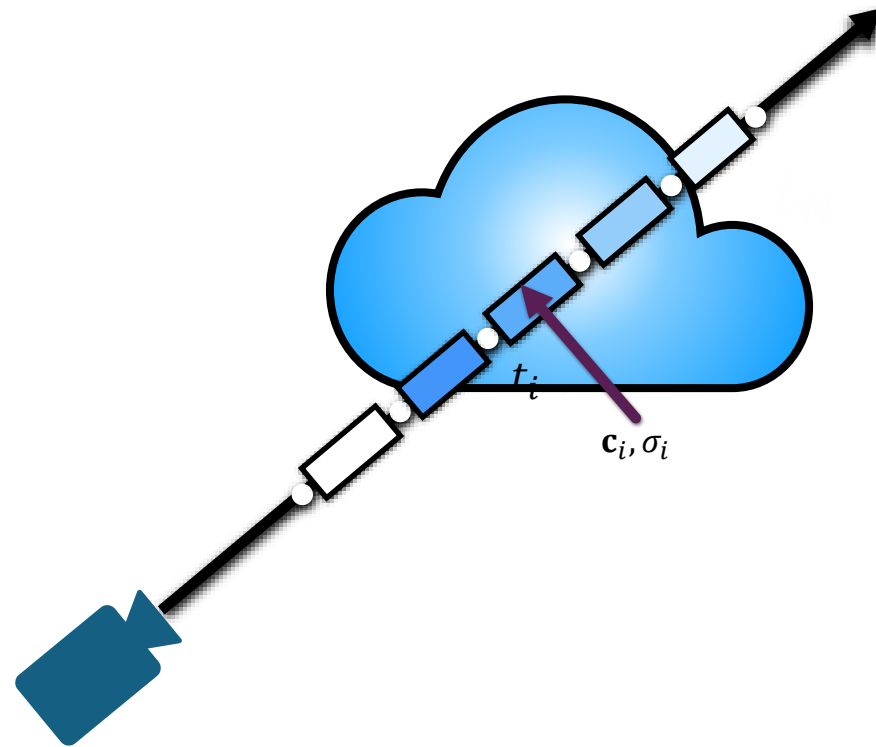
We use quadrature to approximate the nested integral, splitting the ray up into n segments with endpoints $\{t_1, t_2, \dots, t_{n+1}\}$

Approximating the nested integral



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

Approximating the nested integral



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

Summary: volume rendering integral estimate

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

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

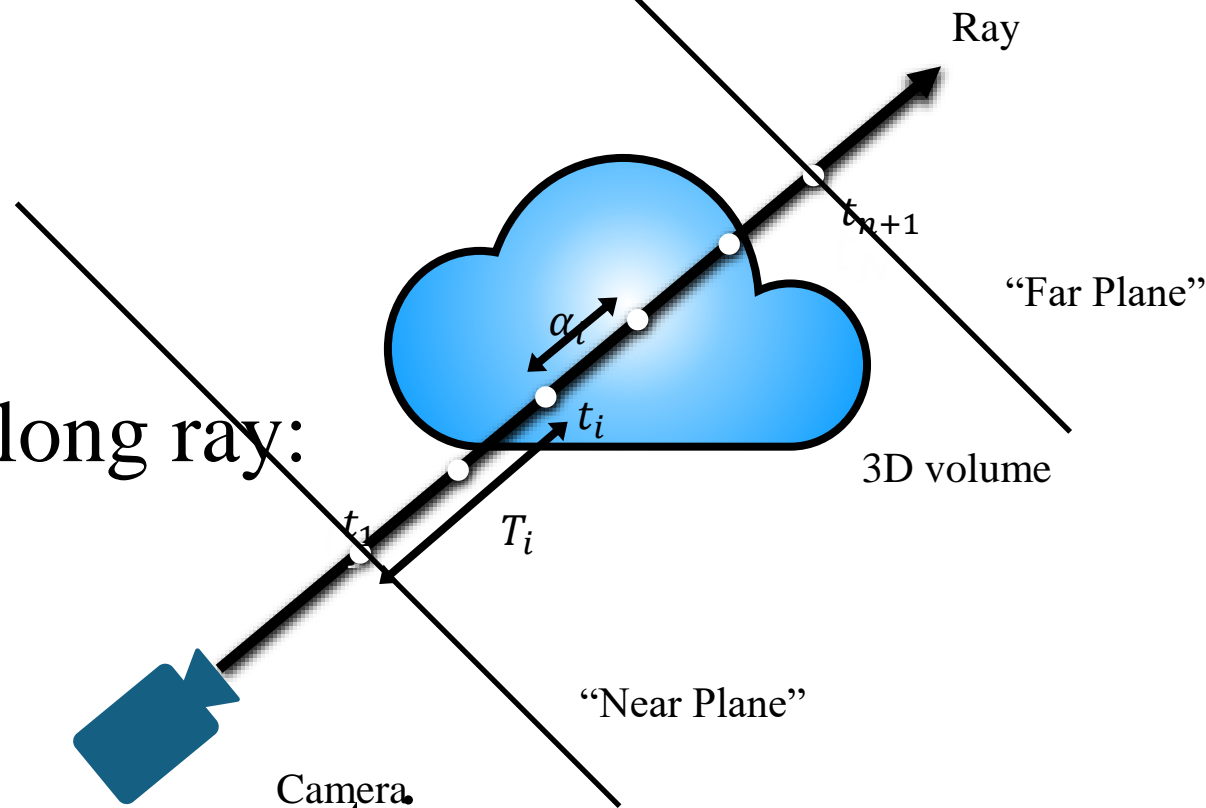
weights colors

How much light is blocked earlier along ray:

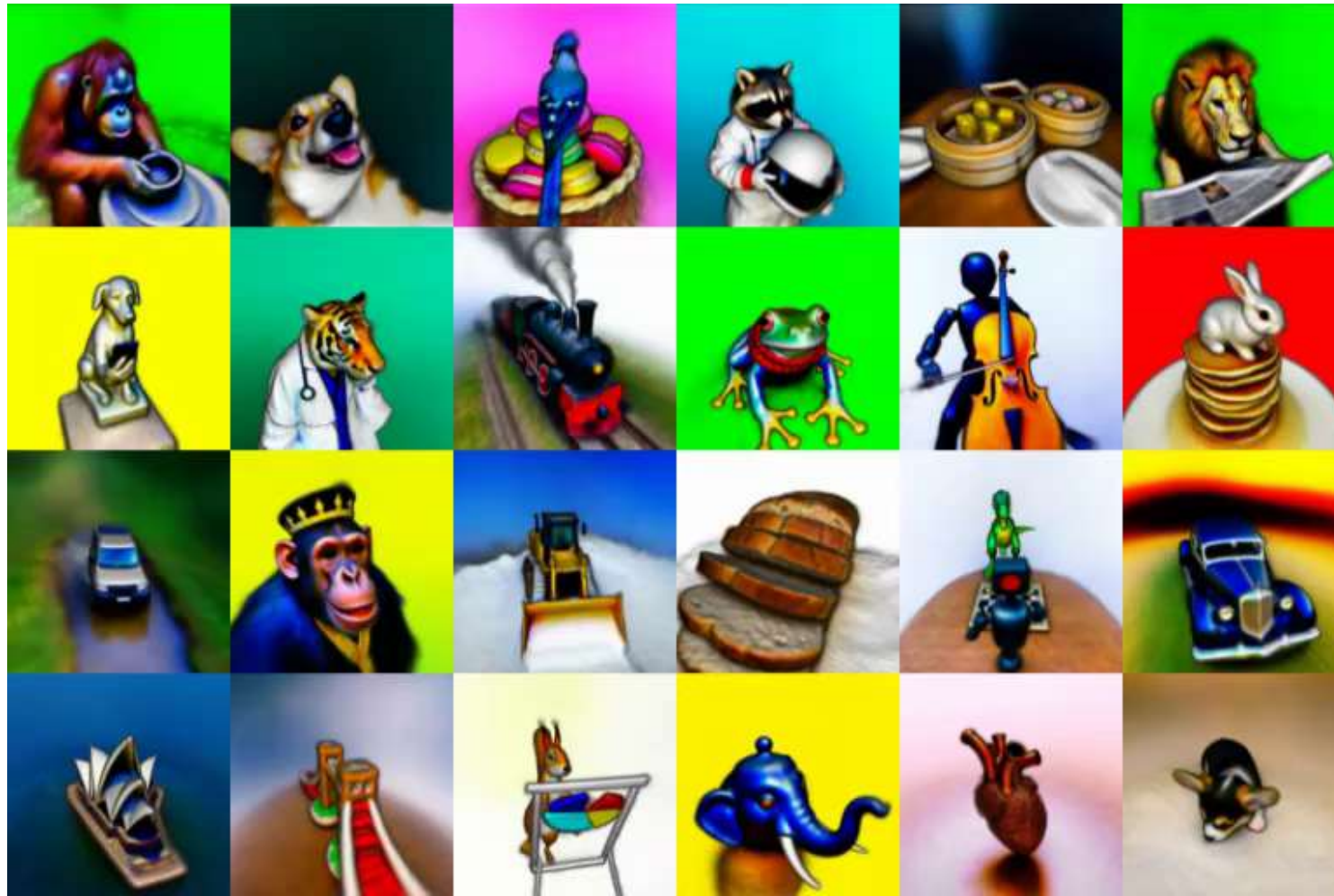
$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$$

How much light is contributed by ray segment i :

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



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



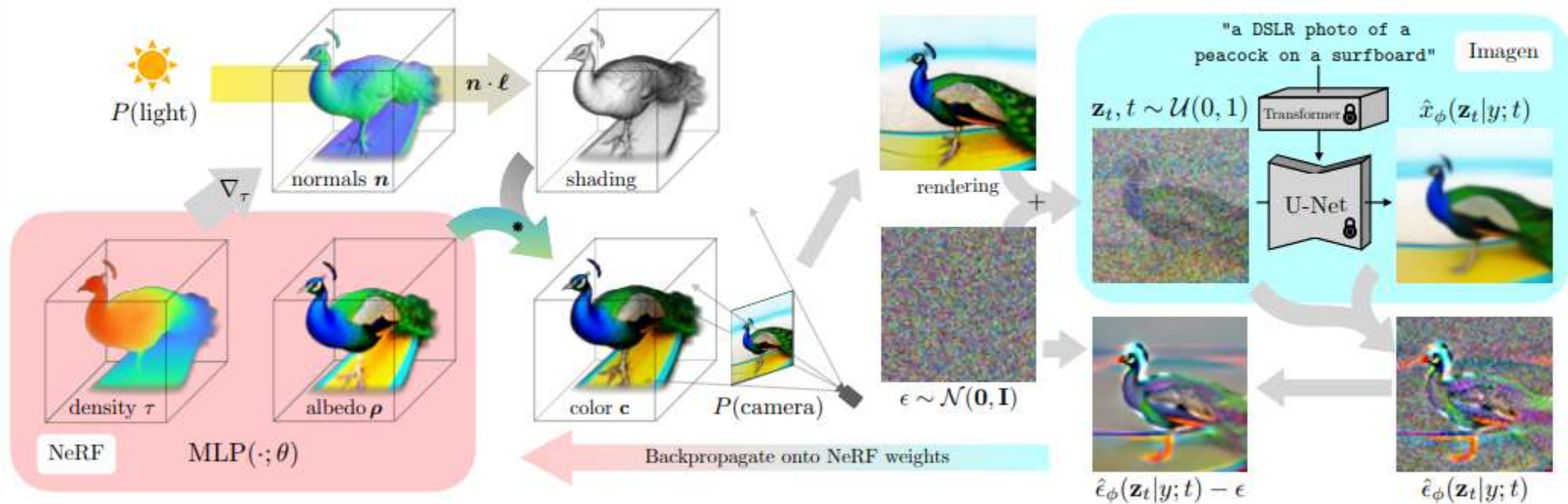
DreamFusion: Text-to-3D using 2D Diffusion

"a DSLR photo of a peacock on a surfboard"

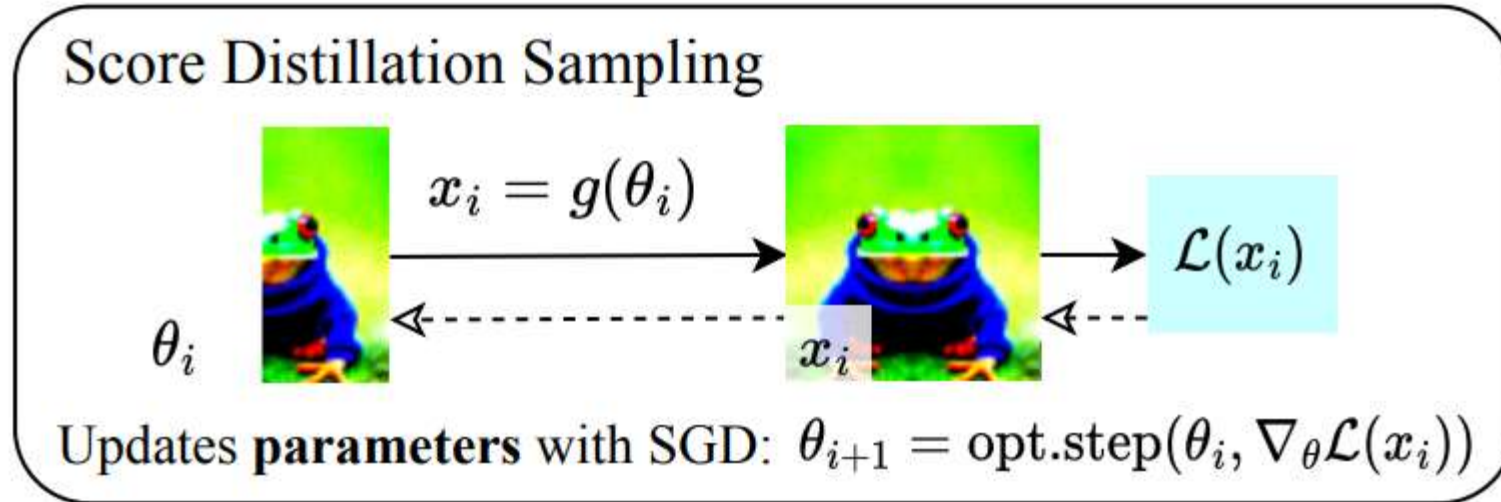
DreamFusion
Automatic text-to-3D



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



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

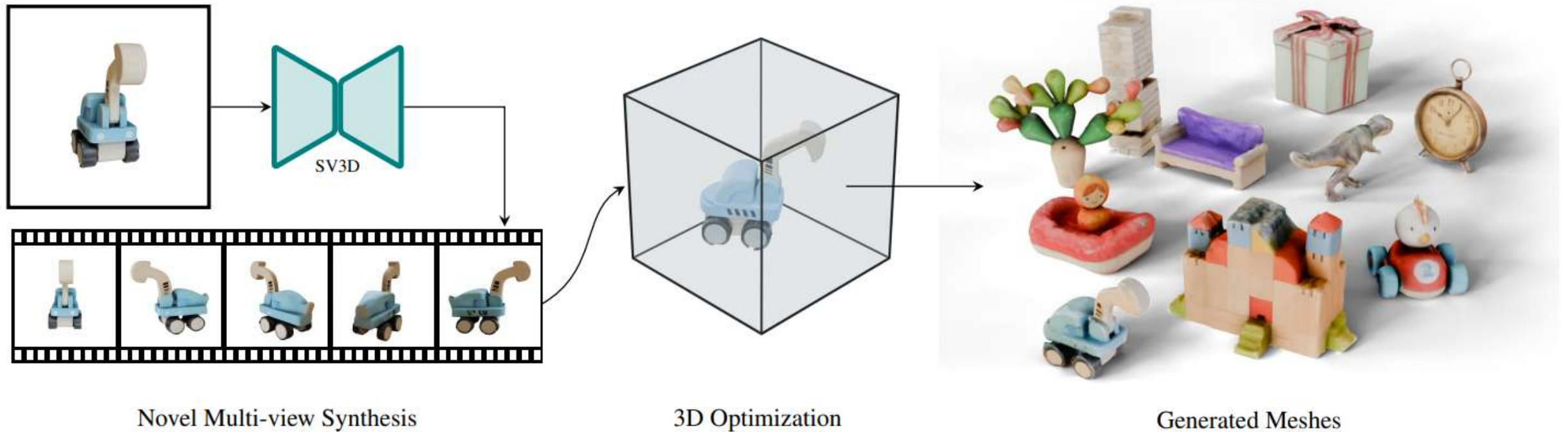


$$\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}) = \mathbb{E}_{t \sim \mathcal{U}(0,1), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [w(t) \|\epsilon_{\phi}(\alpha_t \mathbf{x} + \sigma_t \epsilon; t) - \epsilon\|_2^2]$$

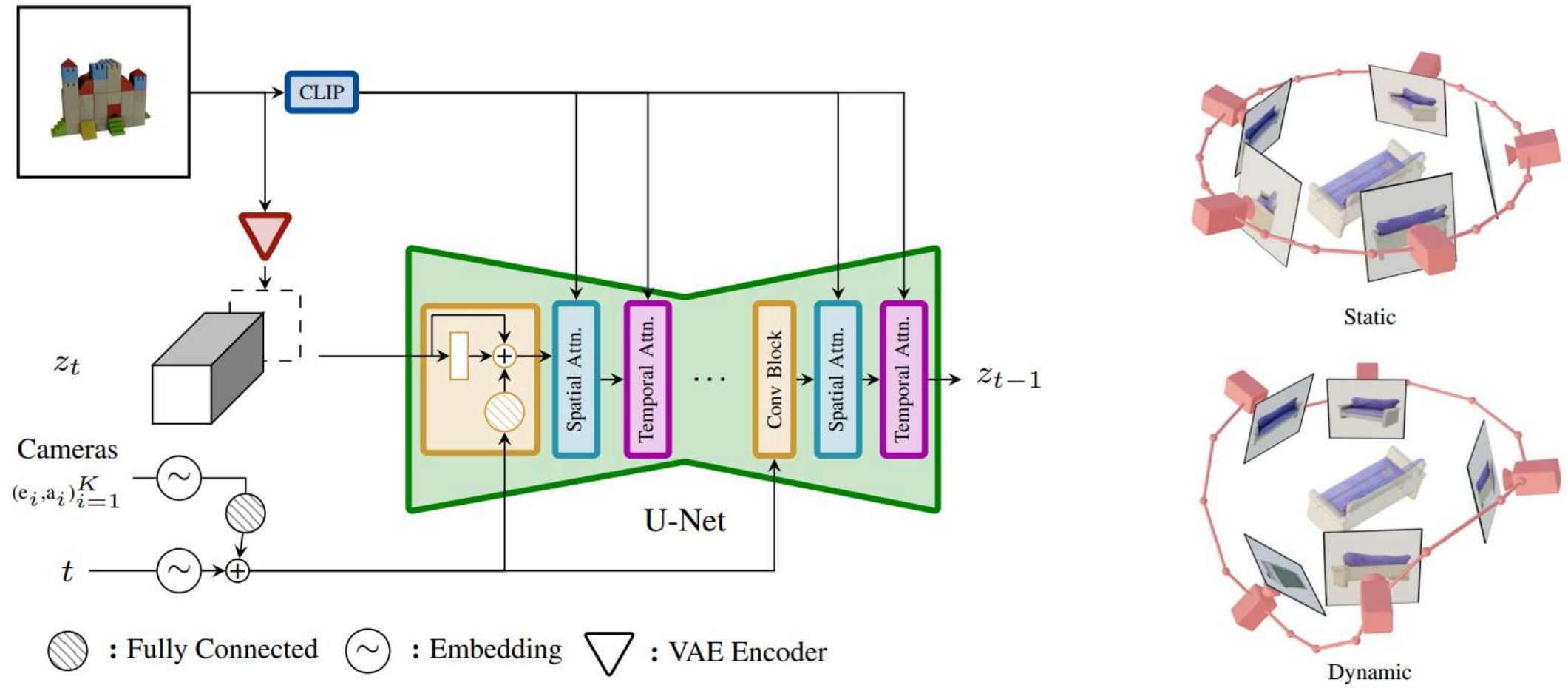
SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion



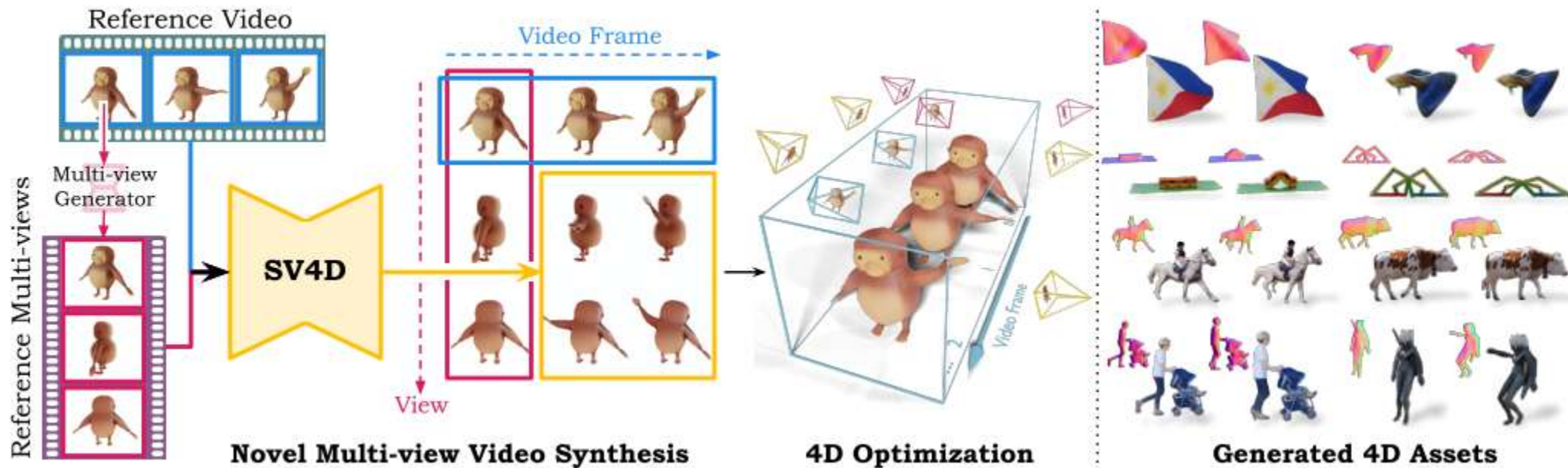
SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion



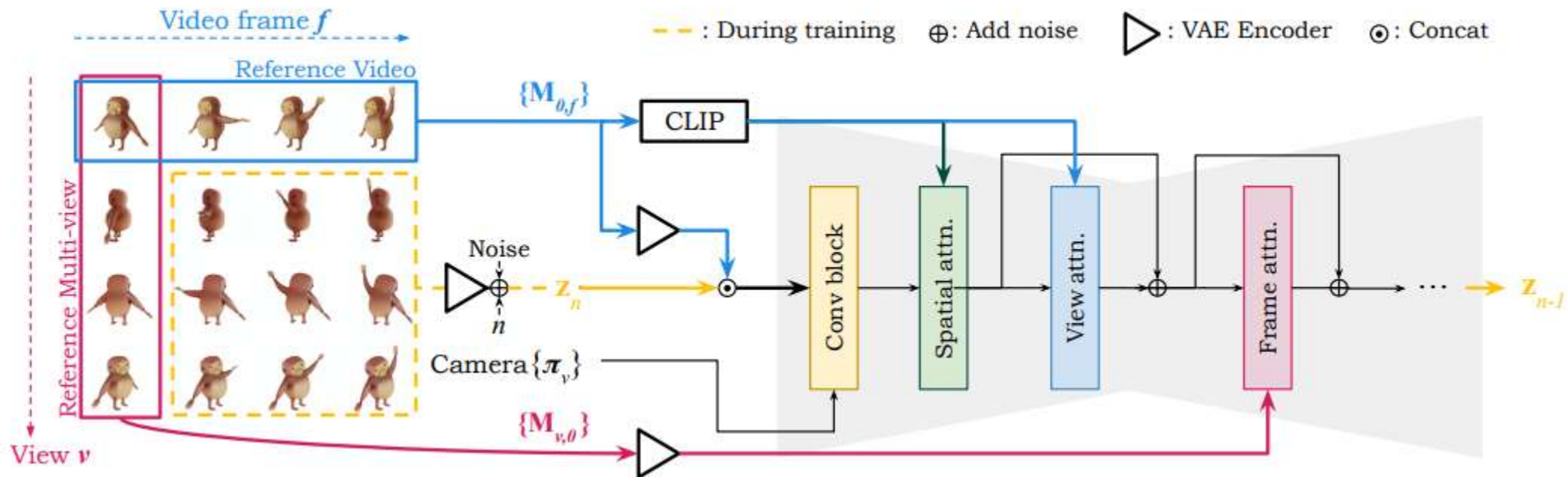
SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion



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



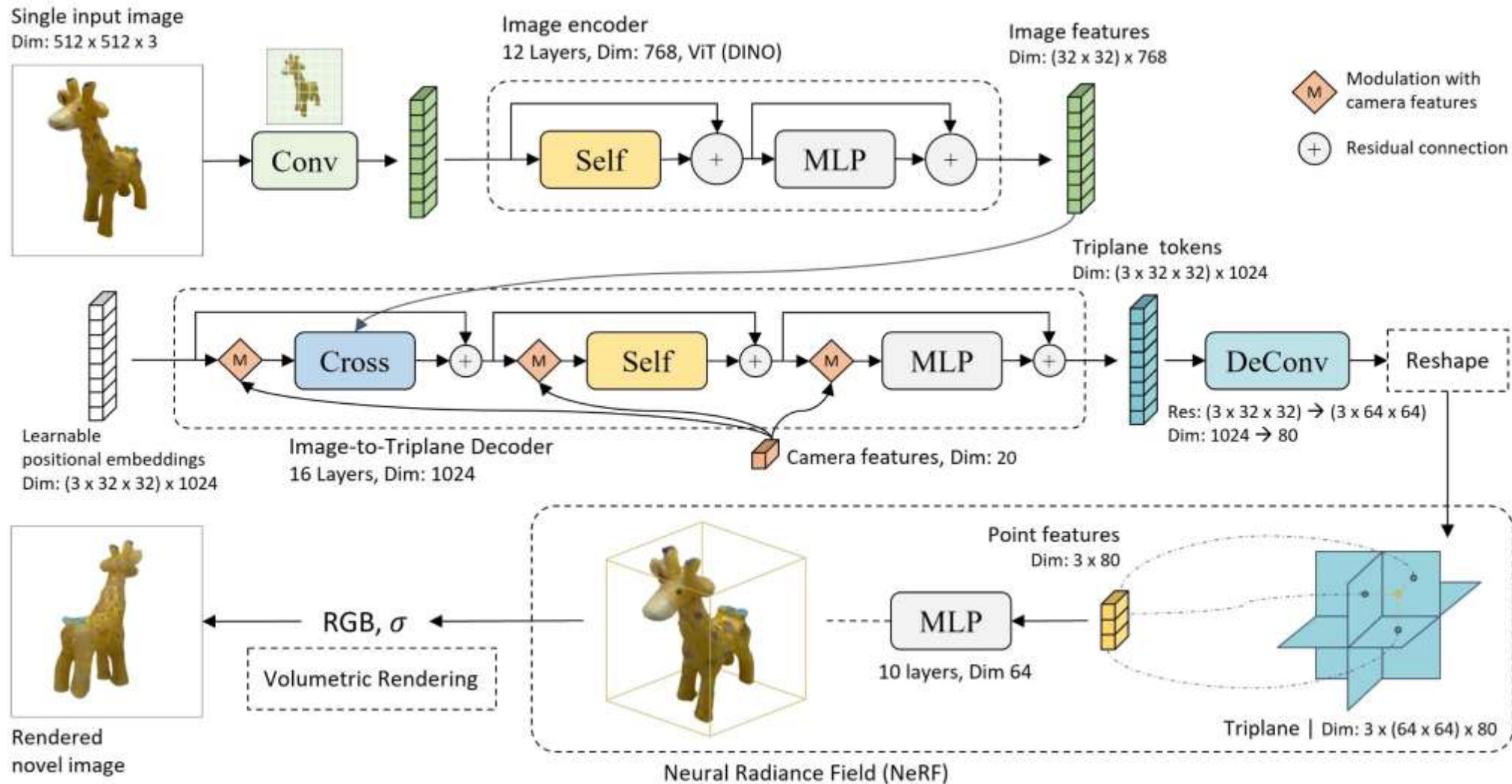
SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency



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



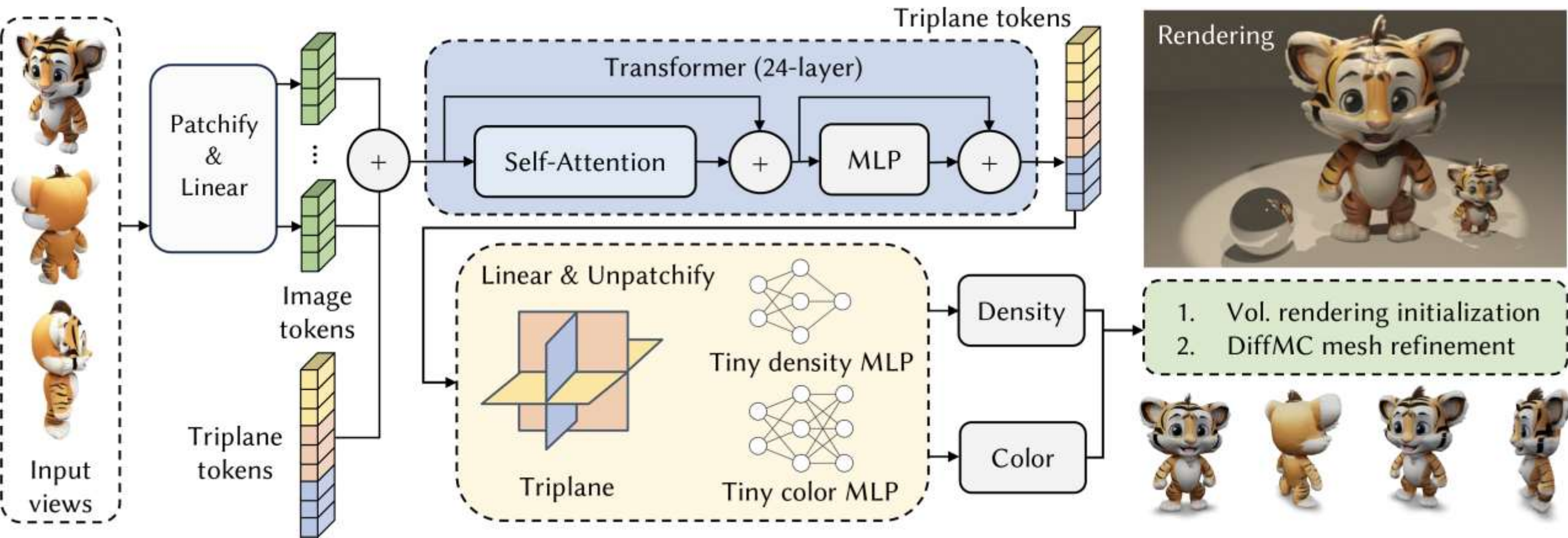
LRM: LARGE RECONSTRUCTION MODEL FOR SINGLE IMAGE TO 3D



LRM: LARGE RECONSTRUCTION MODEL FOR SINGLE IMAGE TO 3D



MeshLRM



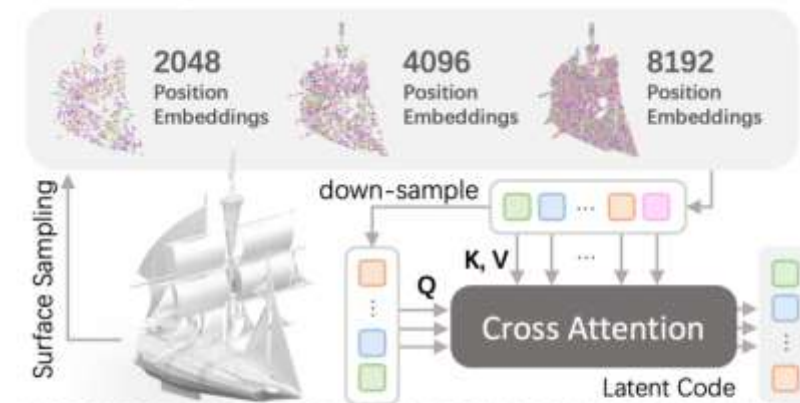
MeshLRM



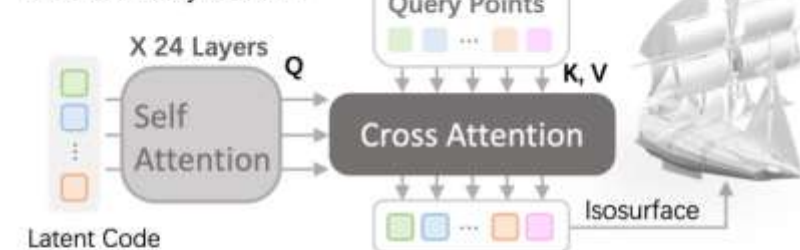
CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets



VAE Geometry Encoder



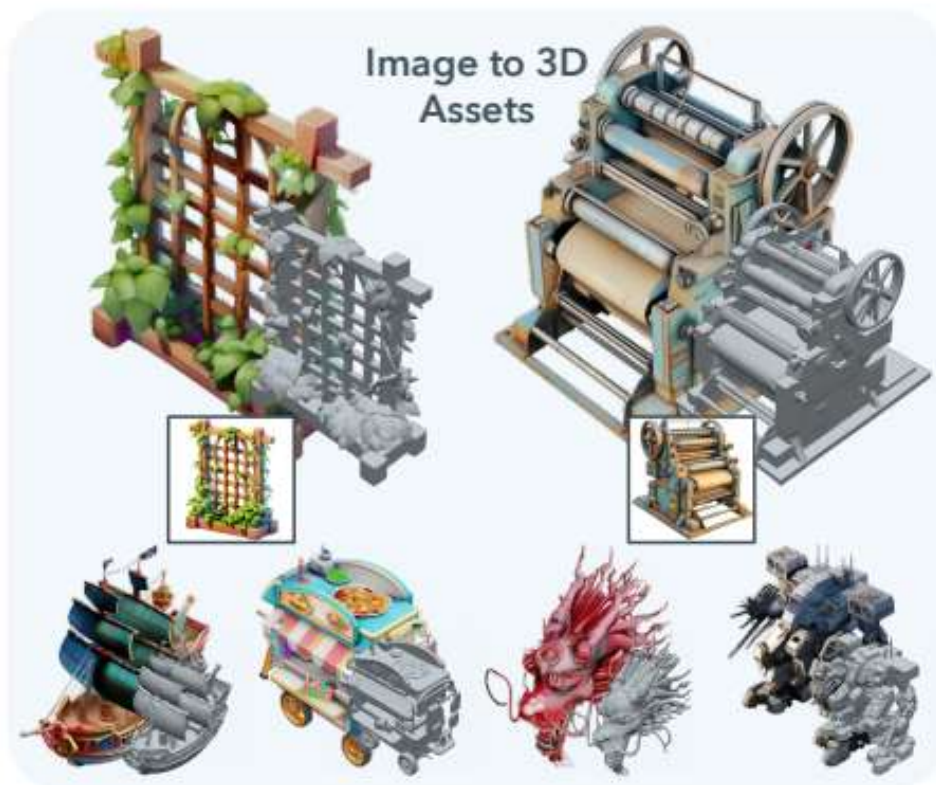
VAE Geometry Decoder



Latent Diffusion Model



TRELLIS

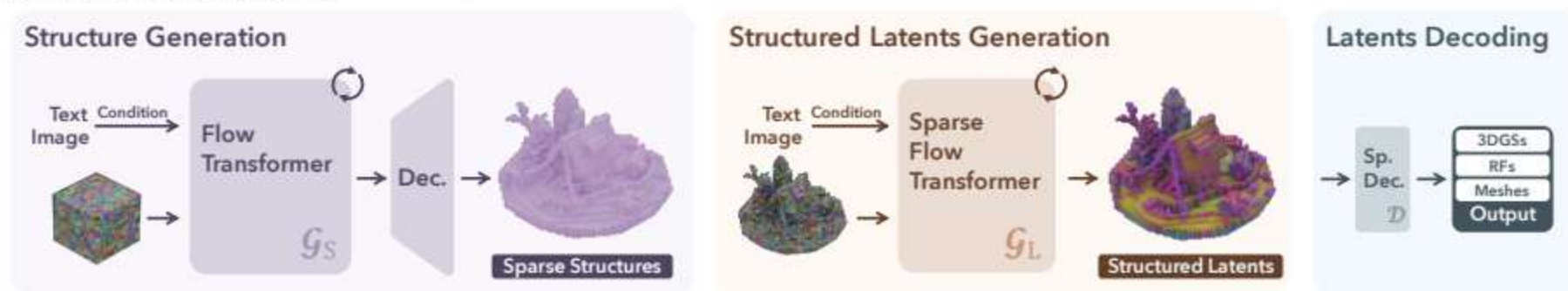


TRELLIS

3D Assets Encoding & Decoding



3D Assets Generation



TRELLIS.2

Reconstruction



Generation

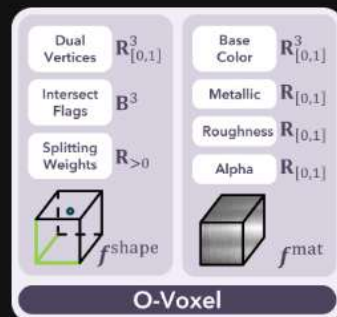


TRELLIS.2

TECH INNOVATIONS



TRELLIS.2's pipeline begins with an Instant Bidirectional Conversion that transforms meshes into our new representation termed O-Voxel. A Sparse Compression VAE then encodes these voxels into a compact Structured Latent space.



O-Voxel: *Omni-Voxel Representation*

O-Voxel is a novel "field-free" sparse voxel structure designed to encode both precise geometry and complex appearance simultaneously.

GEO

Geometry (f_{shape})

Utilizing a Flexible Dual Grids representation to handle arbitrary topologies while preserving sharp edges.

MAT

Appearance (f_{mat})

Supports full PBR attributes (Base Color, Metallic, Roughness, Alpha) to accurately model rich surface materials.

SC-VAE: *Sparse Compression VAE*

We introduce a Sparse Compression 3D VAE, employing a Sparse Residual Autoencoding scheme to directly compress voxel data.

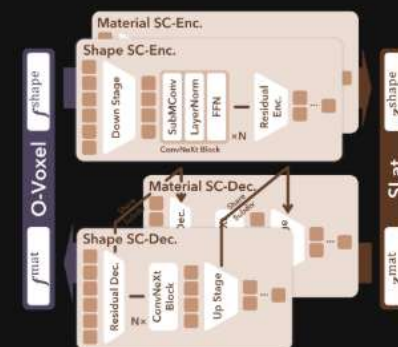
16x

Downsampling

~9.6K

Latent Tokens for 1024^3

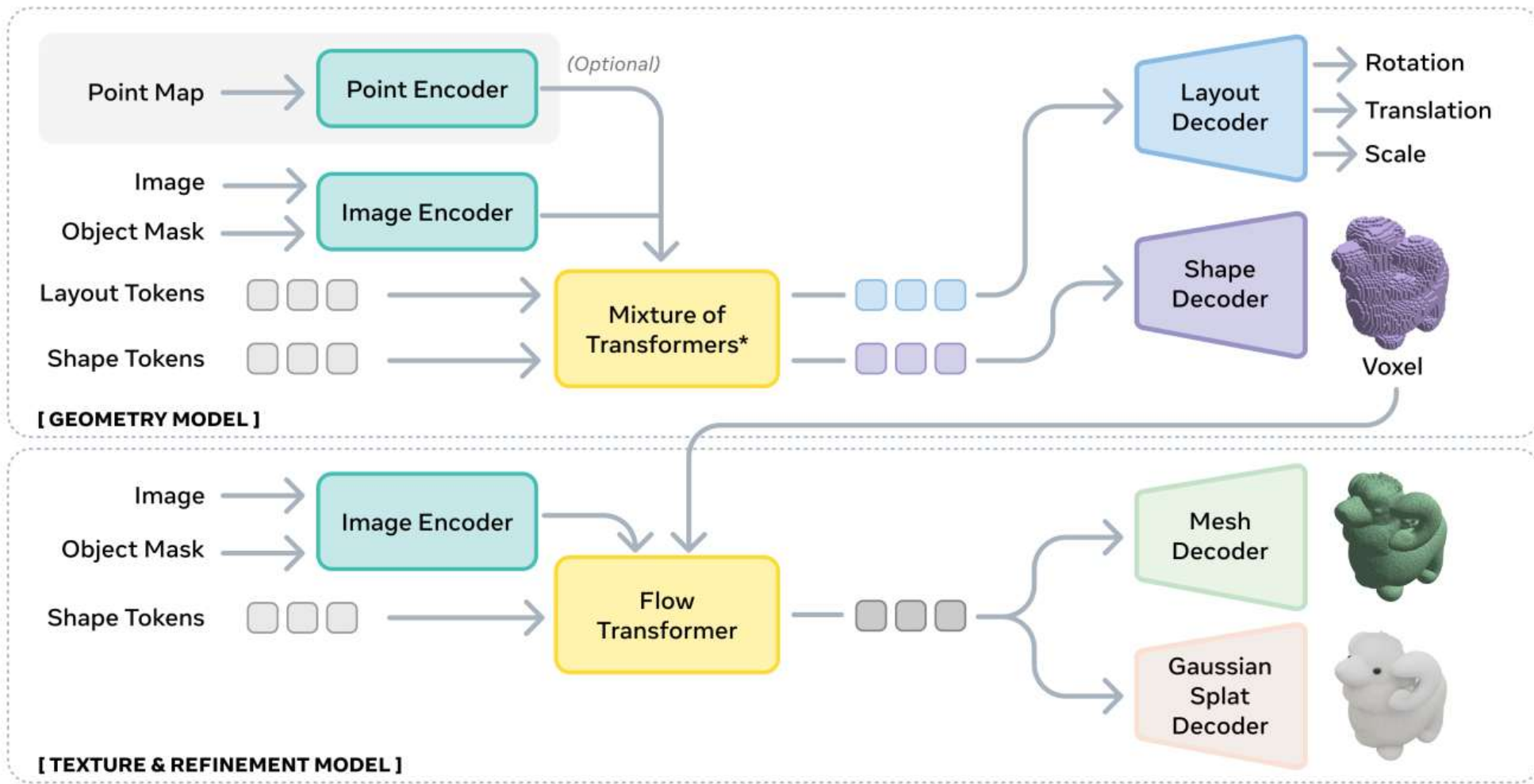
It encodes a fully textured 3D asset into a highly compact representation with negligible perceptual degradation, enabling efficient large-scale generative modeling.



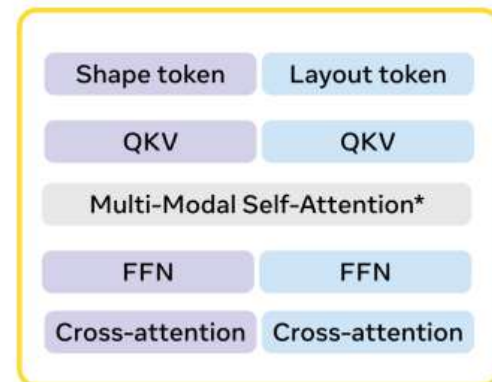
SAM 3D



SAM 3D



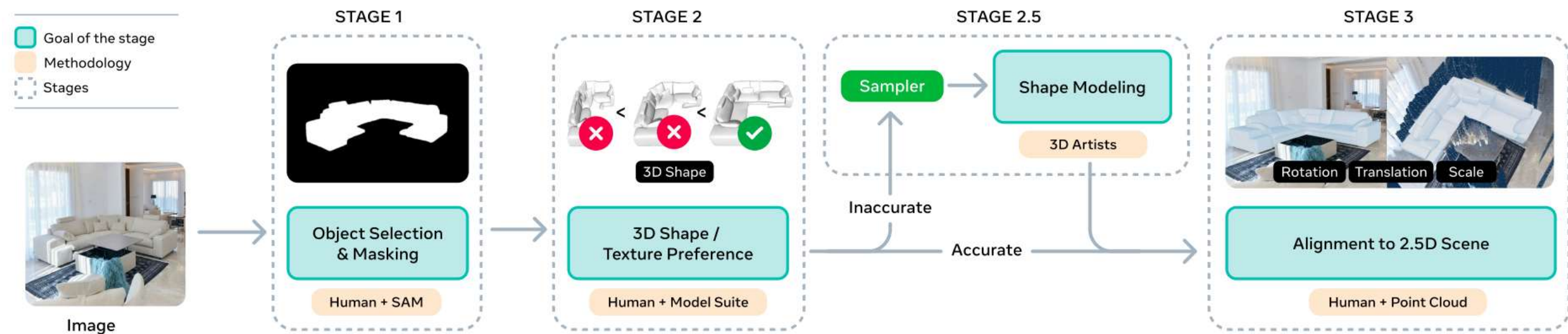
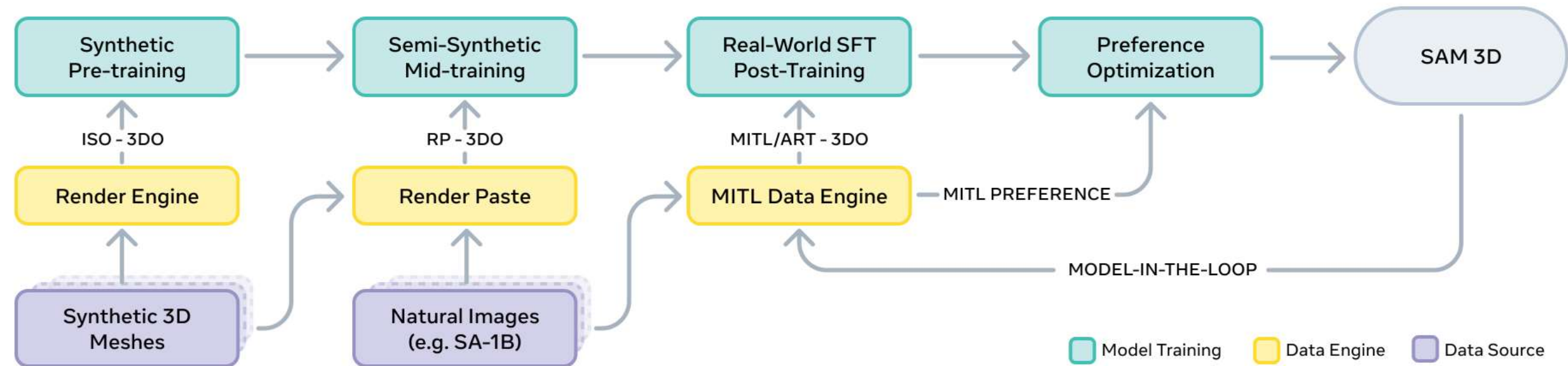
*Mixture of Transformers



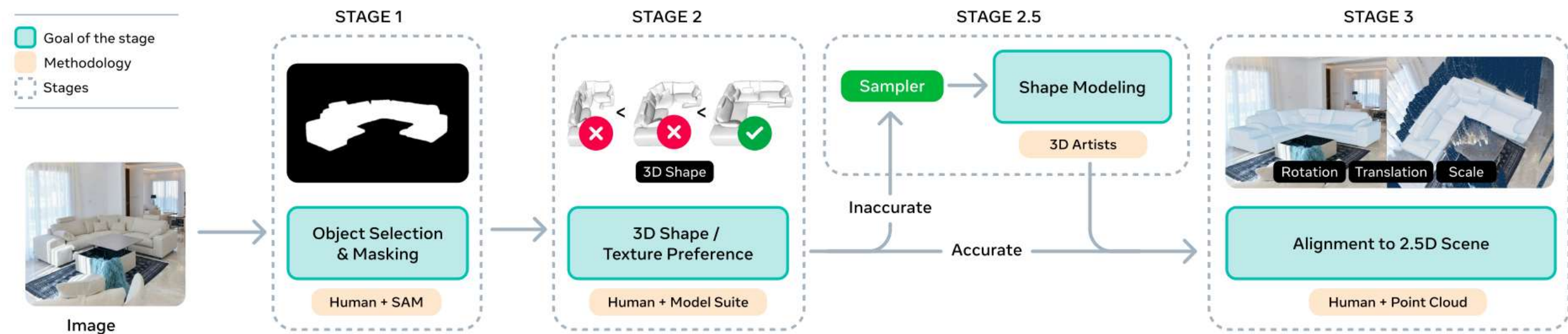
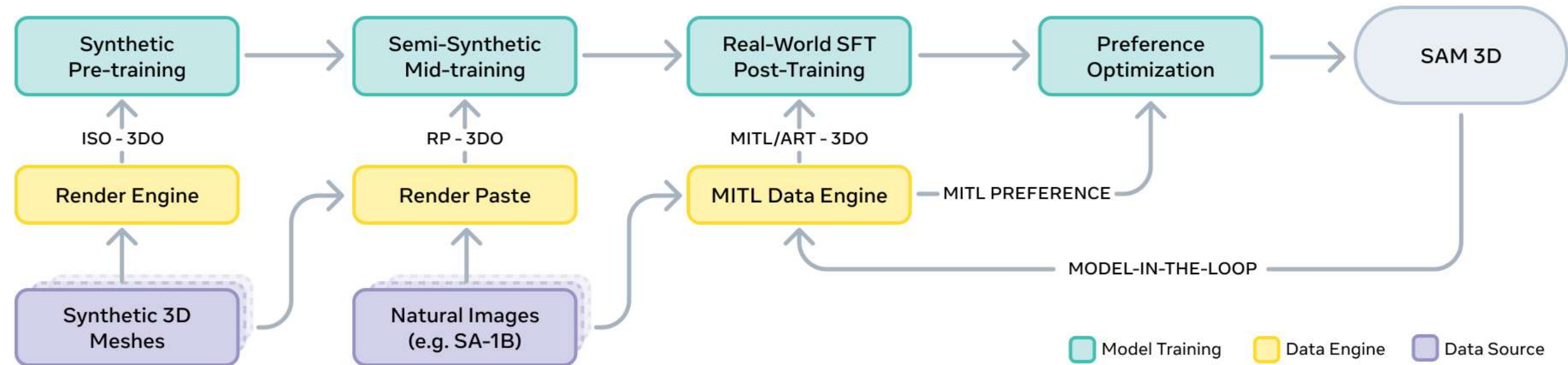
*Multi-Modal Self-Attention Mask



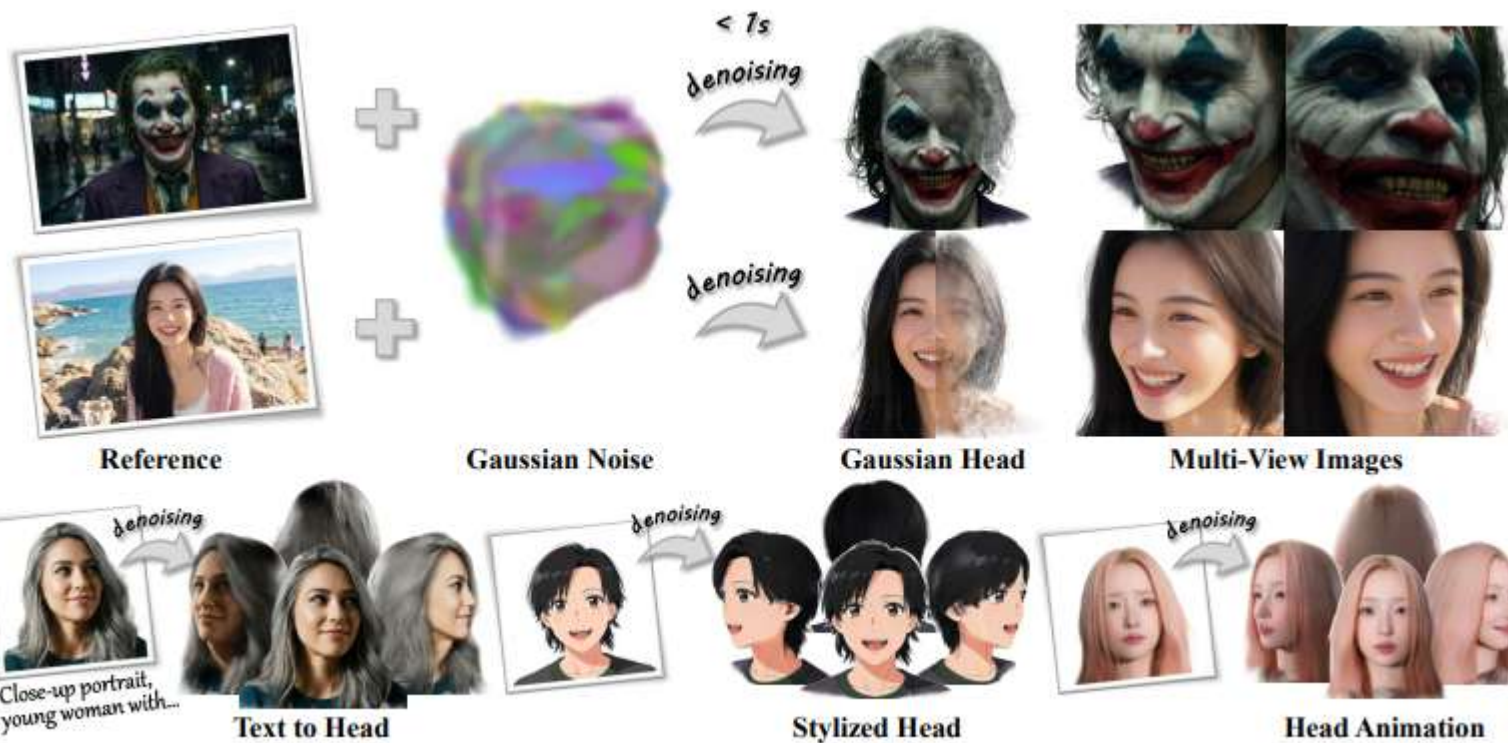
SAM 3D



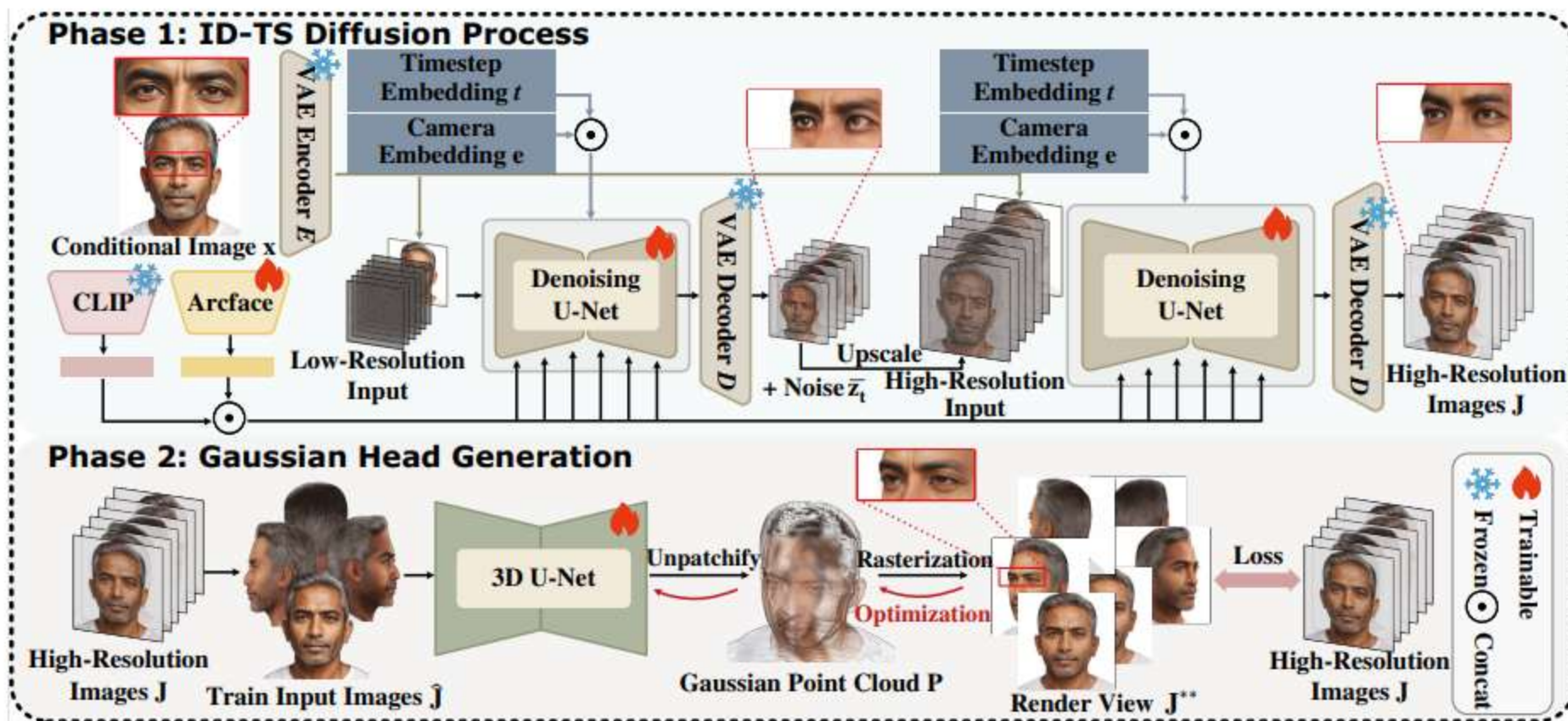
SAM 3D



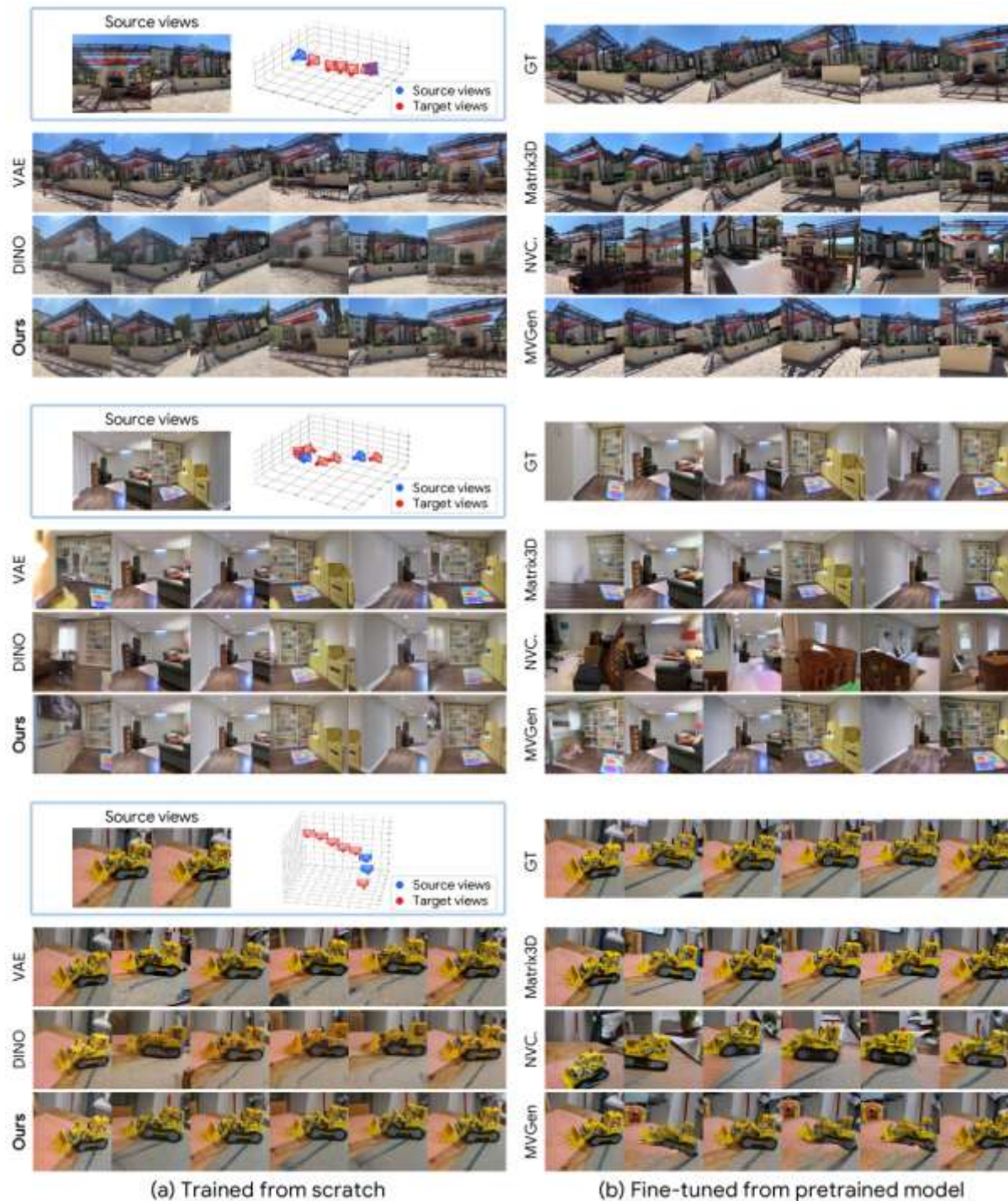
AnyHead & AnyAvatar3D



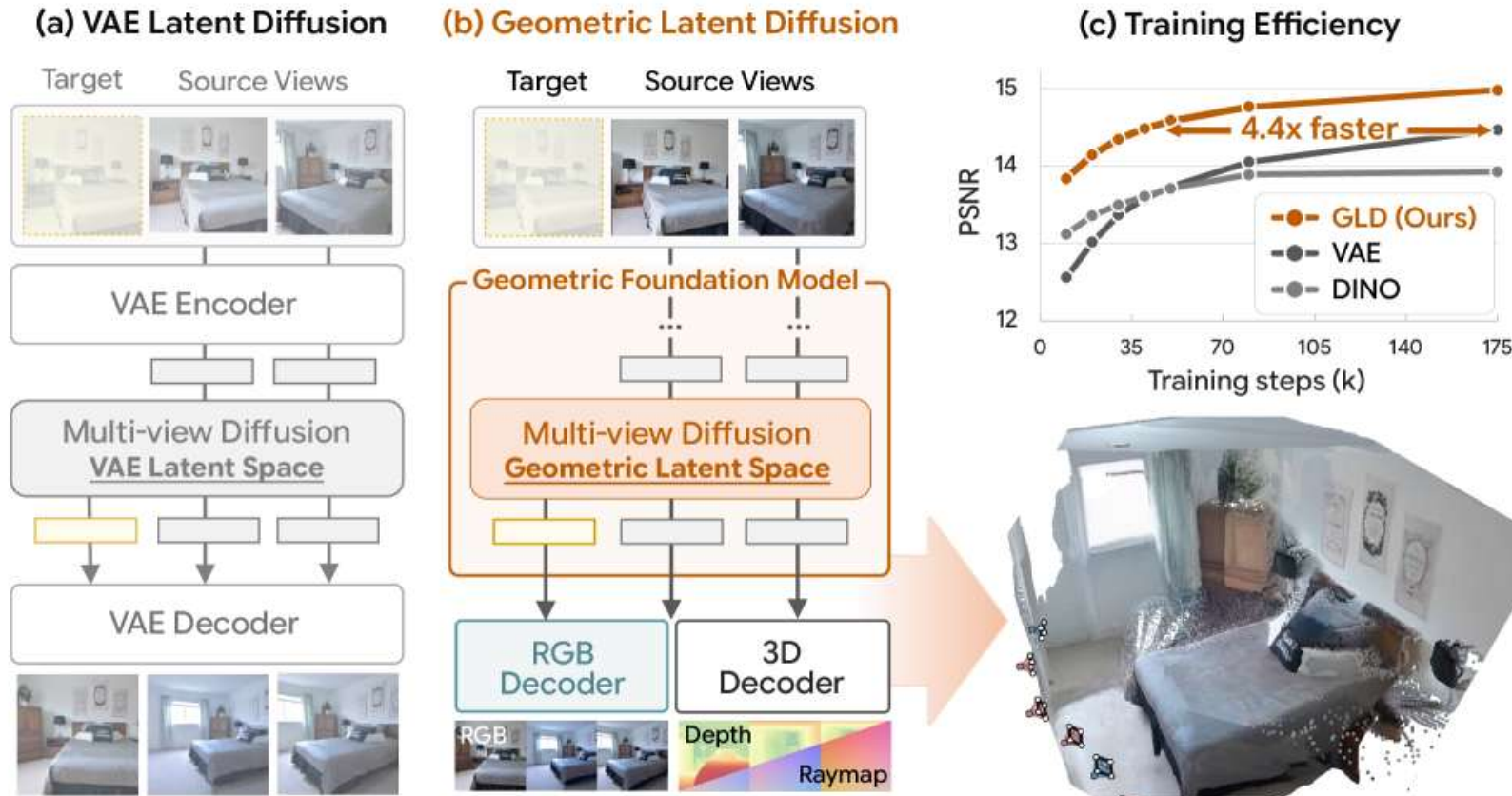
AnyHead & AnyAvatar3D



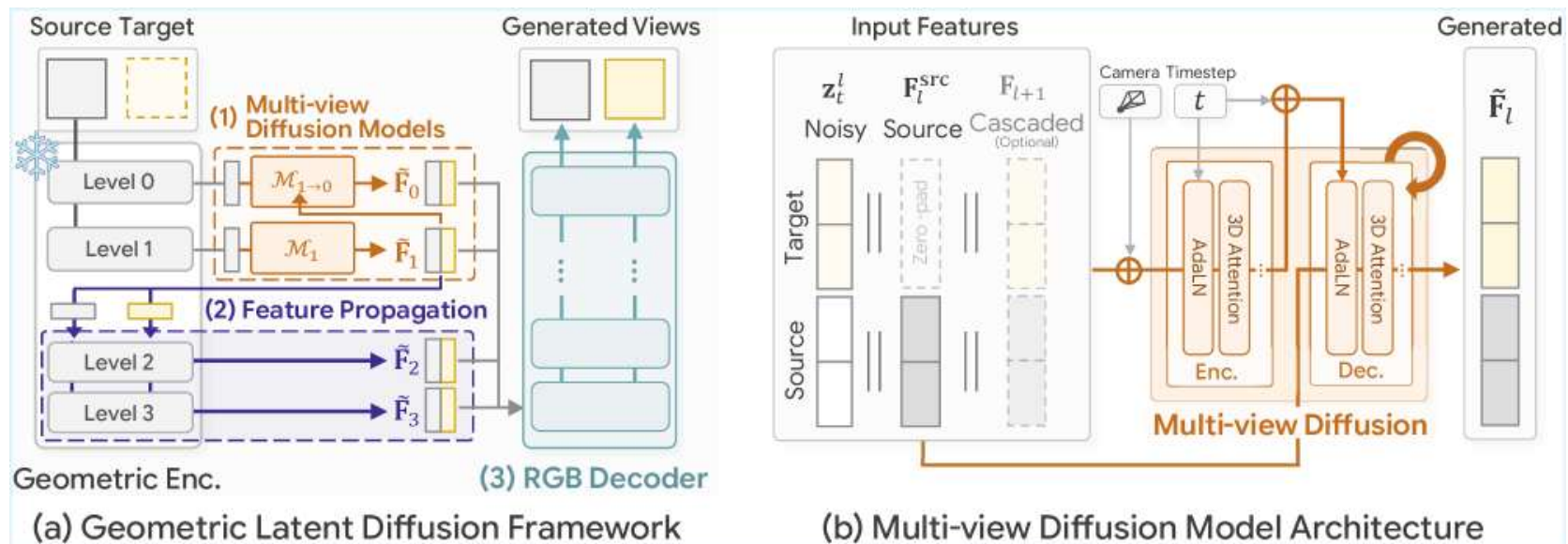
GLD: Repurposing Geometric Foundation Models for Multi-view Diffusion



GLD: Repurposing Geometric Foundation Models for Multi-view Diffusion



GLD: Repurposing Geometric Foundation Models for Multi-view Diffusion

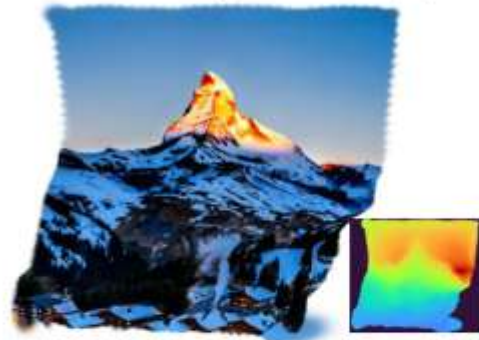


VIST3A: Text-to-3D by Stitching a Multi-view Reconstruction Network to a Video Generator

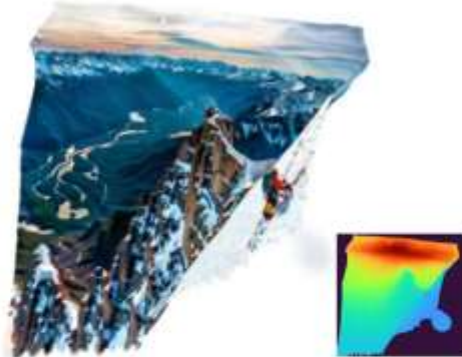


"A wooden rocking horse in a child's playroom"

"A golden retriever with a blue bowtie"



"A majestic view of the Matterhorn at sunrise, ... the scene is surrounded by snowy slopes and serene winter atmosphere."



"An alpinist scaling a dramatic, snow-covered mountain face ... the climber is captured mid-ascent, emphasizing scale and solitude against the immense landscape."

(a) Text-to-3DGS



"A bedroom scene features a large bed adorned with white linens. Two lamps sit on nightstands..."

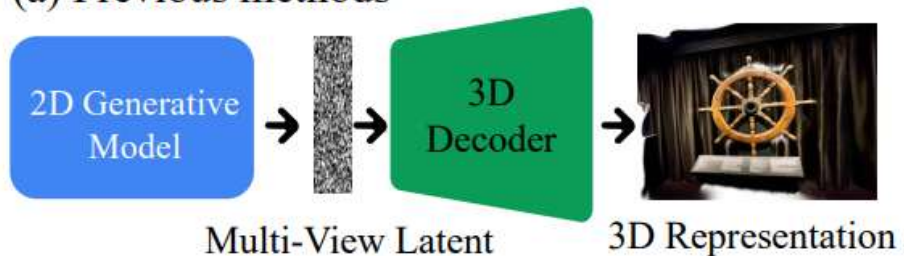


"A gleaming golden trophy with intricate engravings stands too tall to fit inside a small, worn brown leather suitcase..."

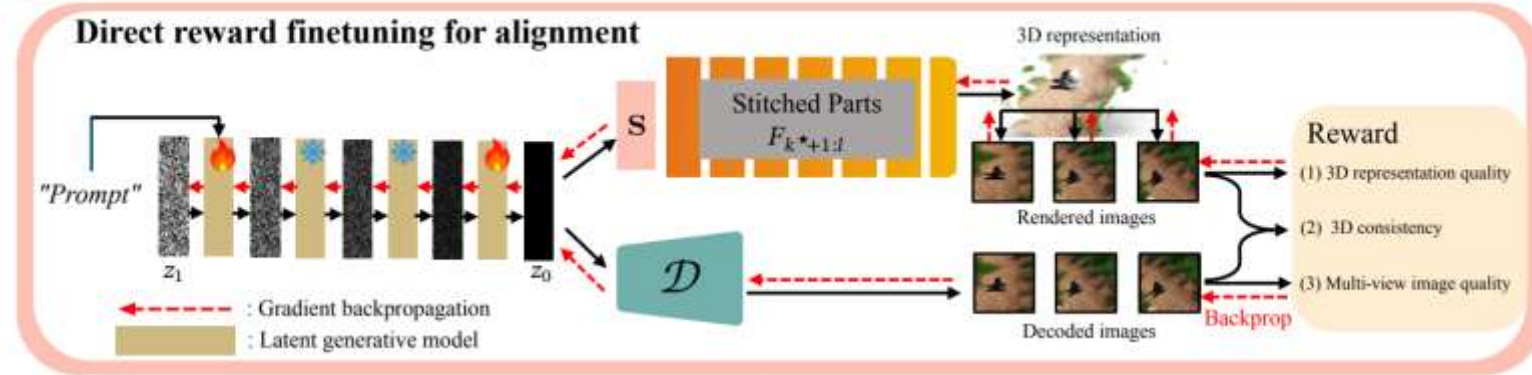
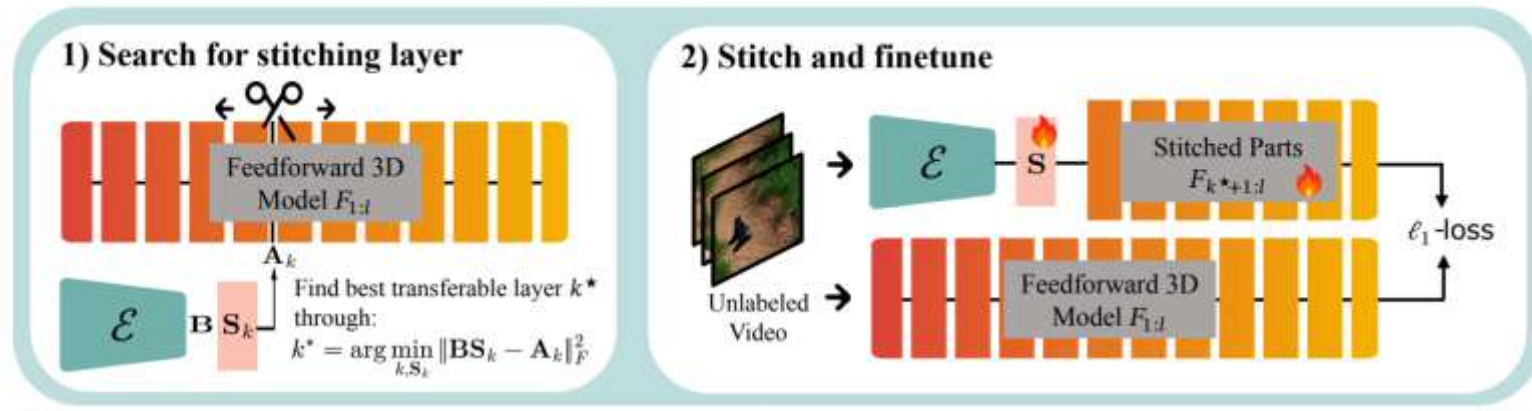
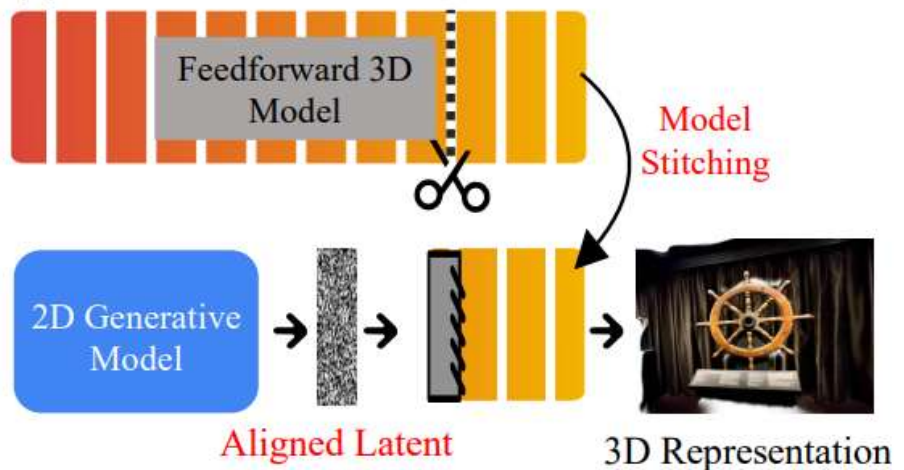
(b) Text-to-Pointmap

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

(a) Previous methods

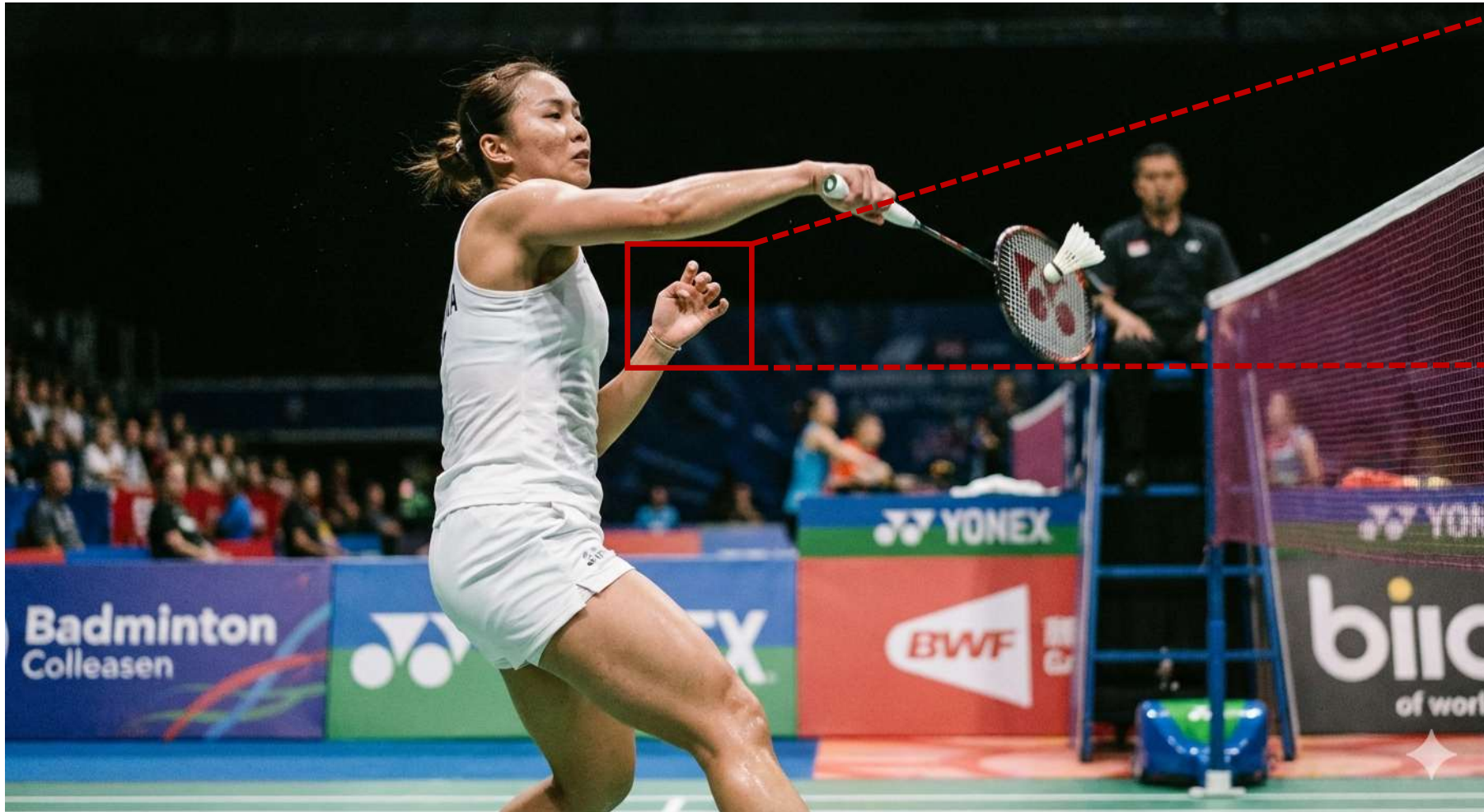


(b) Ours



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.

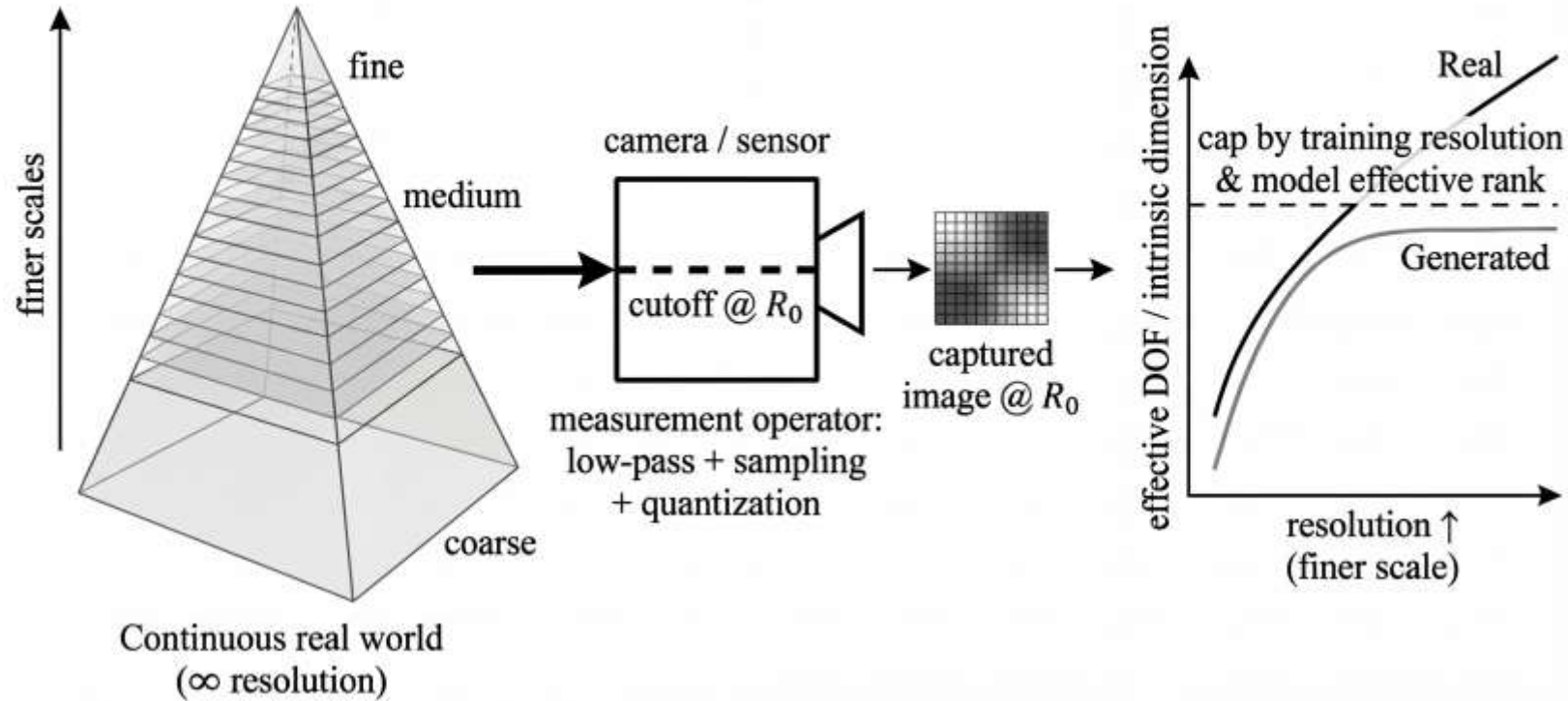
- **Observation.**

- 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?**

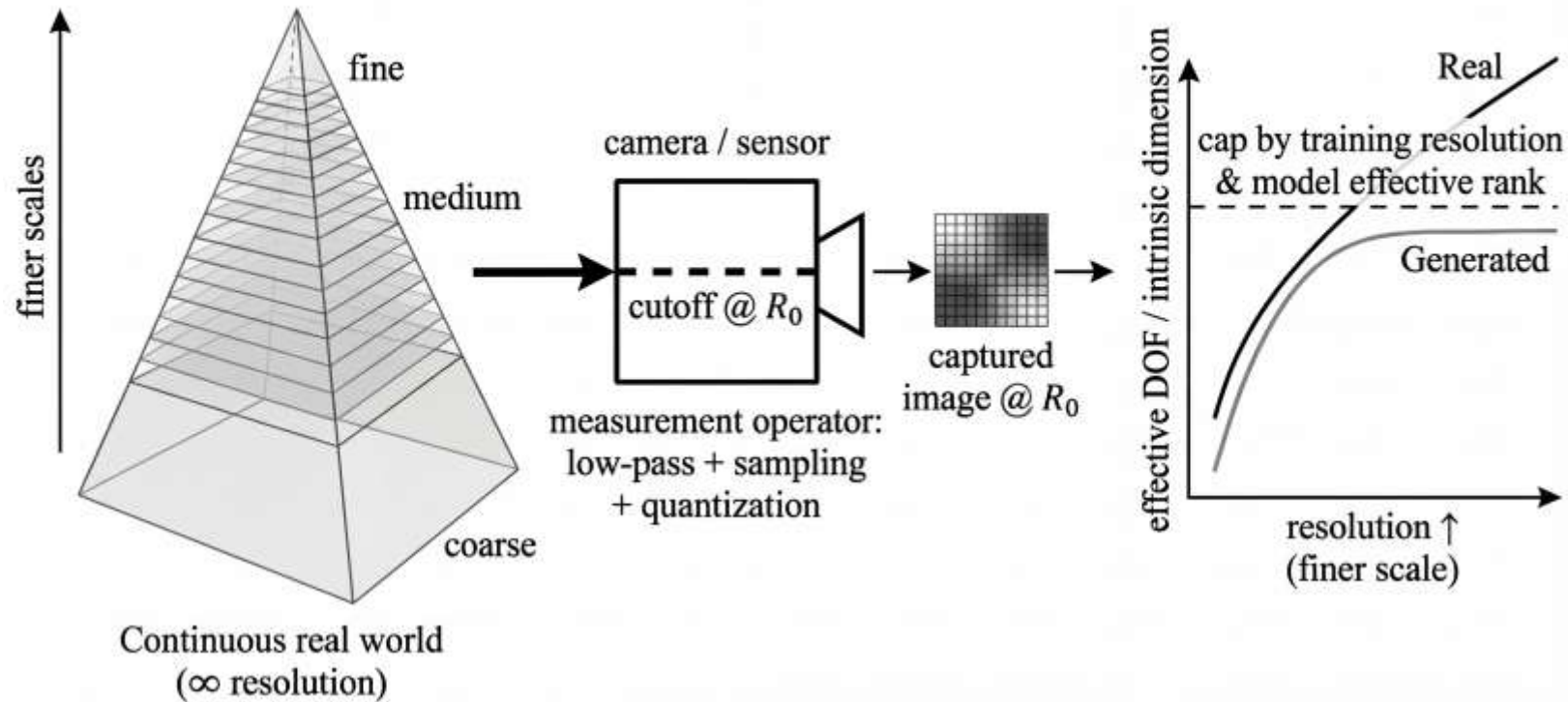
- **Measurement vs. learning mismatch:** the camera provides a finite-resolution *measurement* of a continuous scene (low-pass + sampling). Generators learn a **band-limited approximation** tied to the training resolution.



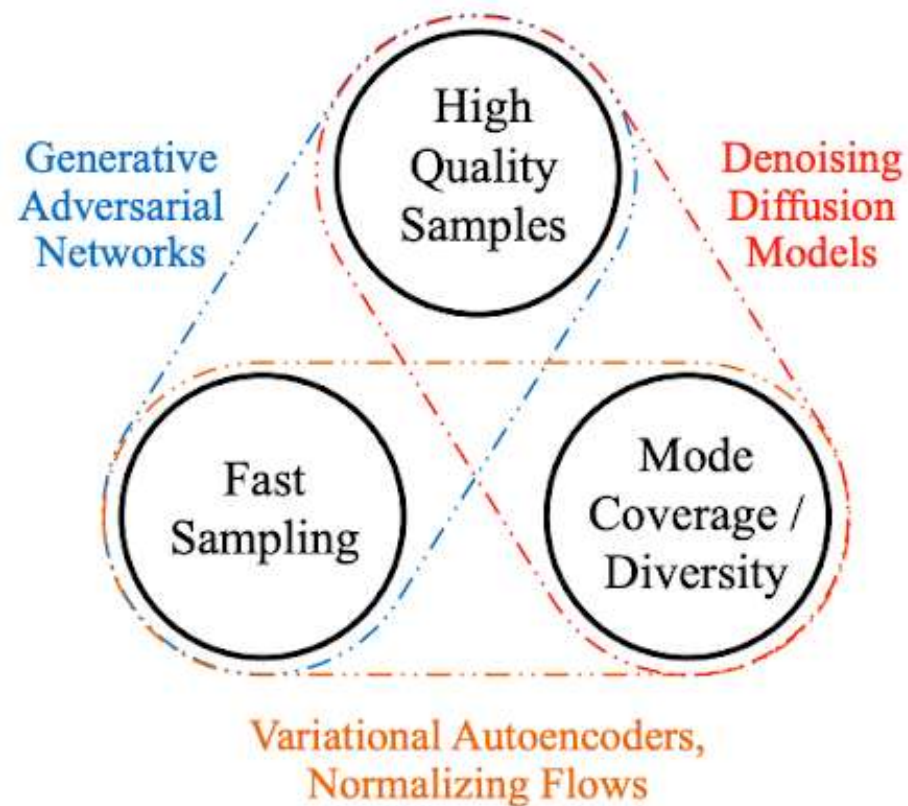
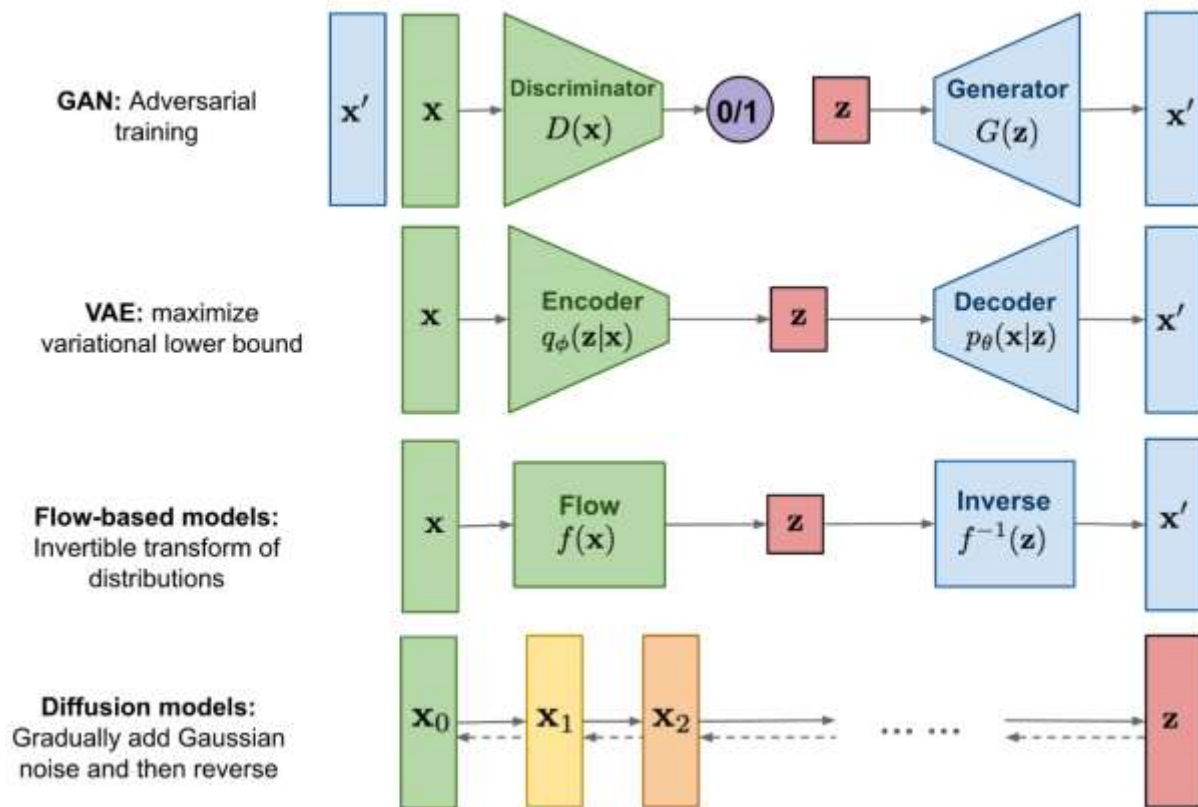
Real vs. Generated Images: A Multi-Scale Geometry Gap

- **Why this happens?**

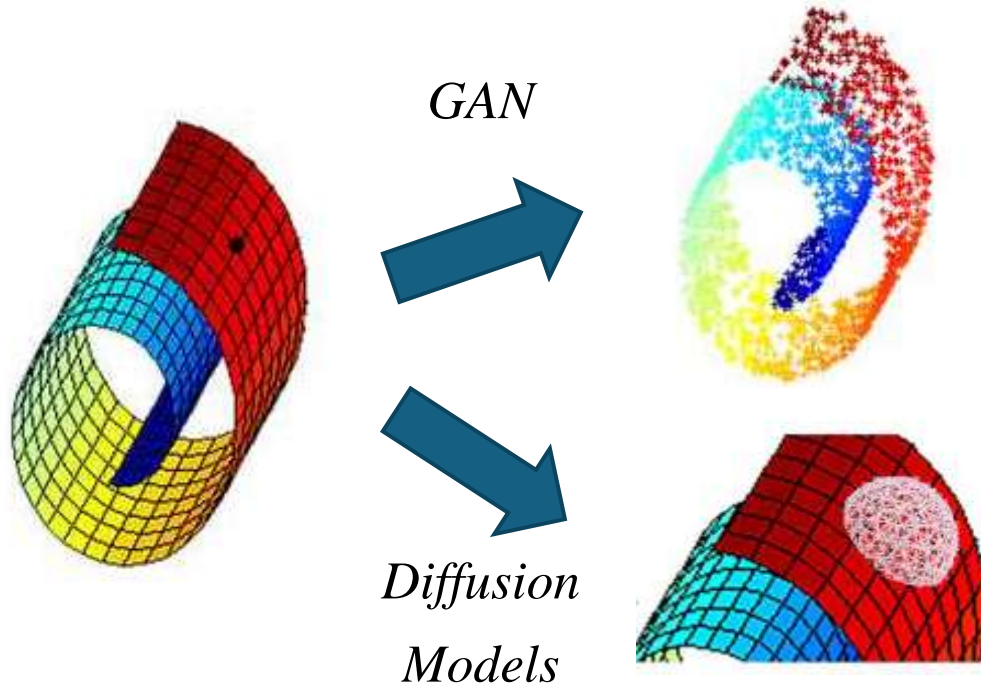
- **Effective DOF cap (model/optimization bias):** the generator parameterization and training objective impose a **bounded effective dimension / thinner manifold**, so extra fine-scale degrees of freedom are either **collapsed** (stable but repetitive) or **fabricated** (detailed but inconsistent).



Impossible Triangles



Diffusion Models' Benefit



Distributional loss

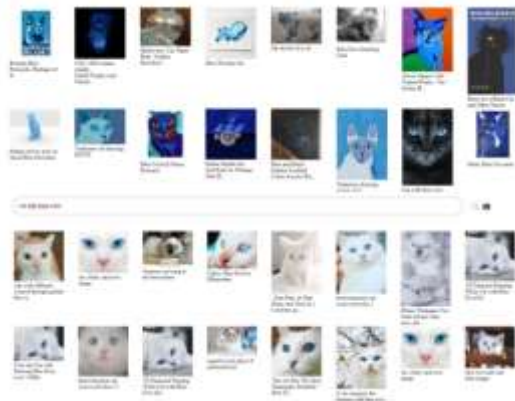
+ Clearer images

- 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!

Diffusion Models' Benefit



Caption generated by GIT base
a man standing in front of a bright light

Caption generated by GIT large
this image may contain clothing apparel human person sleeve long sleeve and man

Caption generated by BLIP base
a man in a black shirt

Caption generated by BLIP large
a man in black shirt standing in front of a triangle

Caption generated by VIT+GPT 2
a man in a black shirt and a white shirt

SD: "A photo of Musk"

Caption: "A man in a black shirt"



Caption generated by GIT base
digital art selected for the #

Caption generated by GIT large
a portrait of [unused] by [unused]

Caption generated by BLIP base
a painting of a man with a mustache

Caption generated by BLIP large
a close up of a painting of a man with a red shirt

Caption generated by VIT+GPT 2
a man with a beard and a cartoon character on his face

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?



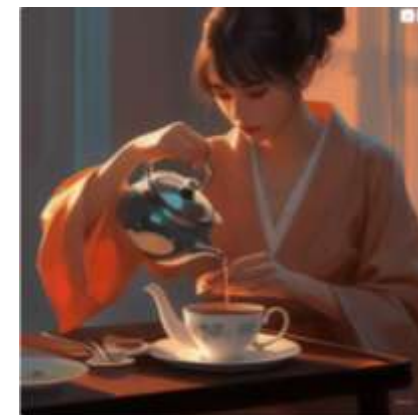
<https://www.midjourney.com/showcase>

+ 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

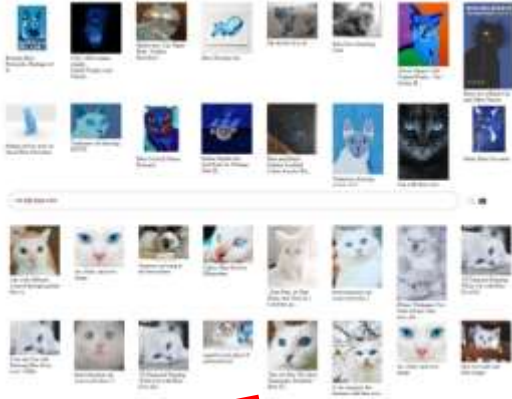
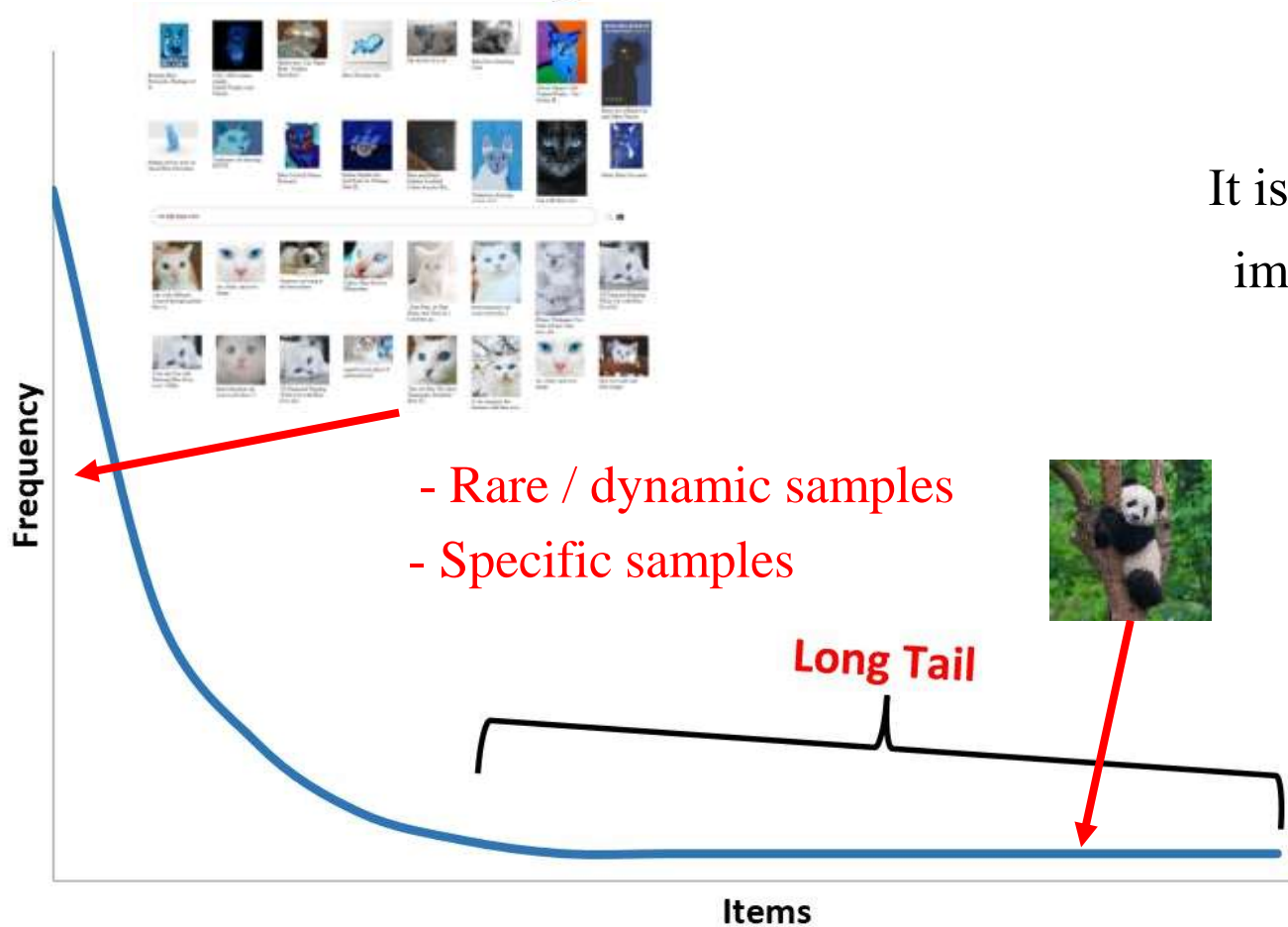


- Physic systems



Are Diffusion Models All You Need?

Long-Tailed Distribution

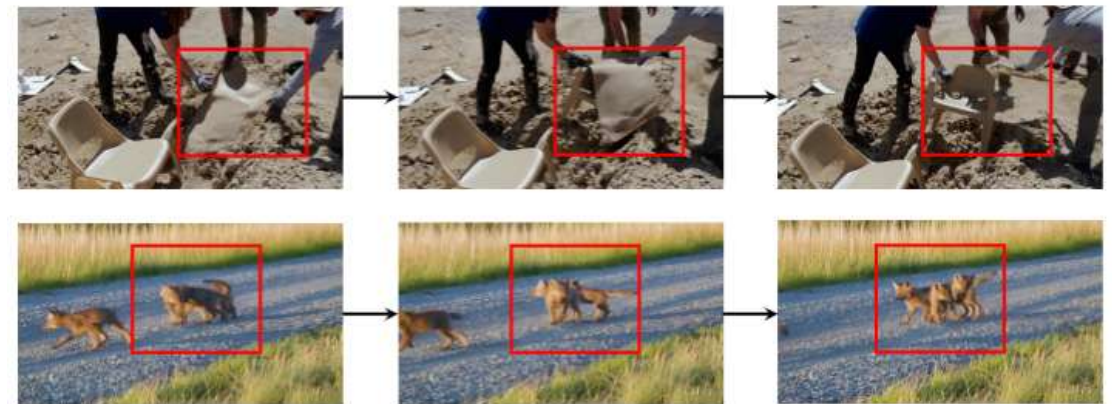


Training Subset	Testing Subset		
	DMs	GANs	Others
DMs	99.7/99.9/99.9	86.4/95.9/89.8	79.3/93.5/84.5
GANs	77.1/83.2/74.3	98.1/99.0/99.6	91.4/97.2/94.6
Others	76.4/77.2/70.1	82.5/96.0/91.2	99.6/99.9/99.9

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?

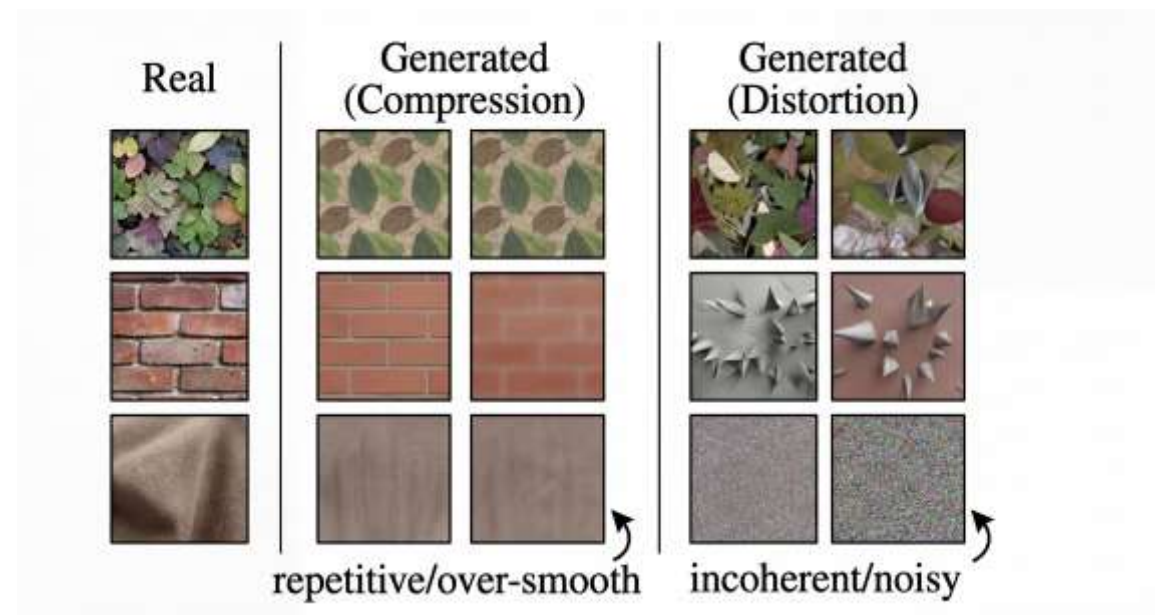
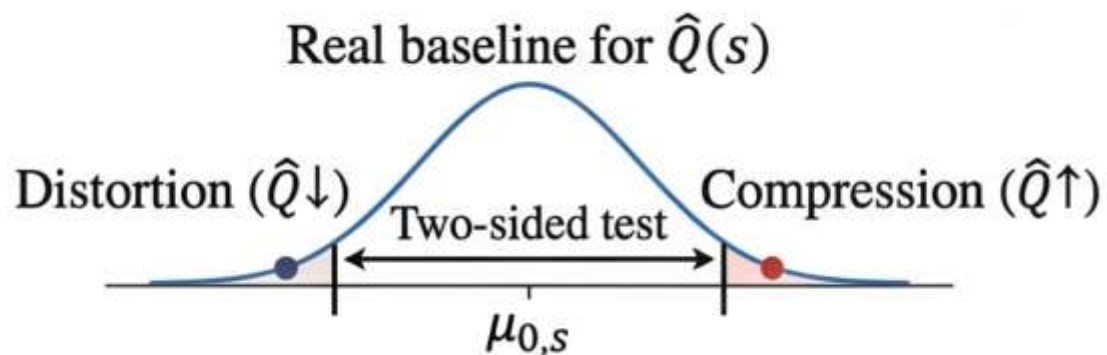


- Comprehensive
- Cross-Modality
- Activity

Detecting generated images with effective dimensions

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

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



Detecting generated images with effective dimensions

(a) Inpainting: Ill-Posed Prob.



Masked Image

Possible Layouts

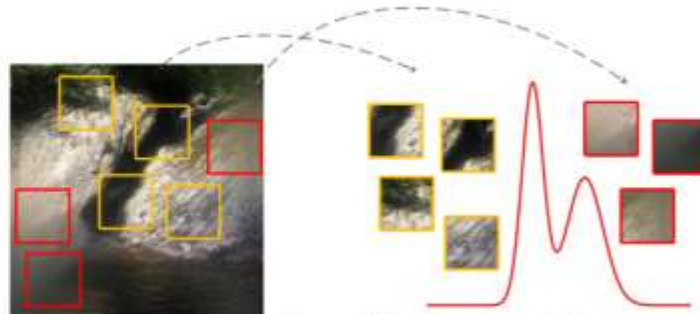
(b) Assessing / Verification:
Ill-Posed Prob.



Masked Image

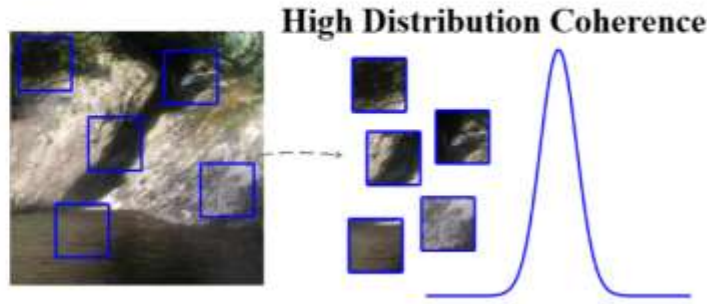
Which is better /
more natural?

(c) Low PHomo Score Images

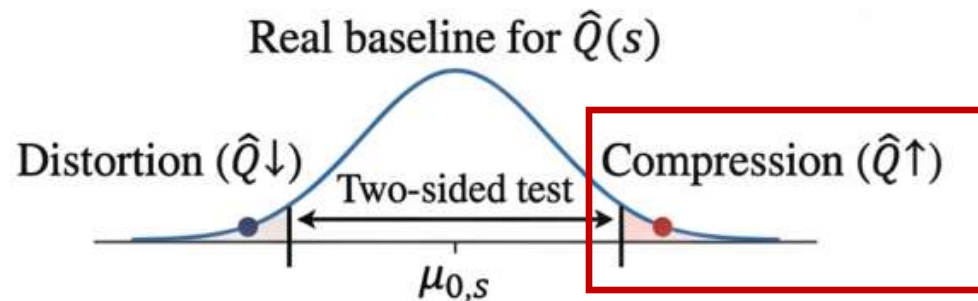


Low Distribution Coherence

(d) High PHomo Score Images



High Distribution Coherence



Real



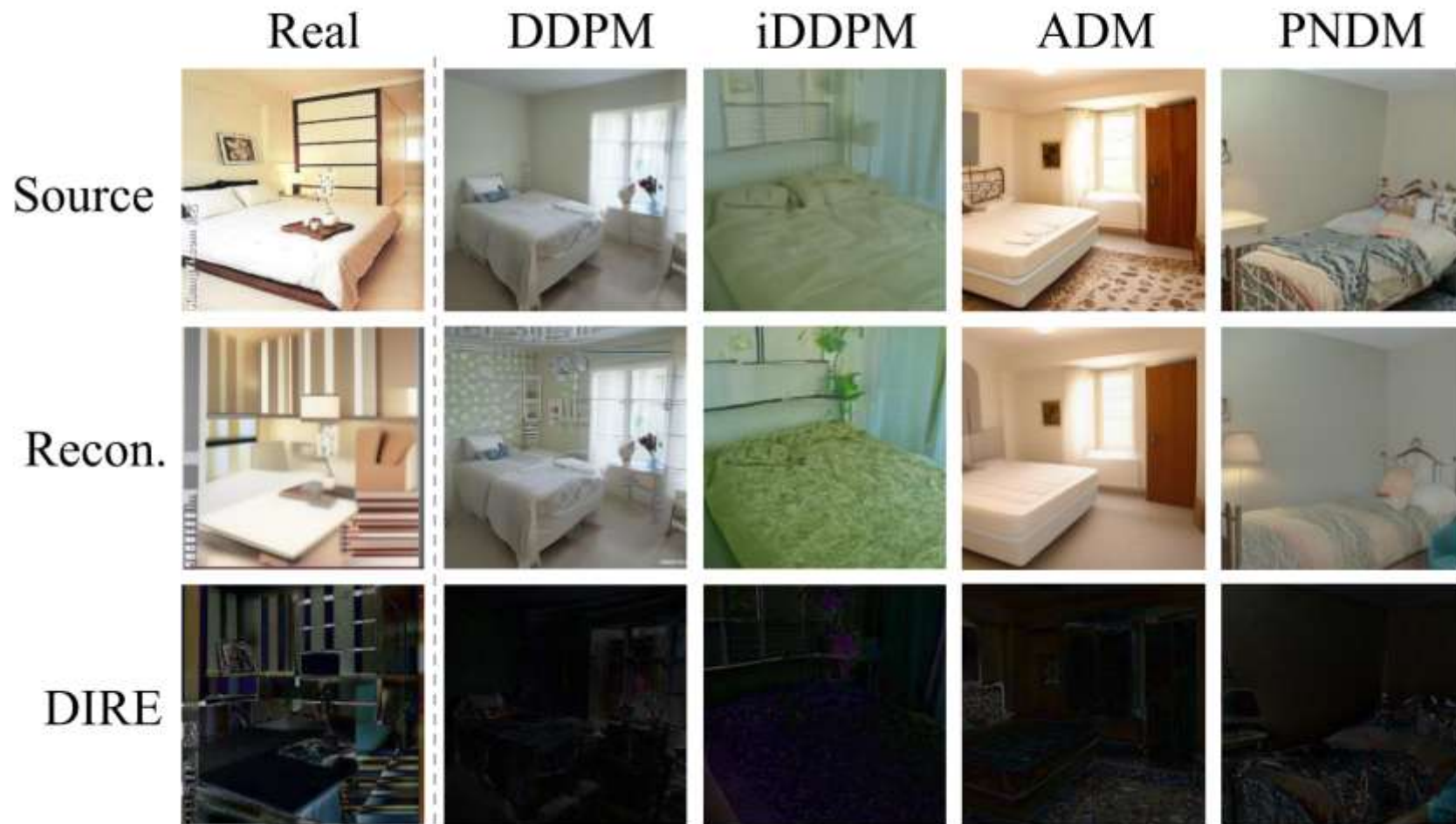
Tail: none

AIGC

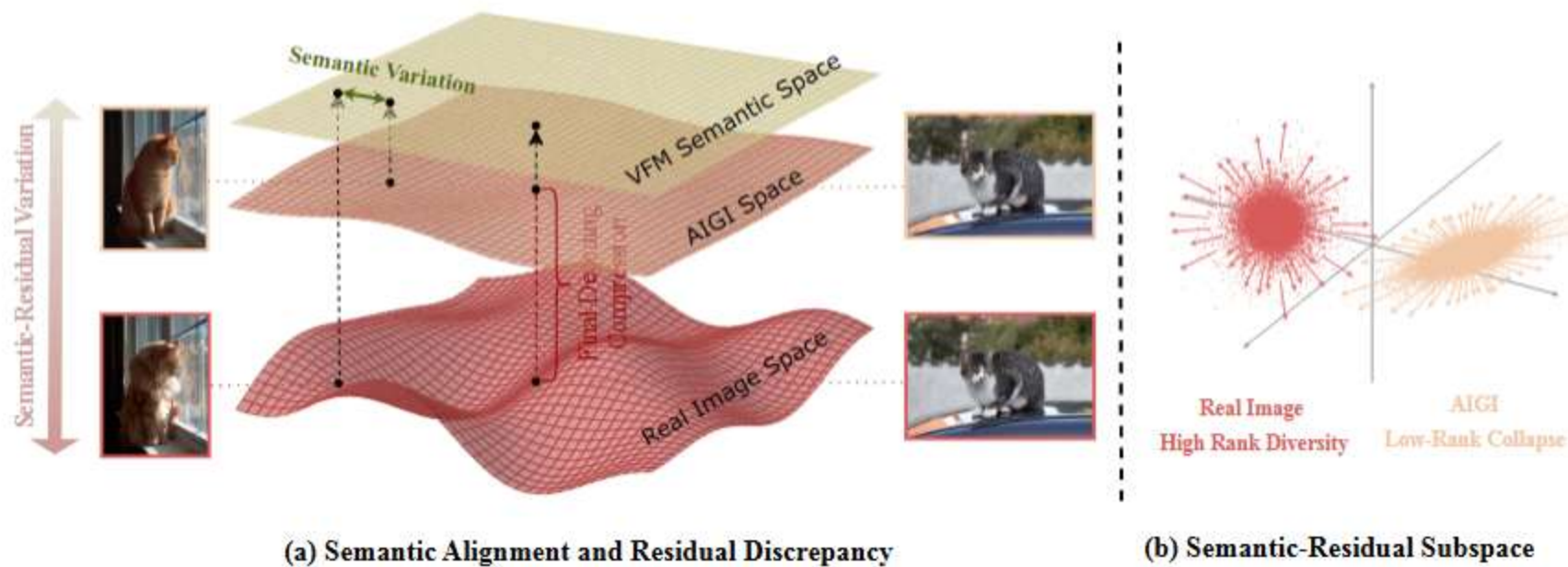


Tail: compression

Detecting generated images with effective dimensions



Detecting generated images with effective dimensions



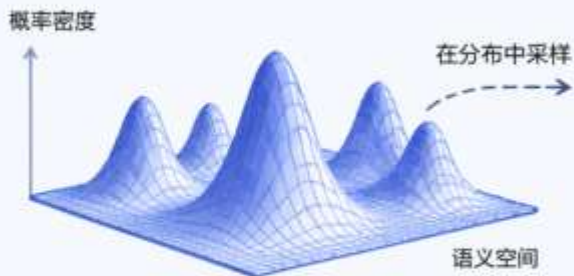
(a) Semantic Alignment and Residual Discrepancy

(b) Semantic-Residual Subspace

生成图像来自有限分布采样，语义自由度更低

“Hallucination”
“Emergence”

生成模型分布 (有 mode, 受温度影响)



分布有若干 mode, 采样集中在这些高概率区域, 可能性 (自由度) 相对更低。

生成图像的典型结果 (桌子上最常见的是花瓶)



温度 (Temperature) 的影响

温度越低, 分布越尖锐, 输出更趋同; 温度越高, 多样性增加, 但仍受分布限制。

VS.

真实世界的可能性 (几乎无限)



桌子上的内容与桌子本身无关, 理论上可以是任意物体, 真实世界的可能性几乎是无限的。

真实世界中桌子上可能出现的任意物体



核心观点

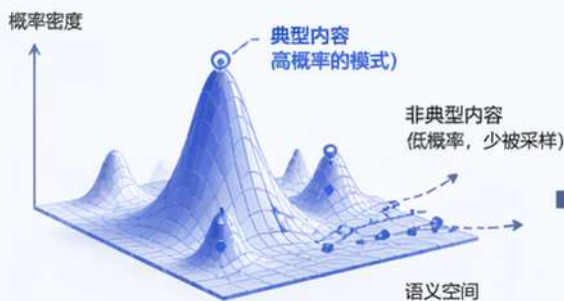
生成图像受限于训练分布和采样机制 (mode + 温度), 语义自由度/可能性更低; 而真实世界的可能性几乎无限, 桌子上出现任何物体都是合理的。

为什么生成图像会出错？以及能力增强如何带来“涌现”与更高自由度

生成模型在分布中采样 → 倾向典型内容 → 自由度受限 → 容易出现破碎/结构性错误；能力增强后突破典型 → 自由度提升 → 整体能力跃升

“Hallucination” “Emergence”

1. 分布中的采样：偏向典型内容



模型训练目标：拟合数据分布 $p_{data}(x)$ ，采样时更可能落在高概率的 mode 附近。

2. 典型性 → 自由度低



3. 自由度受限 → 易出错 (破碎/结构性错误)



! 根本原因：模型在模式内学习到的“典型共现”和“结构先验”很强，但对模式外/长尾组合与精细几何约束的刻画不足 → 一旦要求非典型内容或精细结构，就容易出现破碎或结构性错误。

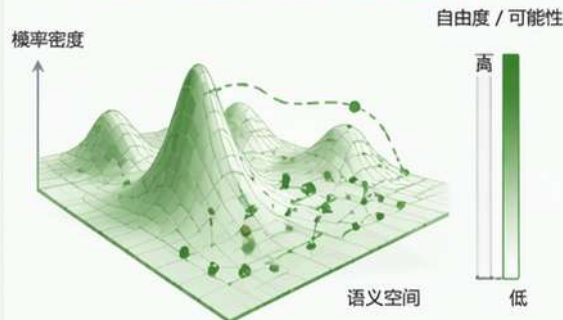
4. 能力增强 → 突破典型 → 自由度提升 → 能力涌现

结果：不止是更高可能性，更是整体能力跃升

模型能力增强 (数据/规模/采构/训练目标改进)



突破典型：探索更多可能性 (更高自由度)



核心观点

生成模型的进步不仅仅是“更像”，更是“能生成更多合理的不同内容”：从只能生成典型结果（自由度低，易出错），到能生成多样且合理的结果（自由度高，整体能力涌现）。



早期模型
(典型、局限)

中等模型
(扩展、改善)



强大模型
(多样、自由、可靠)

“Hallucination” “Emergence”

“由 ChatGPT 内置图像生成工具根据文本提示生成。”

曹操
直播中
10.8万 观看
关注

人气榜第1名

何以解忧?
唯有
古井贡酒!

北方的狼
送小心心 x66

江东小霸王
送赞 x88

谋定天下: 丞相雄才大略 🍌🍌🍌

宁教我负天下人: 丞相威武!

月明星稀: 这酒真不错, 回购多次了

汉室宗亲: 支持丞相, 匡扶汉室!

风起: 已下单, 坐等美酒 🍷

清风徐来: 古井贡酒, 名不虚传!

老兵不死: 丞相带货, 必须支持!

说点什么...

热卖中

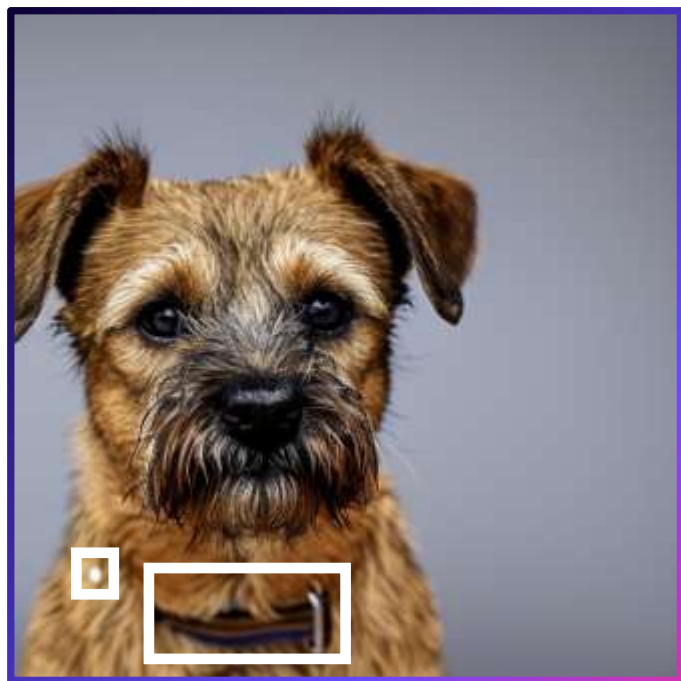
古井贡酒 年份原浆20
浓香型白酒 52度 500ml
官方正品 破损包赔 假一赔十

¥699

立即购买



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。

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。

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。

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

P1 (Defect Query):

...features of AI-generated / real images?
 {Features of AI-generated / real images}
 Sample A Sample B real or generated...?

[Sample A]...shadows are correct...no evident stitching errors...background is properly blurred...**Likely real.**
 [Sample B]...reflections on the diamonds appear slightly exaggerated...unrealistic reflections and perfect surfaces...**Generated.**

P2 (Regional Analysis):

Sample A Sample B ...focus on the main object as ROI A ROI B provided...

[Sample A]...highly reflective pupils...fur lacks depth...unnatural textures... **AI-generated.**
 [Sample B]...diamond reflections behave realistically...skin texture...highly detailed, with visible pores and nail ridges... **Likely real.**

P3 (Common Sense Reasoning):

...analyze Sample A Sample B ...Consider:
 1- Physical proportions...
 2- Spatial relationships...
 3- Natural world physics and rules...
 4- Other logical inconsistencies...

[Sample A]...dog's anatomy appears correct...no extra limbs...lighting exhibits hyper-realistic characteristics...Lacks subtle imperfections...**Likely generated.**
 [Sample B]...diamond reflections and metallic finish appear plausible...text appears slightly out of place...lighting and shadows align correctly...**Likely real.**

P4 (Few-Shot):

Exemplar A Exemplar B ...real or generated? Why?
 {preset reason}
 Sample A Sample B real or generated...?

[Sample A]...a cartoon-like dog...overly smooth fur, overly defined reflections, and uniform patterns...**AI-generated.**
 [Sample B]...maintains an appropriate balance of realism, imperfections, and natural variation...retains natural imperfections...**Likely real.**

P5 (Structural Analysis):

Sample A Sample B ...class_name?
 {class_name} Payload 1
 ...list the key components of {class_name}...?
 {components} Payload 2
 Sample A Sample B ...shows a {class_name}. Show me a list of its key components...
 {components} Payload 3
 ...examine for these structural inconsistencies:
 1- Repetitive patterns that are too uniform
 2- Missing essential components
 3- Incorrect object placement

[Sample A]...No missing facial features...No misplaced shadows...background is clearly not on the same layer...**could be real.**
 [Sample B]...ring edges are unnaturally smooth...diamonds are placed with mechanical precision...too perfect...background shows repetitive pattern...**Likely generated.**


P6 (Stereotype Matching):

Sample A Sample B ...main subject depicted?
 {class_name} Payload 1
 ...what are common stereotypes or patterns of {class_name}...?
 {stereotype_response} Payload 2
 Sample A Sample B ...shows {class_name}... has the following stereotypes: {stereotype_response}...Examine for these attributes... Payload 3


[Sample A]...symmetry in the dog's eyes and face is strikingly perfect...fur exhibits uniformity...depth of field and background blurring seem more **Like AI render.**
 [Sample B]...diamonds exhibit strong sparkle...aligns with expected optical behavior...metal band has smooth but not overly uniform texture...**Likely real.**

Fusion:


...Use the previous responses effectively, combine the information, and make a decision...
 ...make a decision based on the information provided...
 [Sample A] All previous analysis indicate that the image is likely generated rather than real. The strongest evidence comes from...
Final Conclusion: AI-Generated
 [Sample B] After evaluating the image using all six paradigms, there are no clear signs of AI generation...
Final Decision: Real Image




Sample A




ROI A




Sample B



ROI B



Exemplar A



Exemplar B

⚙️ System 👤 User 👉 Assistant 📄 Verdict

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.

AIGI Detection with Reliable Explanations

LLaVACoT (Open Source)



GPT-4o (Proprietary)



Dataset Construction

Fine-tuning MLLM

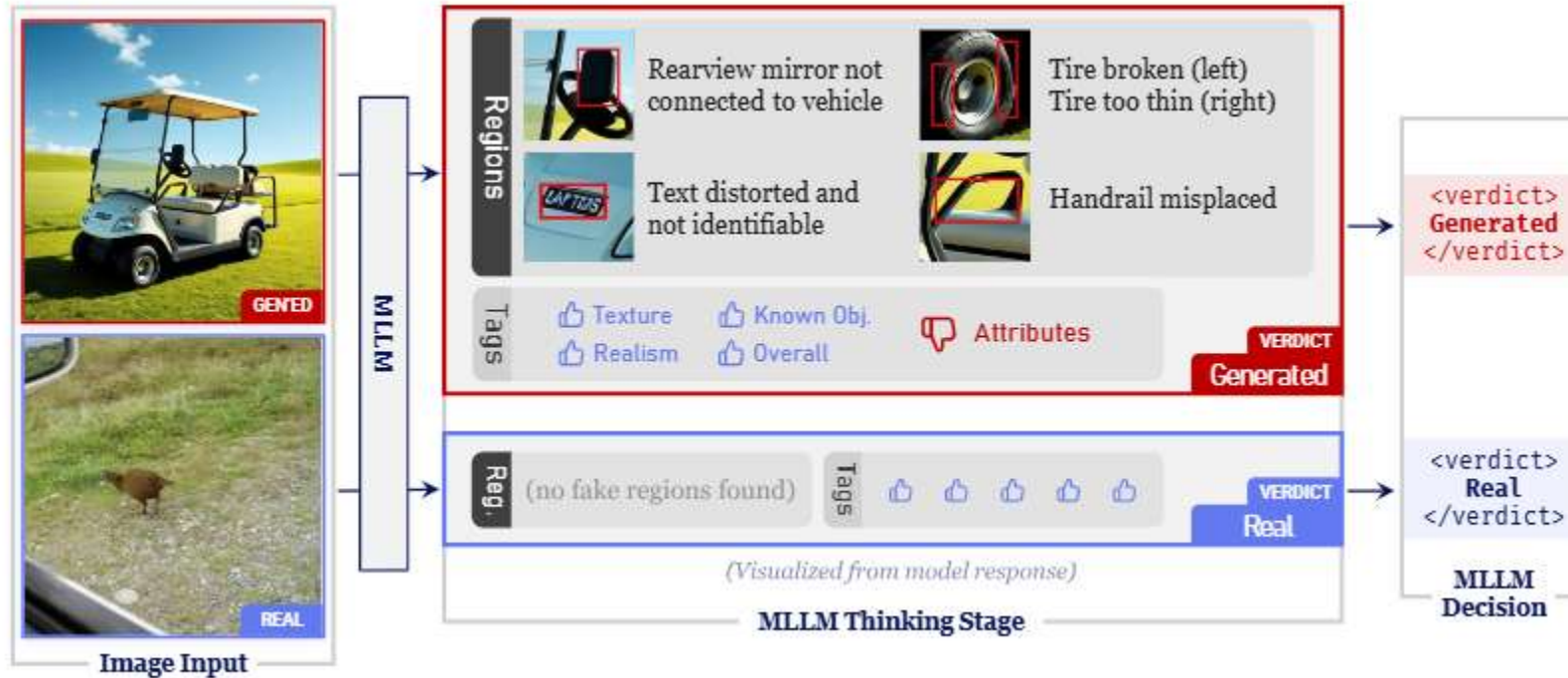
FakeXplainer

Trained on the dataset with

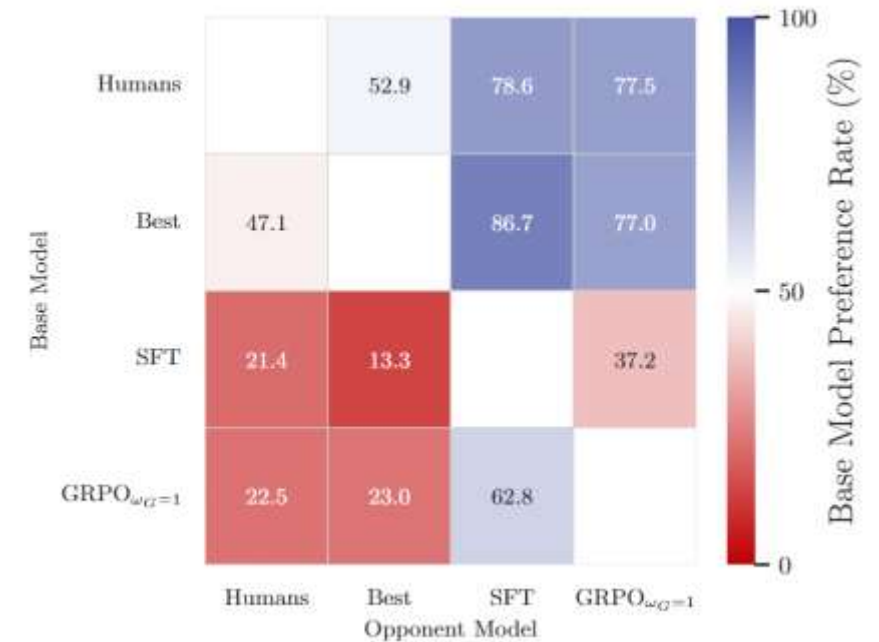
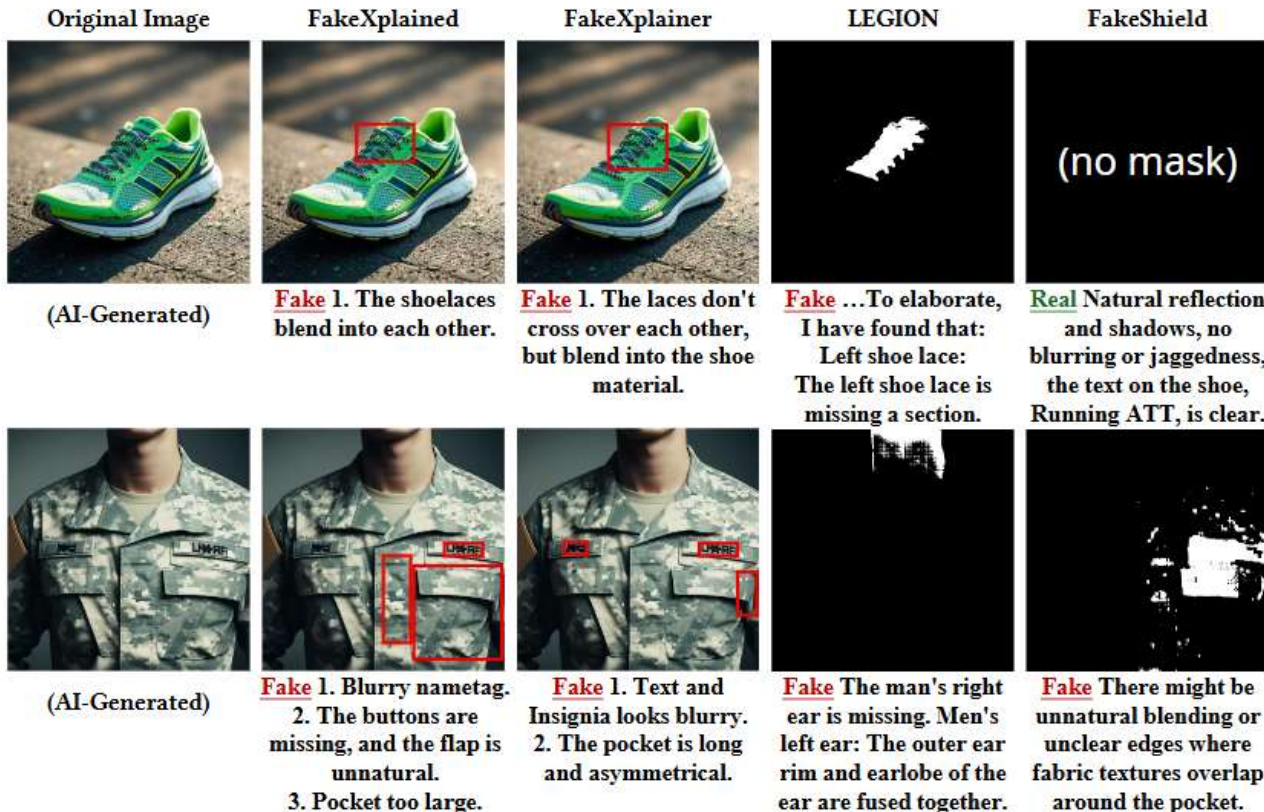
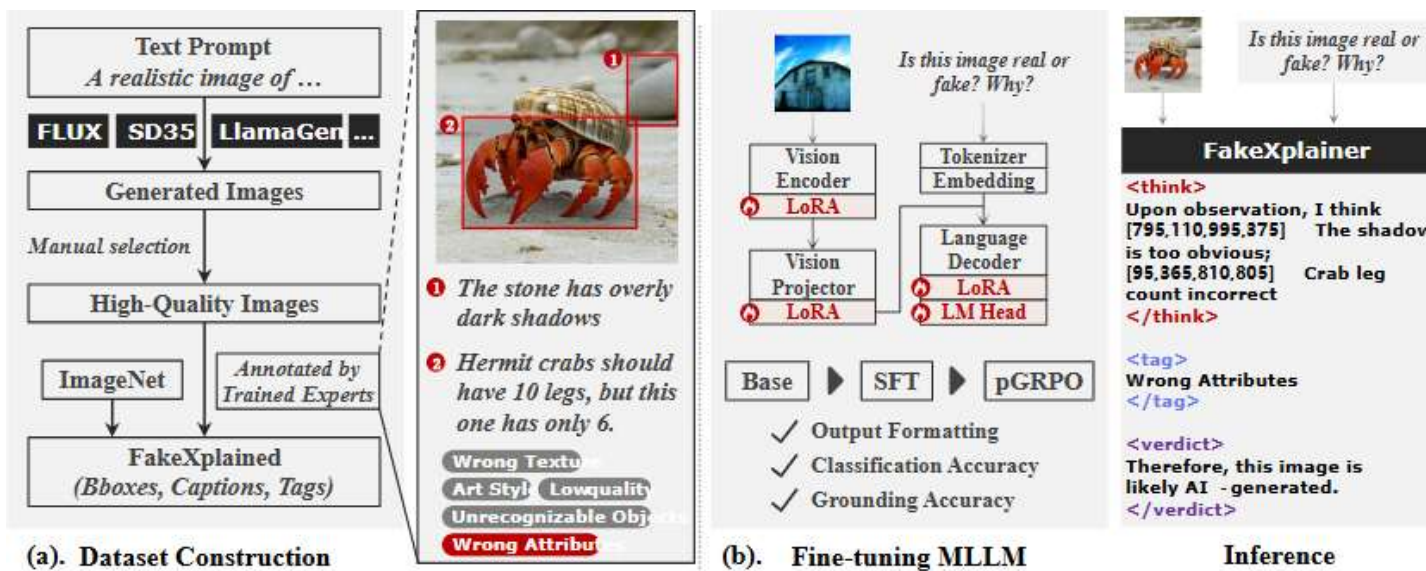
SFT + *Progressive* GRPO

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

AIGI Detection with Reliable Explanations



AIGI Detection with Reliable Explanations



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



Generated by FLUX.1-Schnell

AIGI Detection with Reliable Explanations



AIGI Detection with Reliable Explanations

Query 1: Explanation Generation

S You are an expert in classifying **AI-generated** images and **real** images. Please carefully observe the image that the user provided, decide whether the image is **real** or **AI-generated**. Justify your verdict by listing reasons that support your decision.

U Is this image **real** or **generated**?

A This image appears to be {label}. (pre-filled)
If you wish to understand my justifications, just let me know.

U Yes, please provide your justifications in detail.



A Q1 - Response

This image is likely **AI-generated** due to the **dog's unnatural, human-like pose** and the **seamless integration of text into the scene**.

Additional evidence includes **inconsistent details on the Jenga blocks**, **overly uniform lighting**, and the **synthetic appearance of the background elements**.

Query 2: Spatial Grounding

The user will show you a {label} image with *explanations* on why this image looks {label}. Please annotate *bounding boxes* for key regions mentioned in the justification.

S Your response should form several bullet points, each starts with a *bounding box*, followed by a short caption of *how this region can help determine the authenticity of the image*. Your annotated bounding boxes must be completely based on the rationale. Focus on regions that are most likely to be **fake**.

U This image is {label} because: {previous_response}.
Focus on regions that are likely to have **generative** artifacts or cues.

A Q2 - Response

Based on your detailed analysis, here are some key regions...











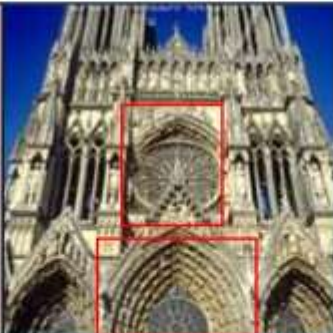

* Dog's Paw Interaction with Blocks (270, 398, 485, 630)

* Jenga Blocks (·)



Roles: **S** System **U** User **A** Assistant

AI-Generated Image Detection with Reliable Explanations

LTE Refined						LTE Unchanged
from <i>Real</i> to <i>Fake</i>			from <i>Fake</i> to <i>Real</i>			
						
The fan hanging above has a wrong geometry	Title mismatch; the towel should not reflect	Extra claw; missing eye	Clear blueberry contour, natural camera blur	The CAD diagram is clean with identifiable text	Branding consistent, correct geometry	
						
Missing petals replaced by leaves	The gates are too far away from each other	Peanuts not cracked and poorly-textured	Precise metallic keyrings; consistent shadows	Symmetrical structure with intricate details	Natural fog and scenery	

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



优化方法

可通过 GAN 框架进行优化。



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。



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）。



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。



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



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

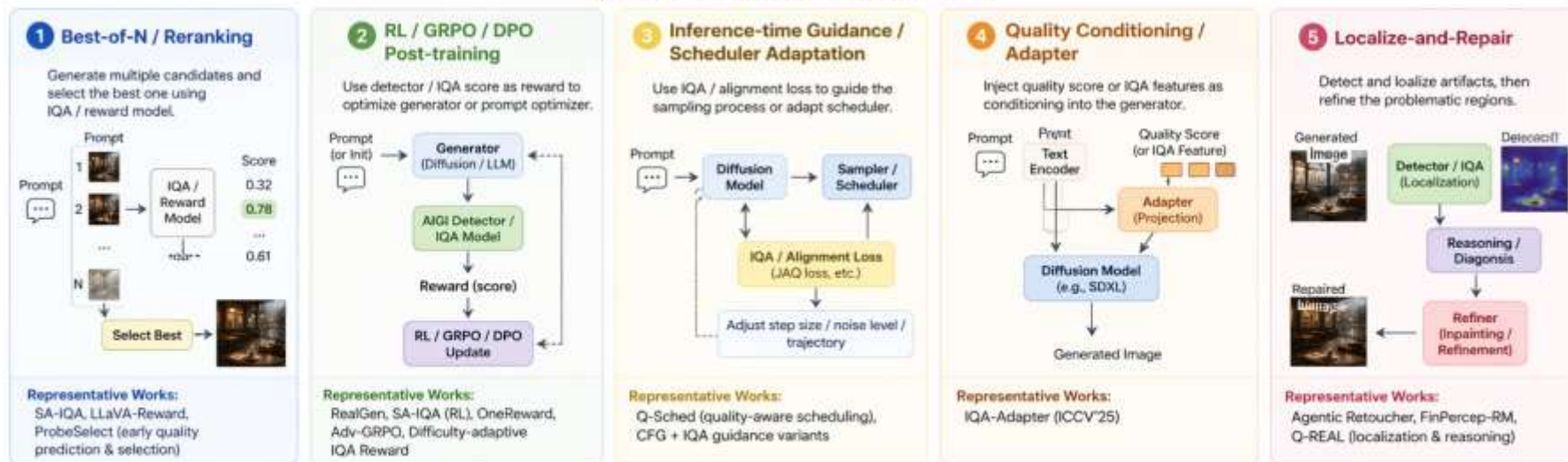
真实性解决“是不是假的”问题，真实度解决“有多像真的”问题。二者互补，共同推动生成模型的进步。

Re-Edit

Unified Pipeline: From AIGI Detection / IQA to Better Generation



Five Feedback Mechanisms in Detail



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.

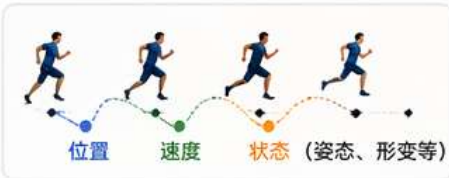
为什么视频生成更困难？—— 需要更高能力，自由度更低，协调更复杂

Video & Music

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：

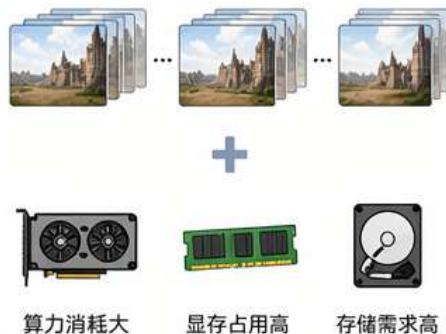
- ✓ 镜头语言设计（景别、角度、运动）
- ✓ 时间线与节奏安排
- ✓ 场景/角色的一致性保持
- ✓ 剪辑点的合理衔接



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，对硬件资源更求极高，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

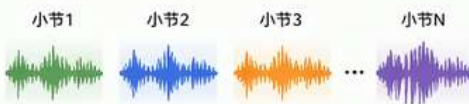
音乐通常以小节 (bar) 为基本单位组织。



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

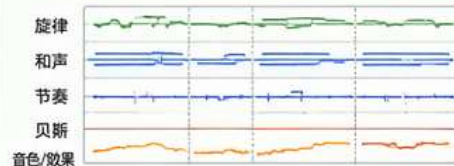
只要和声/节奏/风格匹配，就可以自然衔接。



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）



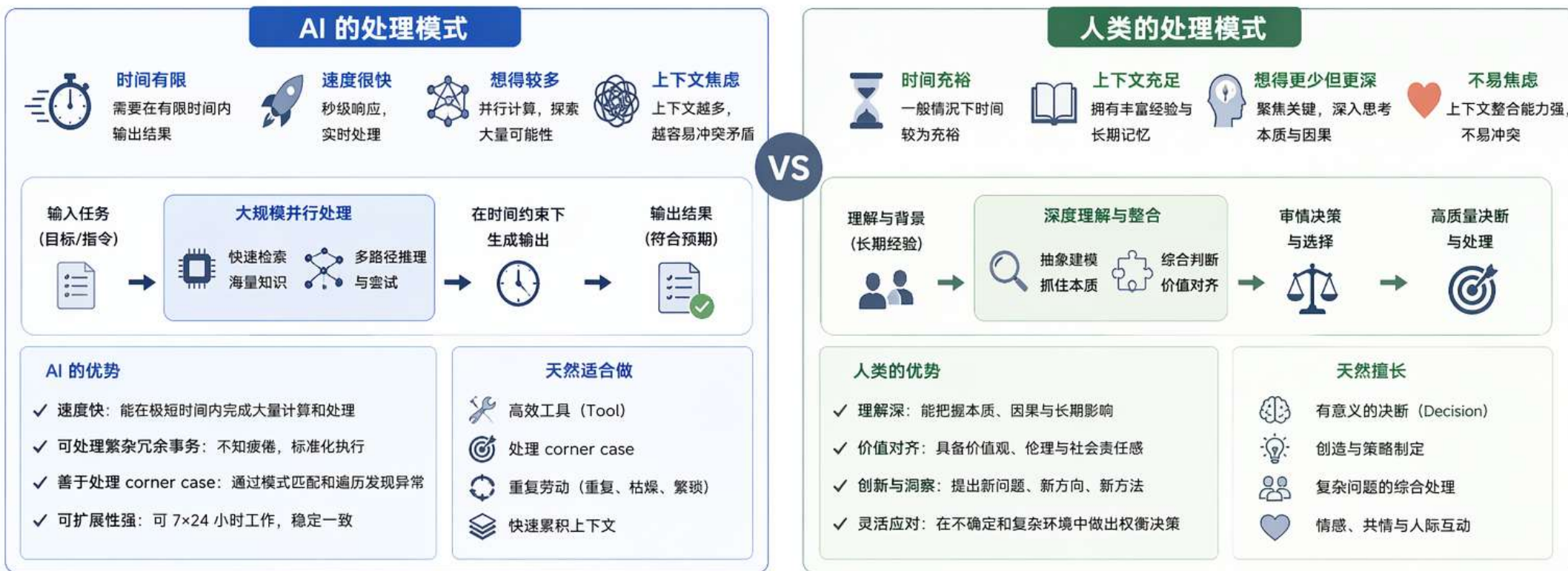
- ✓ 结构清晰（小节/段落）
- ✓ 可拼接、可扩展、易编辑
- ✓ 多轨叠加，信息表达丰富
- ✓ 自由度高，创作空间大



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。

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。



Fin.

互补协同：AI 为人类赋能，人类引导 AI



核心 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。

最终目标



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.