Multi-Modal Generative AI with Foundation Models

Ziwei Liu 刘子纬

Nanyang Technological University

2023

By ~~2027~~, creators won't have to be technical, just creative, thanks to automation tools.

AI-Generated Content

Movie

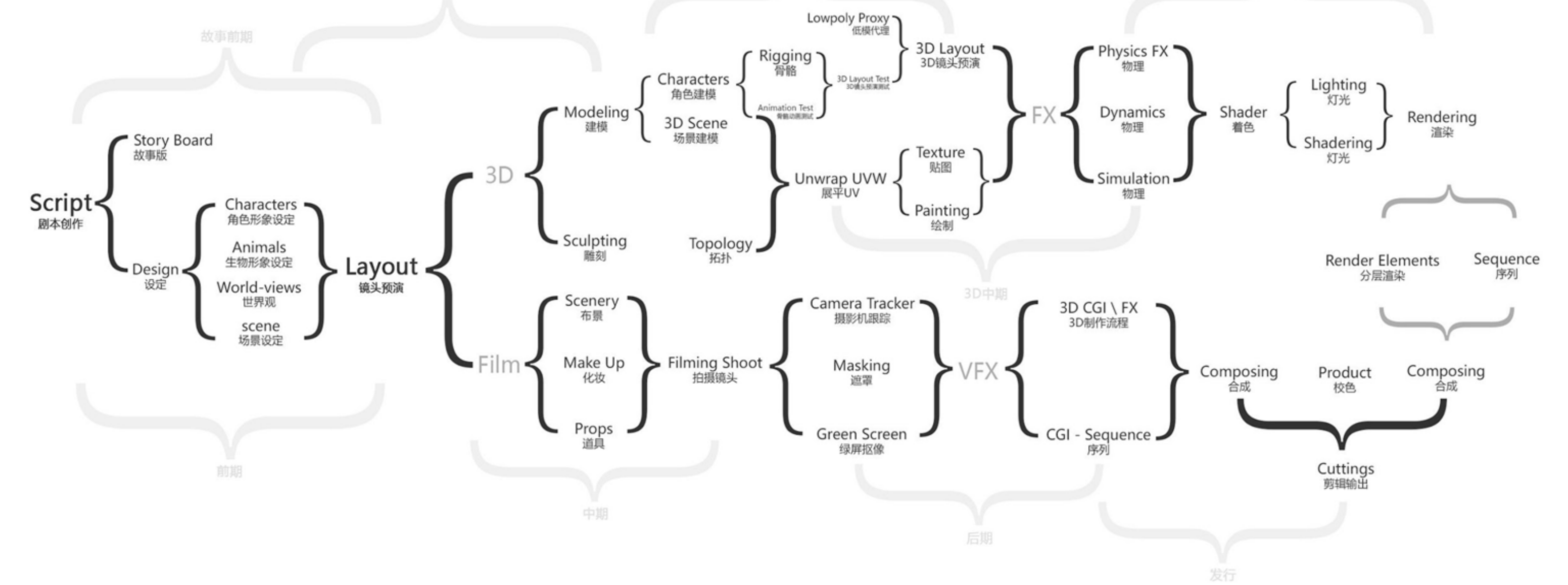
Game

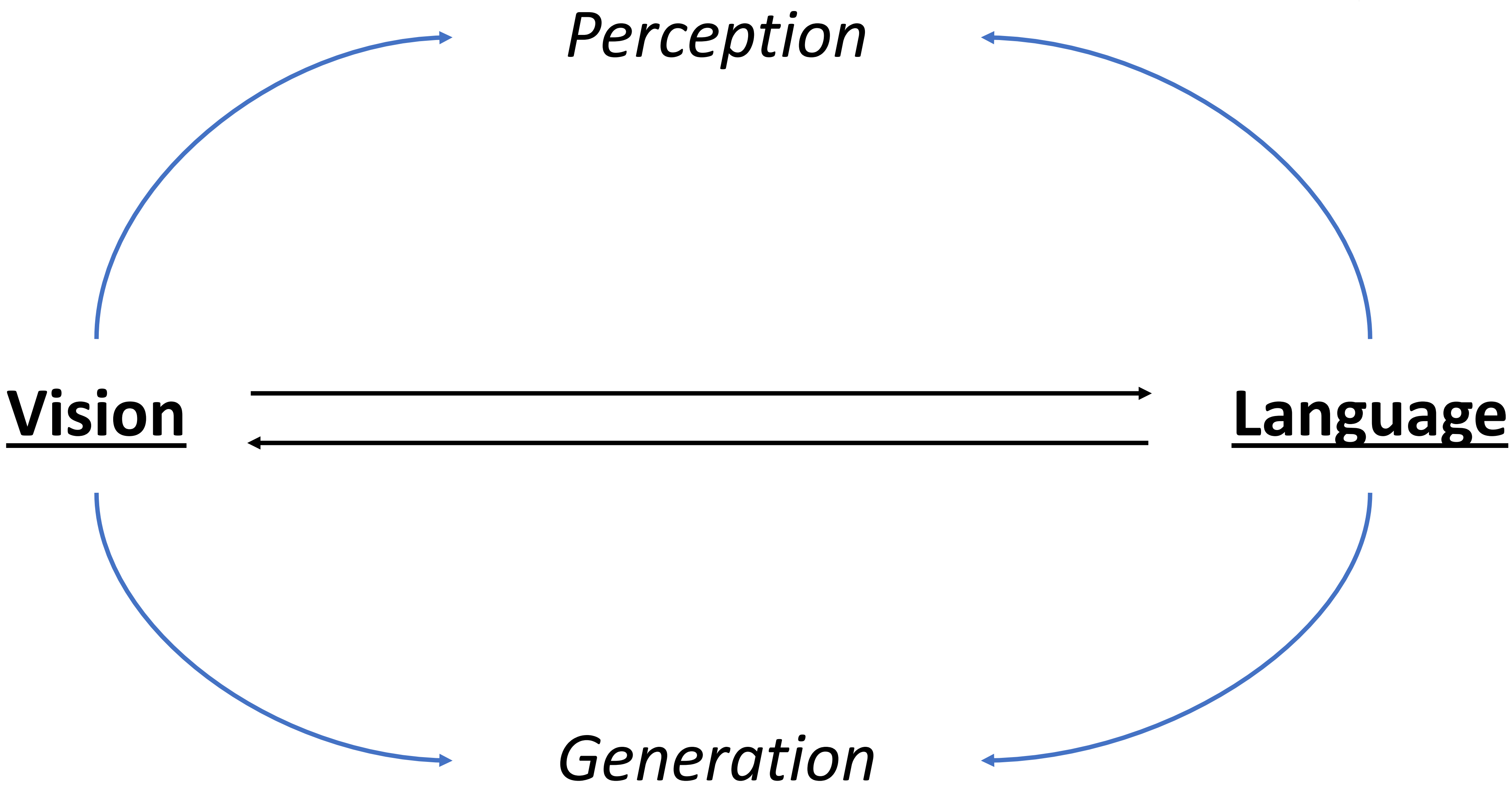
Anime

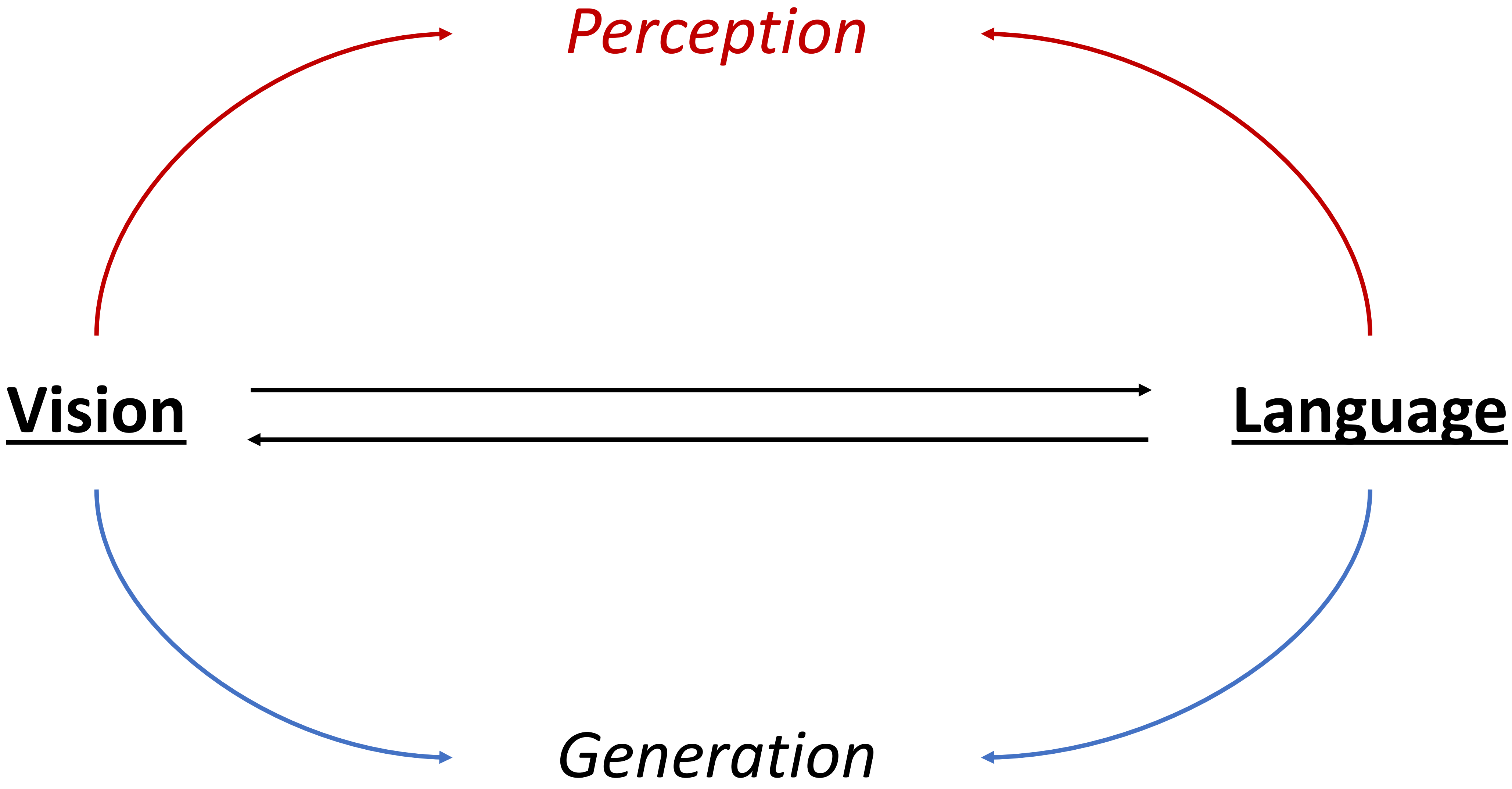
VTuber

Virtual Beings

Creative Industry



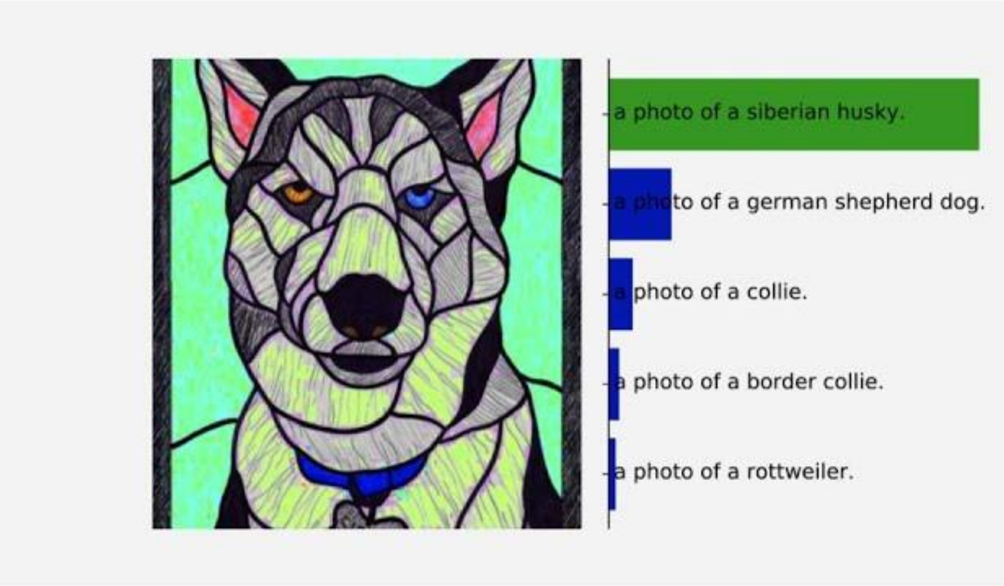

Perception
Vision
Language
Generation

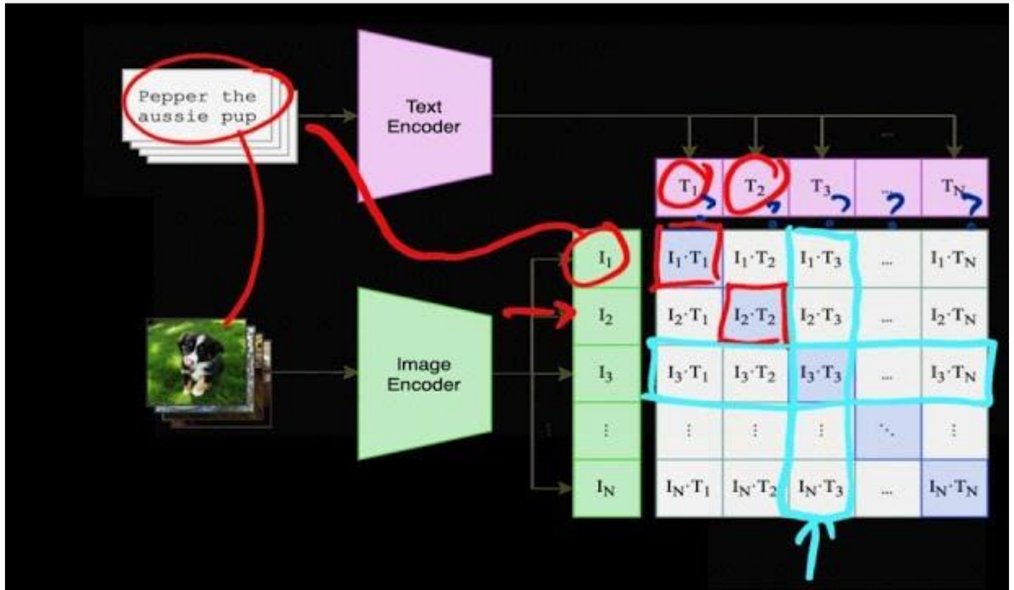

Perception
Vision
Language
Generation

Training, Deployment and Evaluation of Foundation Models

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE





User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

The pathway: From Language Models to Language Assistant

GPT-2	GPT-3	GPT-3.5	ChatGPT
OpenAI	OpenAI	OpenAI	OpenAI

Industrial

Open-source

BERT	LLaMA/T5	Vicuna/Flan-T5	Open Assistant
	 T5	 	
Zero-shot learning	Zero-shot learning In-context learning	Zero-shot learning In-context learning Instruct following	Zero-shot learning In-context learning Instruct following Human alignment

The pathway: From Multi-modal Models to Multi-modal Assistants

Industrial

CLIP

OpenAI

Flamingo

Google DeepMind

Open-source

OpenCLIP

LAION

Zero-shot learning

OpenFlamingo

LAION

Zero-shot learning
In-context learning

The pathway: From Multi-modal Models to Multi-modal Assistants

CLIP

OpenAI

Flamingo

Google DeepMind

Industrial

Open-source

OpenCLIP

LAION

OpenFlamingo

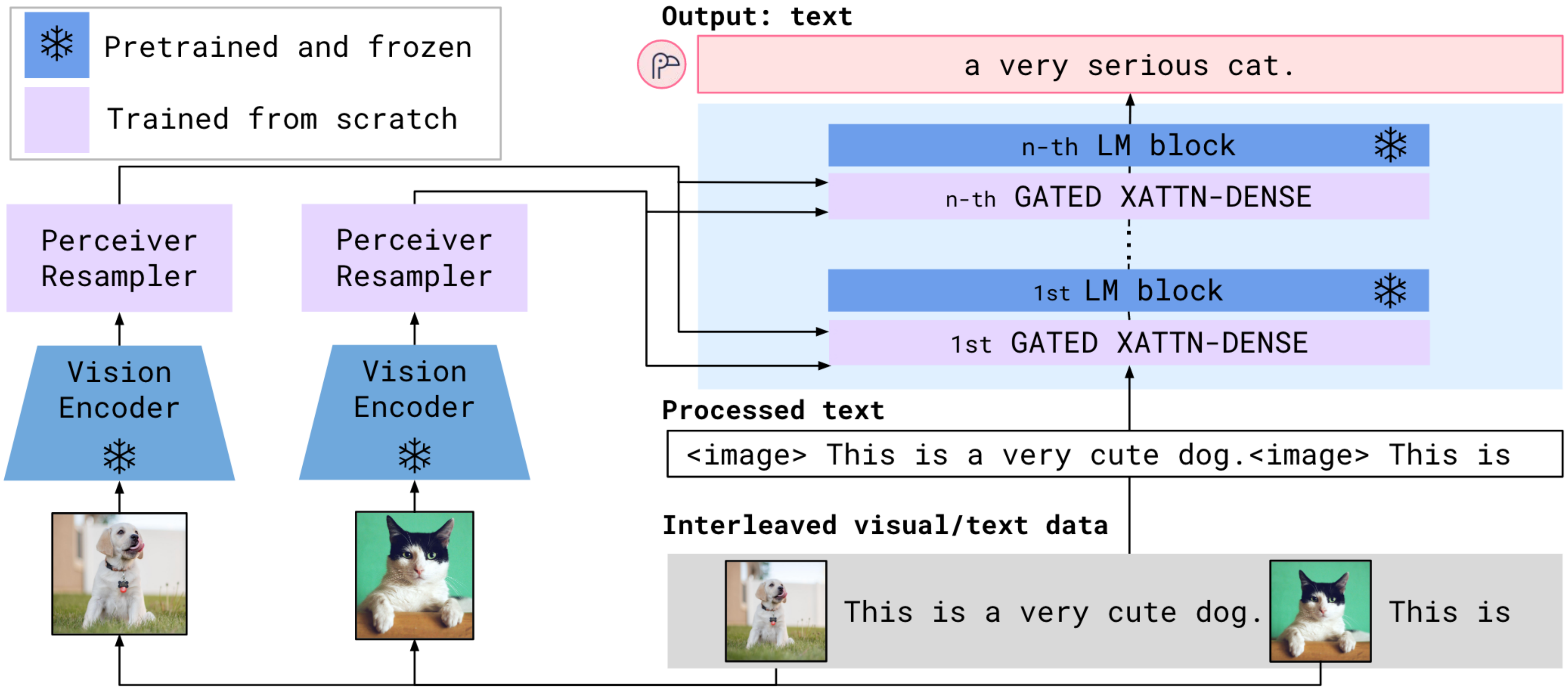
LAION

Otter

Zero-shot learning

Zero-shot learning
In-context learning

Flamingo: a Visual Language Model for Few-Shot Learning



Alayrac et. al. Flamingo: a visual language model for few-shot learning. 2022

Perceiver: versatile to multiple images and in-context examples

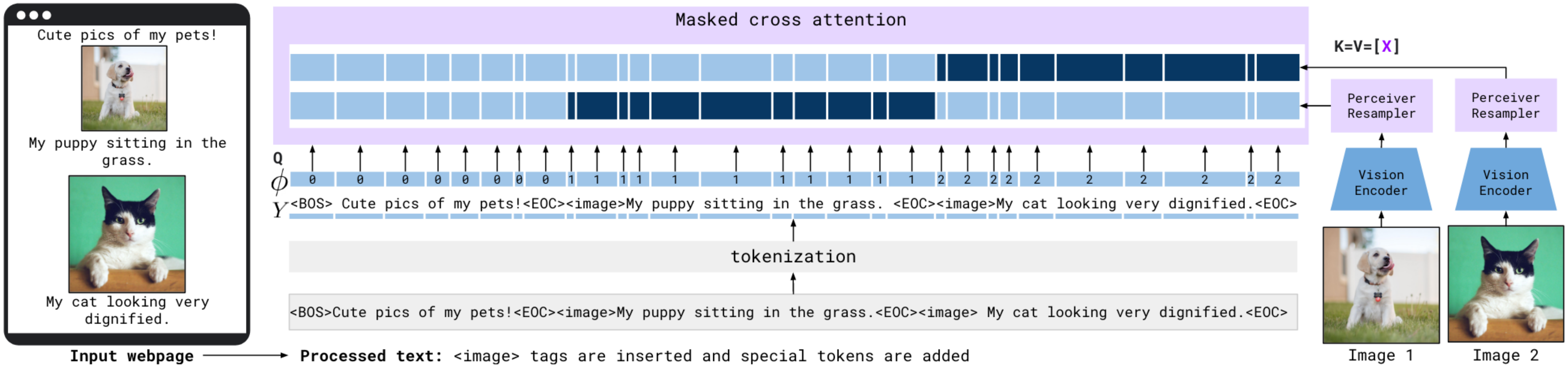


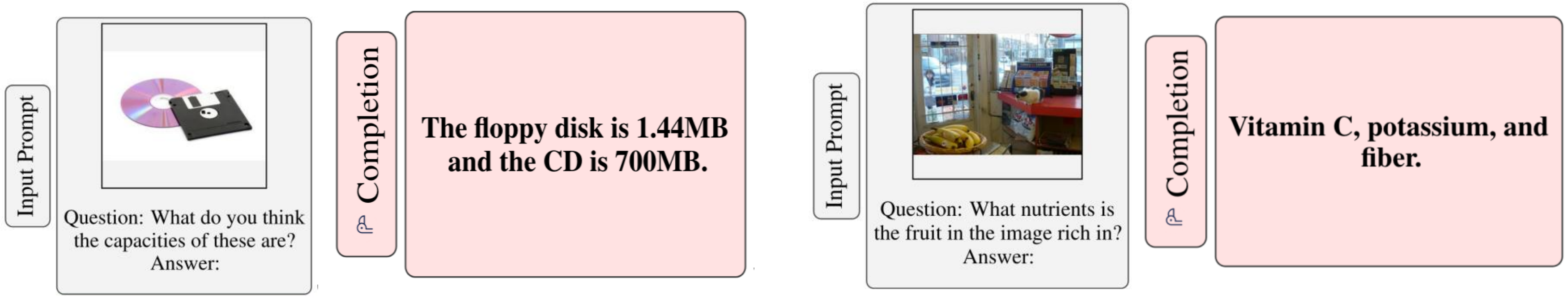
Image-Text Pairs dataset
[N=1, T=1, H, W, C]

Video-Text Pairs dataset
[N=1, T>1, H, W, C]

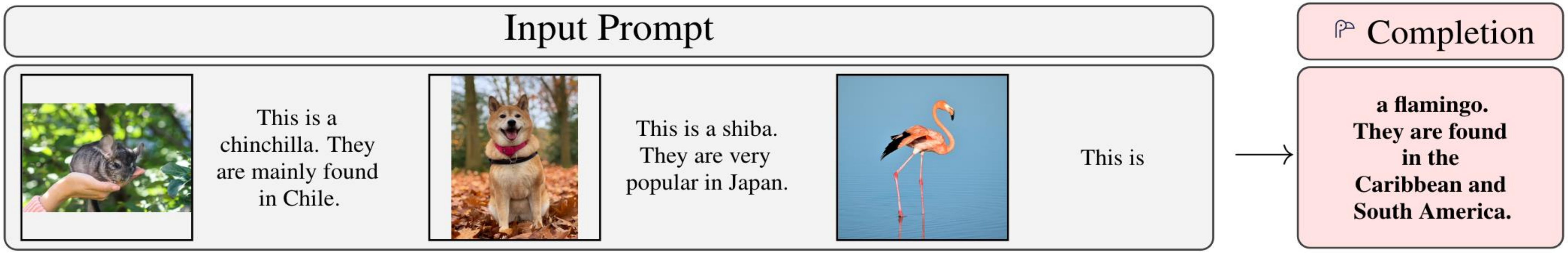
Multi-Modal Massive Web (M3W) dataset
[N>1, T=1, H, W, C]

Flamingo Application

Zero-shot learning



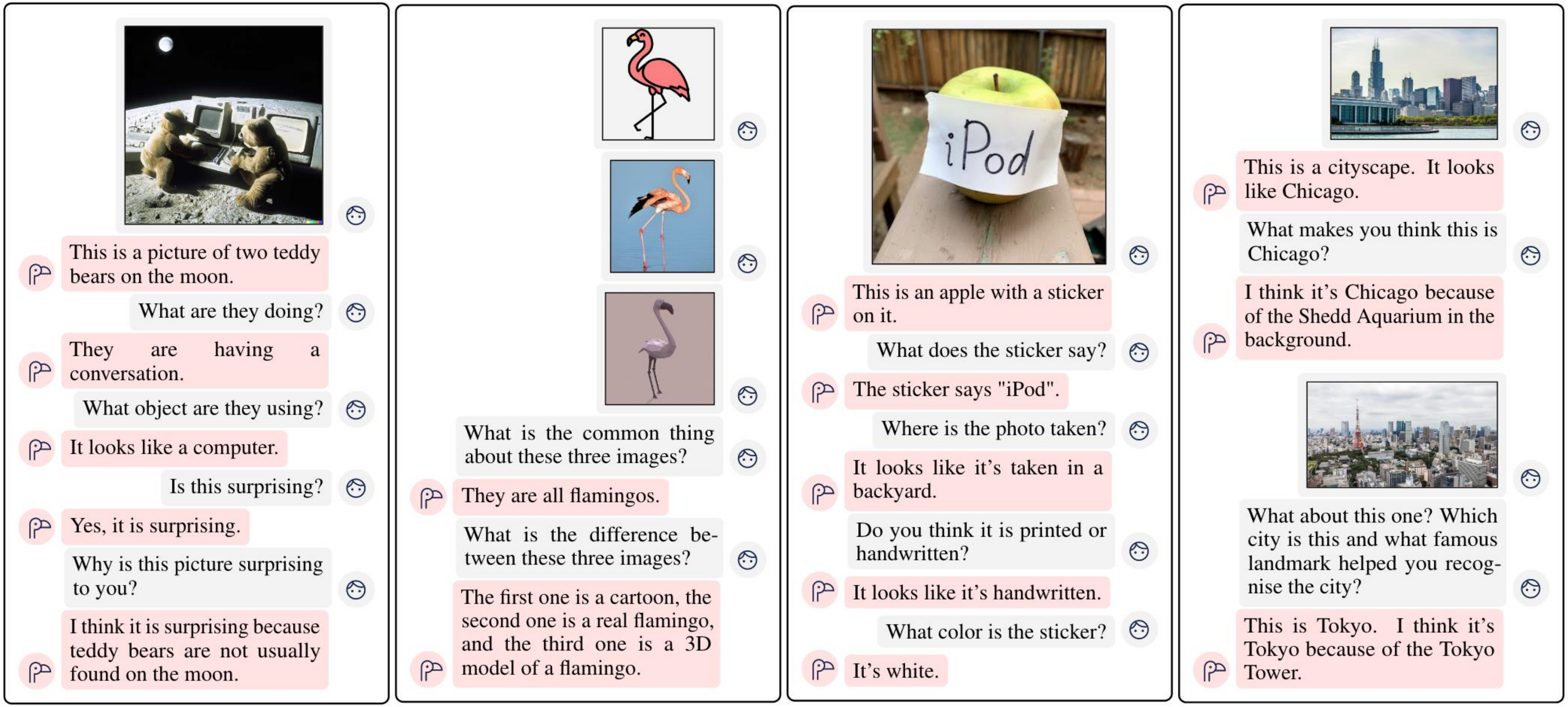
In-context learning



Video Understanding

Flamingo Application

multi-image visual dialogue



Flamingo ≠ Multi-modal Assistants

OpenFlamingo simply completes the next reasonable sentence.

e danger of this sport?

OpenFlamingo*: What is the danger of playing baseball? What is the danger of this sport? What might be the danger of this sports?

Flaming (trained in the SSL manner) are not aligned with user intent and serve as a Chatbot.

*OpenFlamingo is the open-source version of Flamingo, enabling community research with a strong interleaved data pretrained model

Flamingo ≠ Multi-modal Assistants

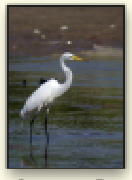
Question: What is the danger of this sport?

Human Expected: The sport involves players running and trying to catch the ball while others are standing in the grass, which can lead to collisions or accidents.

Flaming (trained in the SSL manner) are not aligned with user intent and serve as a Chatbot.

MMC4: Image-text interleaved data for OpenFlamingo Pretraining

[..., "Check out Shane Driscoll's take on sustainable communities and how his photograph fits this year's Green Cities theme.", ...,



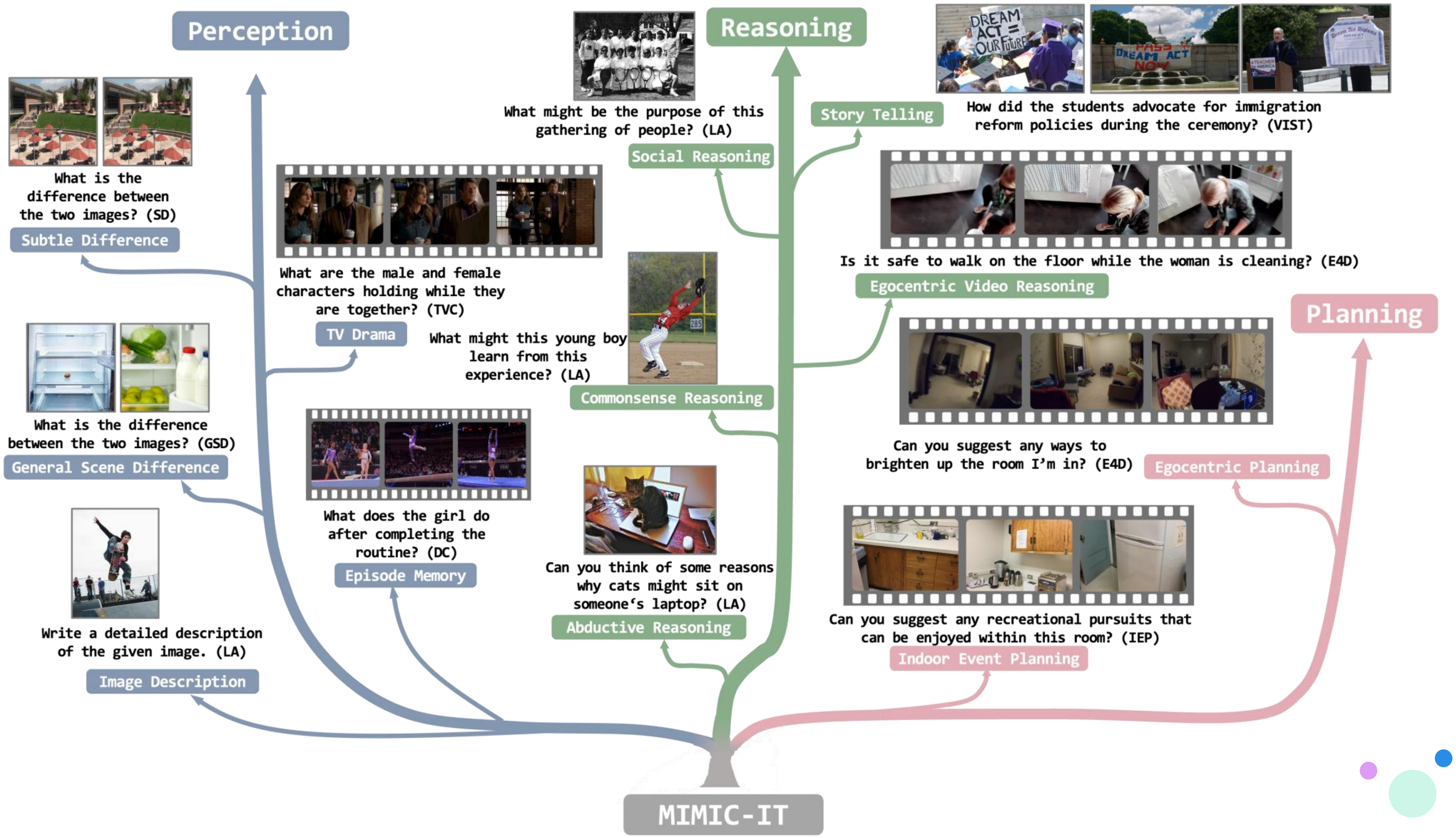
,"Man-made platforms like the one pictured here allow these fish-eating birds of prey to thrive in developed coastal areas.", "A city surrounded by mountains.", "I took this photo in October on a hike in New Hampshire.",

, "It is looking at Mt. Chicora from the middle sister mountain.", "Getting people out into beautiful places like this is becoming more and more popular, and each time we bring a little piece of nature back with us that inspires us to make our cities better.", ...]

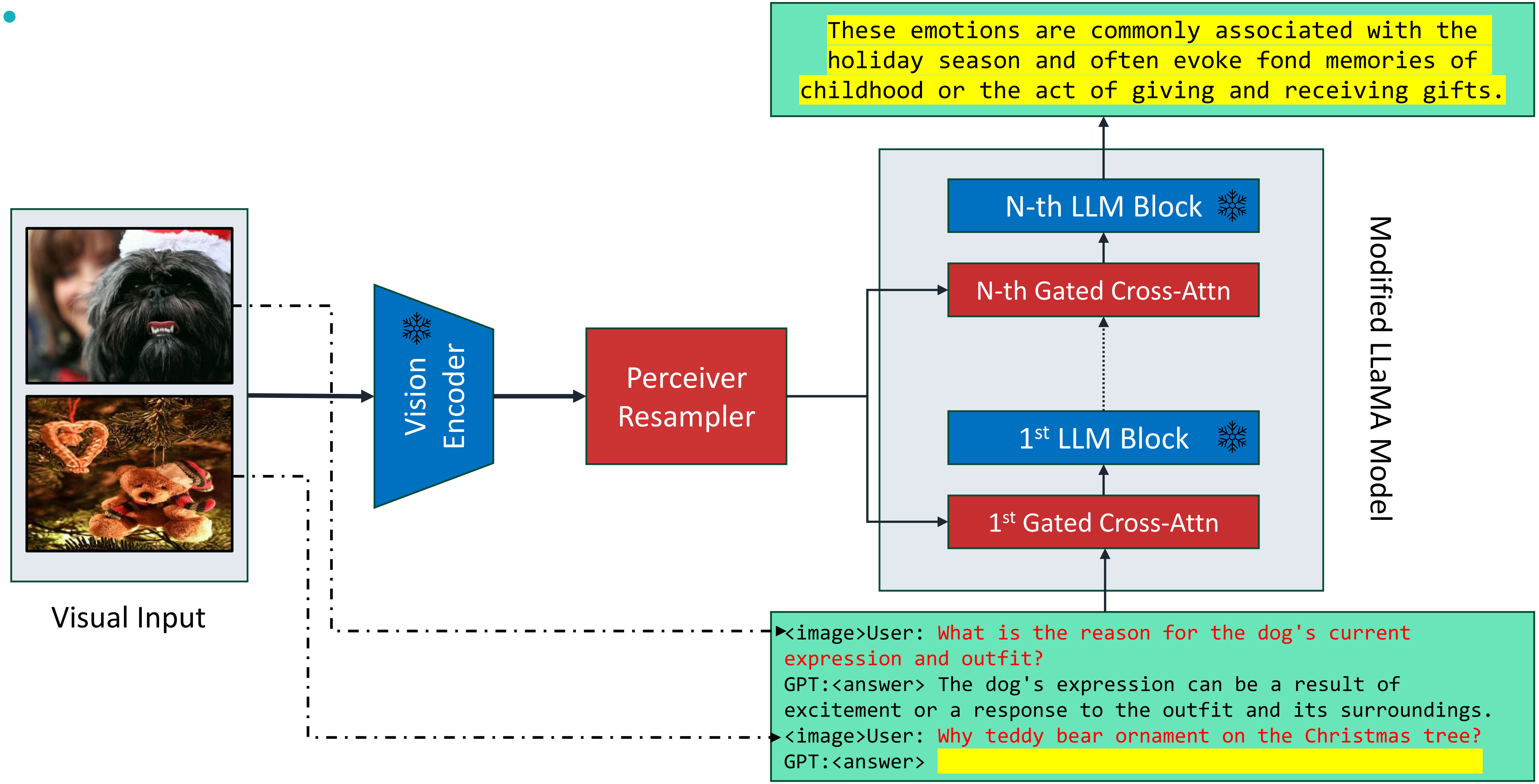
Diverse and large-scale, but lack of Instruct-following scenario

Zhu et. al. Multimodal C4: An open, billion-scale corpus of images interleaved with text. 2023

MIMIC-IT Dataset

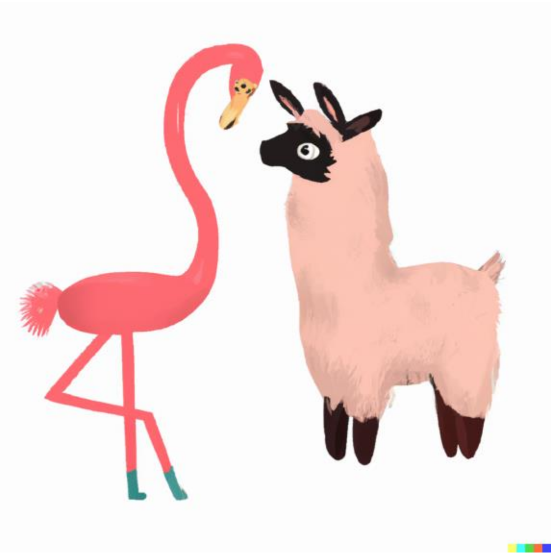


Otter: A Multi-Modal In-context Instruction Tuned Model



From interleaved data pretraining to multi-modal In-context instruction tuning

MMC4
(interleaved pretraining)



OpenFlamingo

MIMIC-IT
(Multi-Modal In-Context Instruction Tuning)

Otter

Otter

Multi-Modal In-Context Learning Model with Instruction Tuning

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Otter

Cognition

Sum of the scores of all cognition subtasks, including commonsense reasoning, numerical calculation, text translation, and code reasoning. The full score of each subtask is 200, and that of all cognition is 800.

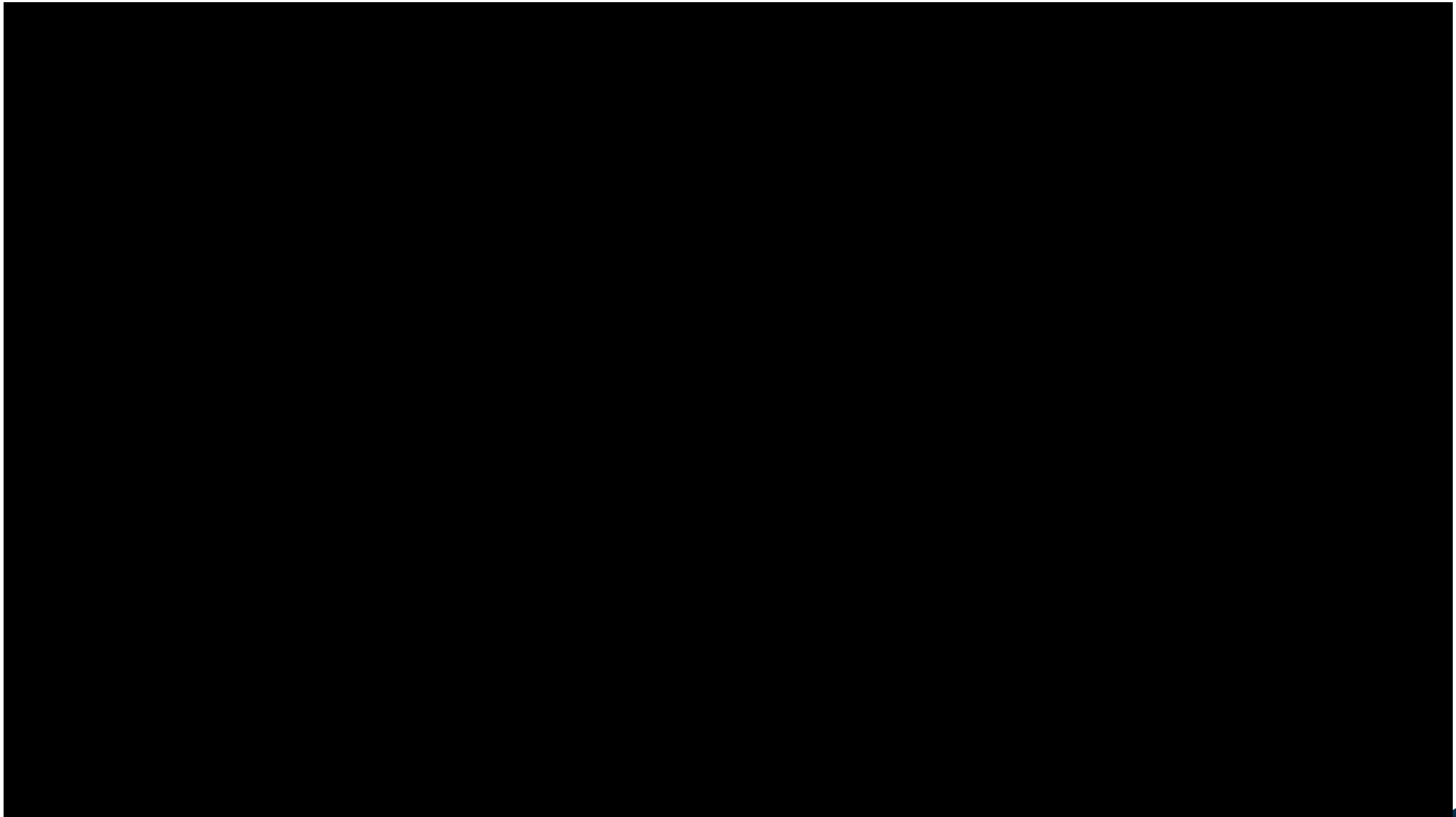
Rank	Model	Version	Score
🥇	**Otter**	**OTTER-Image-MPT7B**	**306.43**
🥈	**MiniGPT-4**	**minigpt4-aligned-with-vicuna13b**	**292.14**
🥉	**InstructBLIP**	**blip2-instruct-flant5xxl**	**291.79**
4	BLIP-2	blip2-pretrain-flant5xxl	290.00
5	mPLUG-Owl	mplug-owl-llama-7b	276.07
6	LaVIN	LAVIN-13B	249.64
7	LLaMA-Adapter V2	LLaMAv2-7B	248.93
8	PandaGPT	pandagpt-7b-max-len-512	228.57
9	Multimodal-GPT	Multimodal-GPT-9B	226.79
10	LLaVA	LLaVA-7B-v0	214.64
11	ImageBind_LLM	imagebind_LLM-7B	213.57
12	VisualGLM-6B	VisualGLM-6B	181.79

Otter

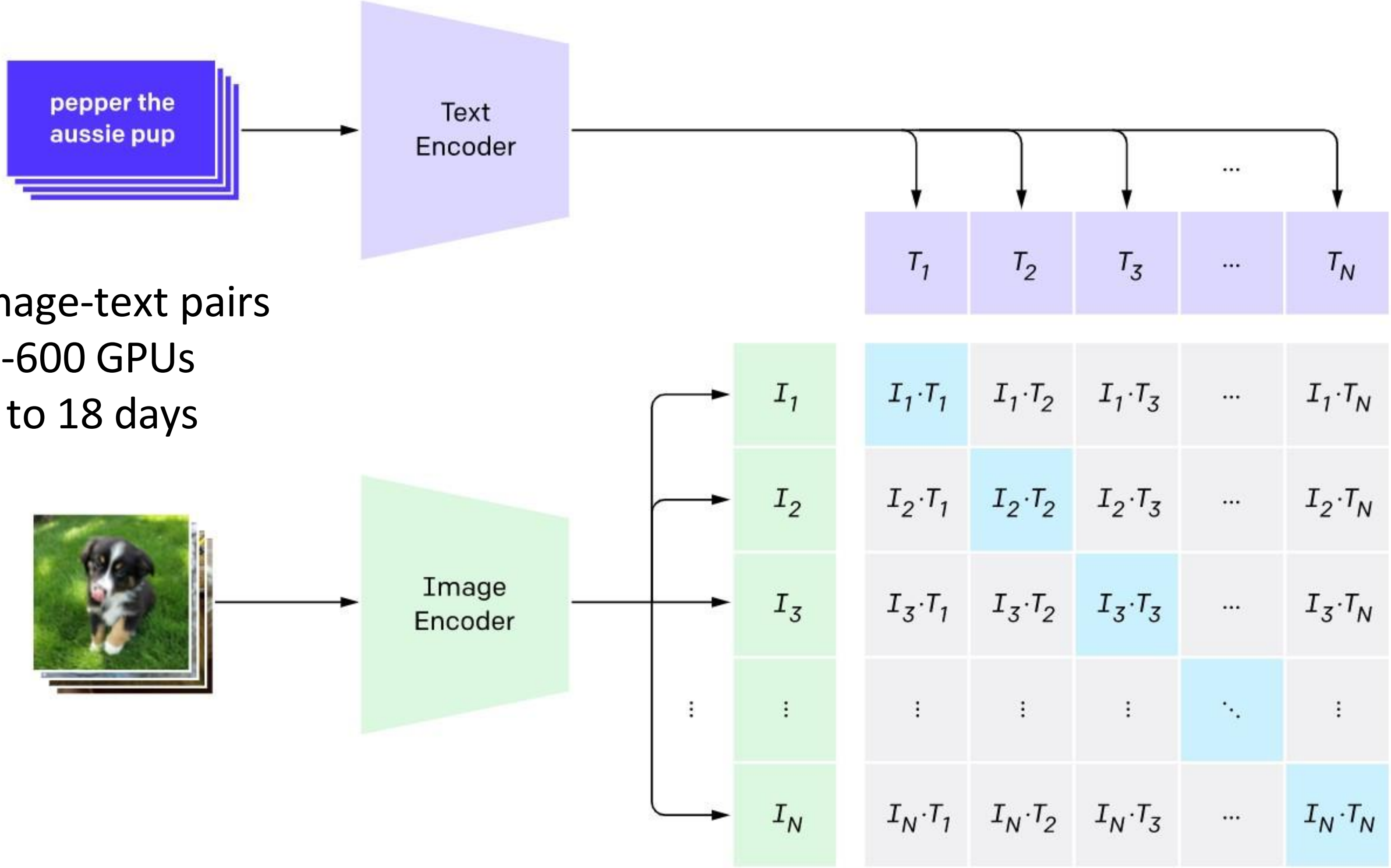
Multi-Modal In-Context Learning Model with Instruction Tuning

Otter

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



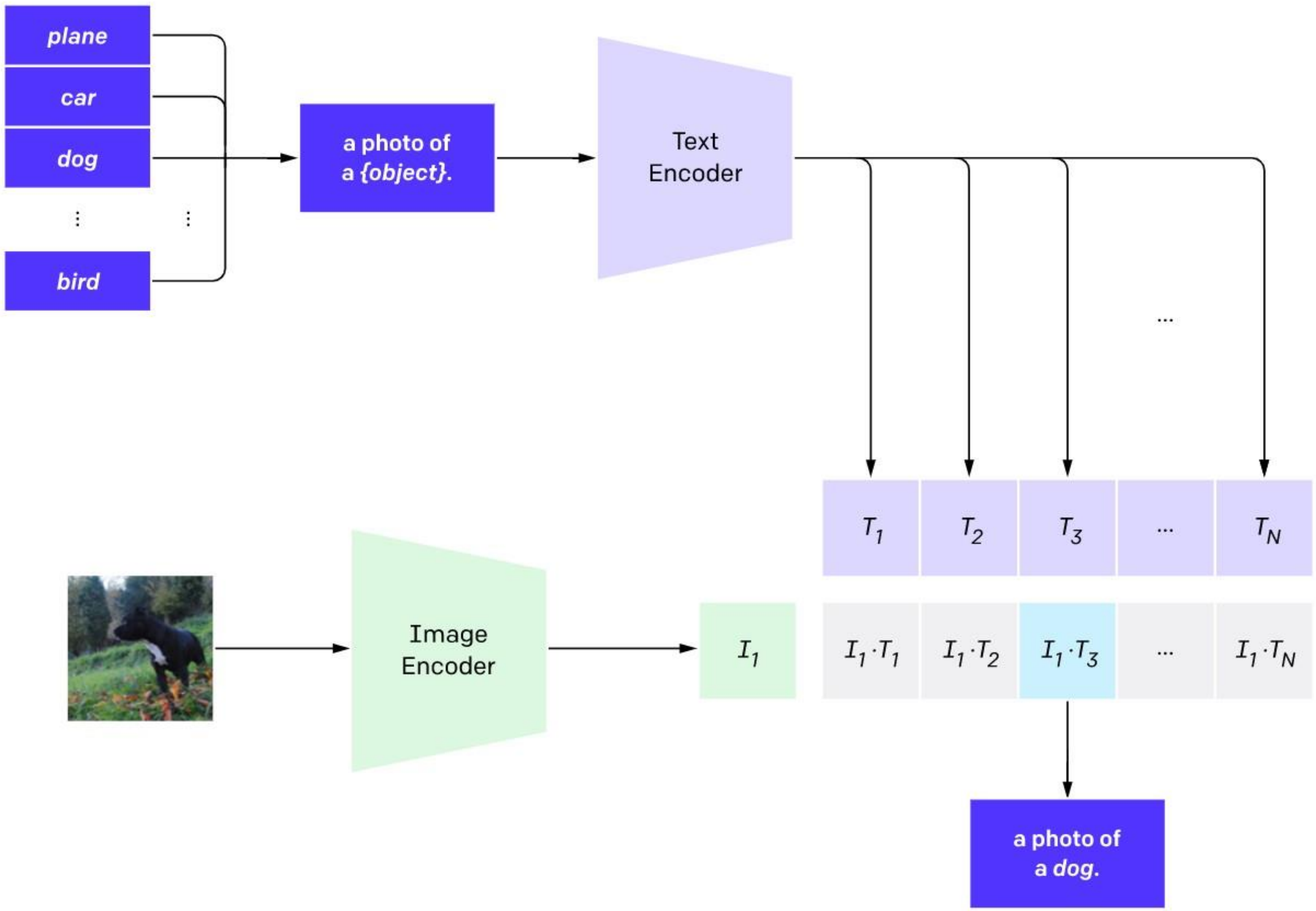
Contrastive Language-Image Pre-training (CLIP)



- Data: 400M image-text pairs
- Compute: 250-600 GPUs
- Train time: up to 18 days

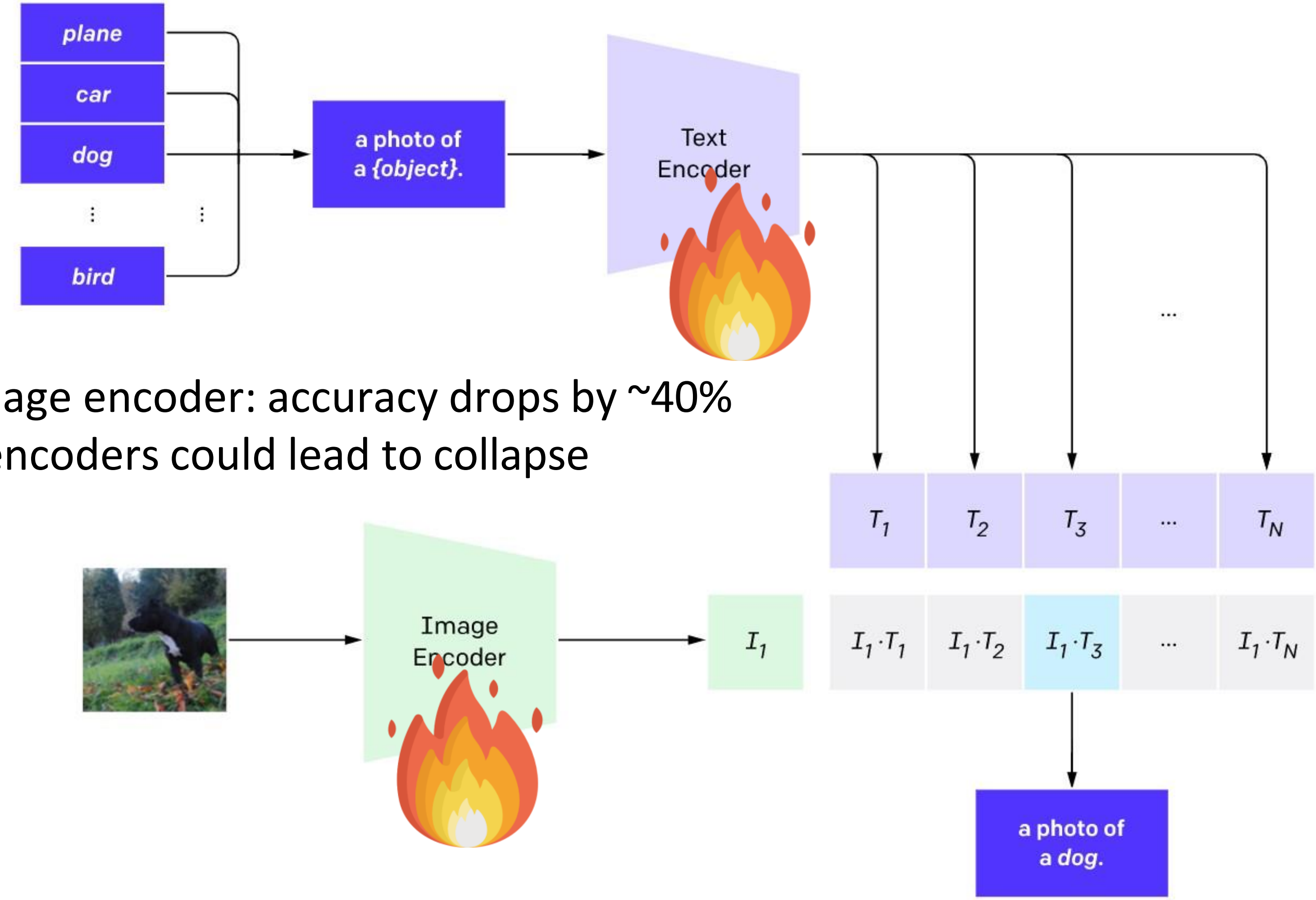
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML’21.

Zero-shot image recognition via prompting



Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.

Fine-tuning might not be a good idea



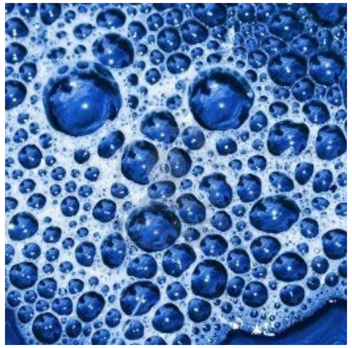
- Fine-tuning the image encoder: accuracy drops by ~40%
- Fine-tuning both encoders could lead to collapse

Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

Prompt engineering is too time-consuming

Caltech101 Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**91.83**

Flowers102 Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**94.51**

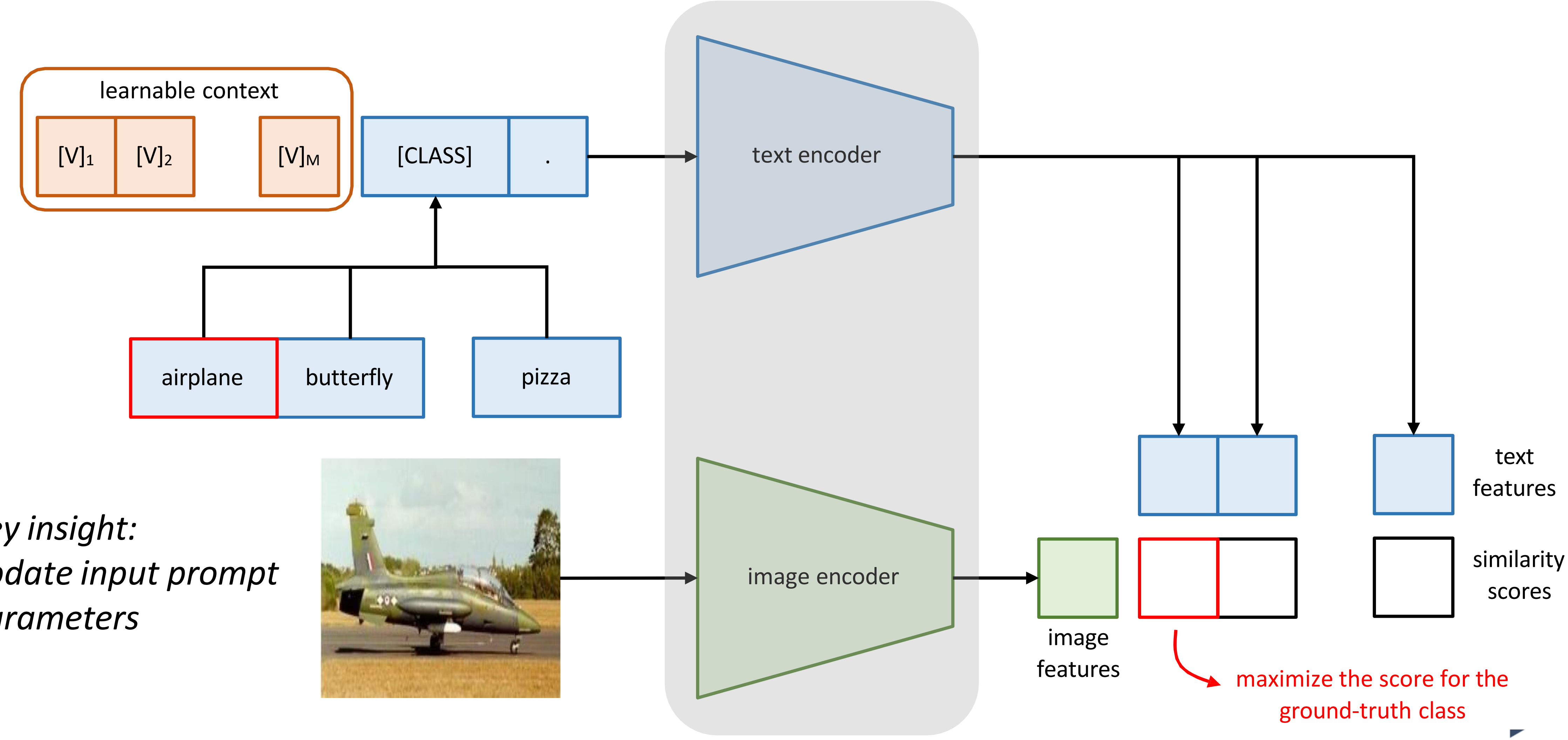


Describable Textures (DTD) Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**63.58**

EuroSAT Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**83.53**

Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

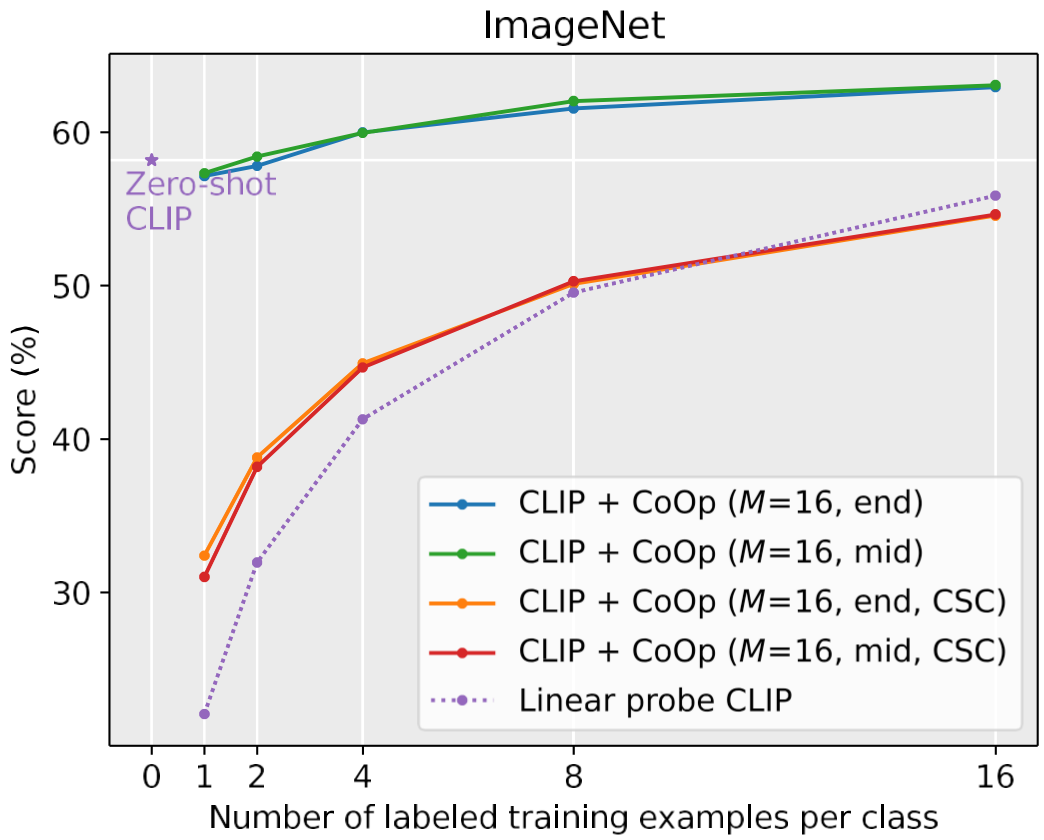
Context Optimization (CoOp)

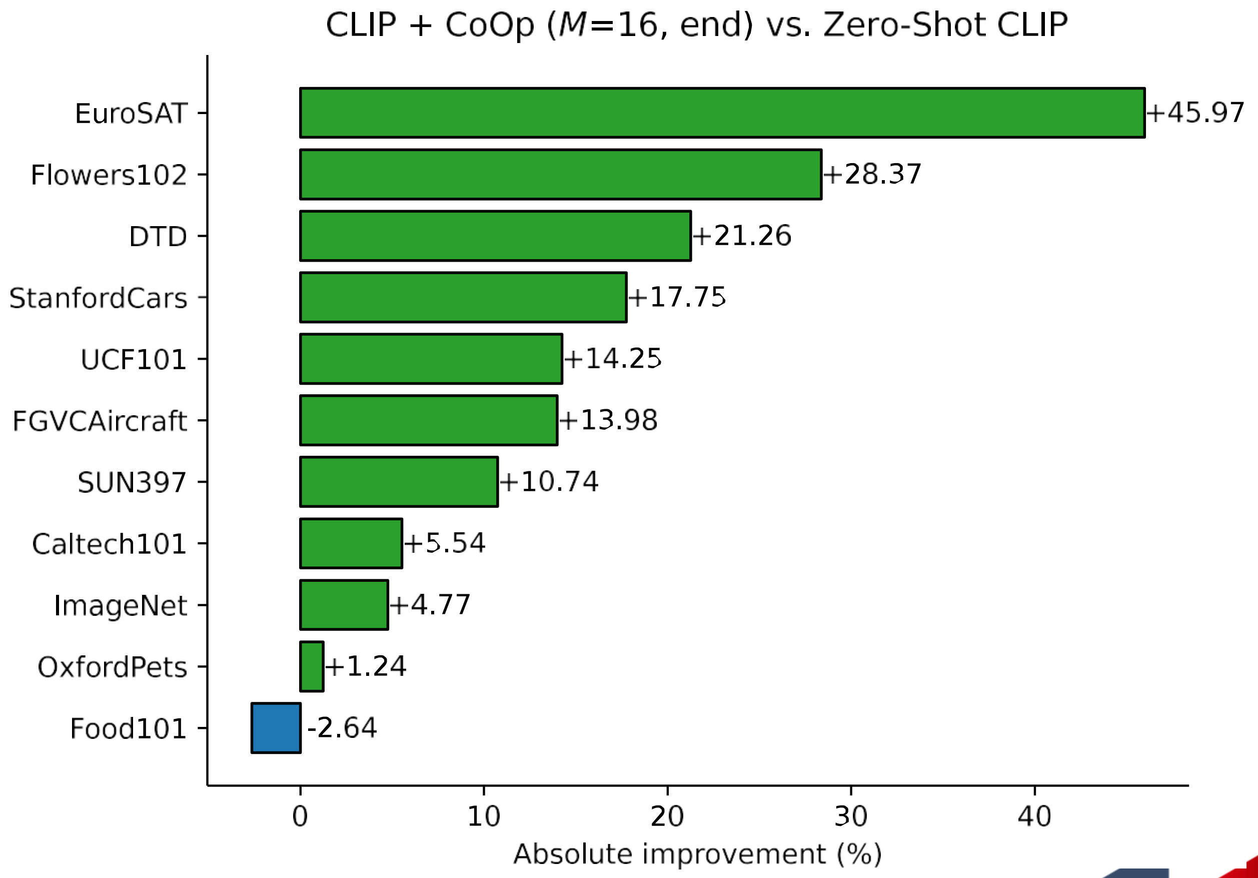


Key insight: update input prompt parameters

Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

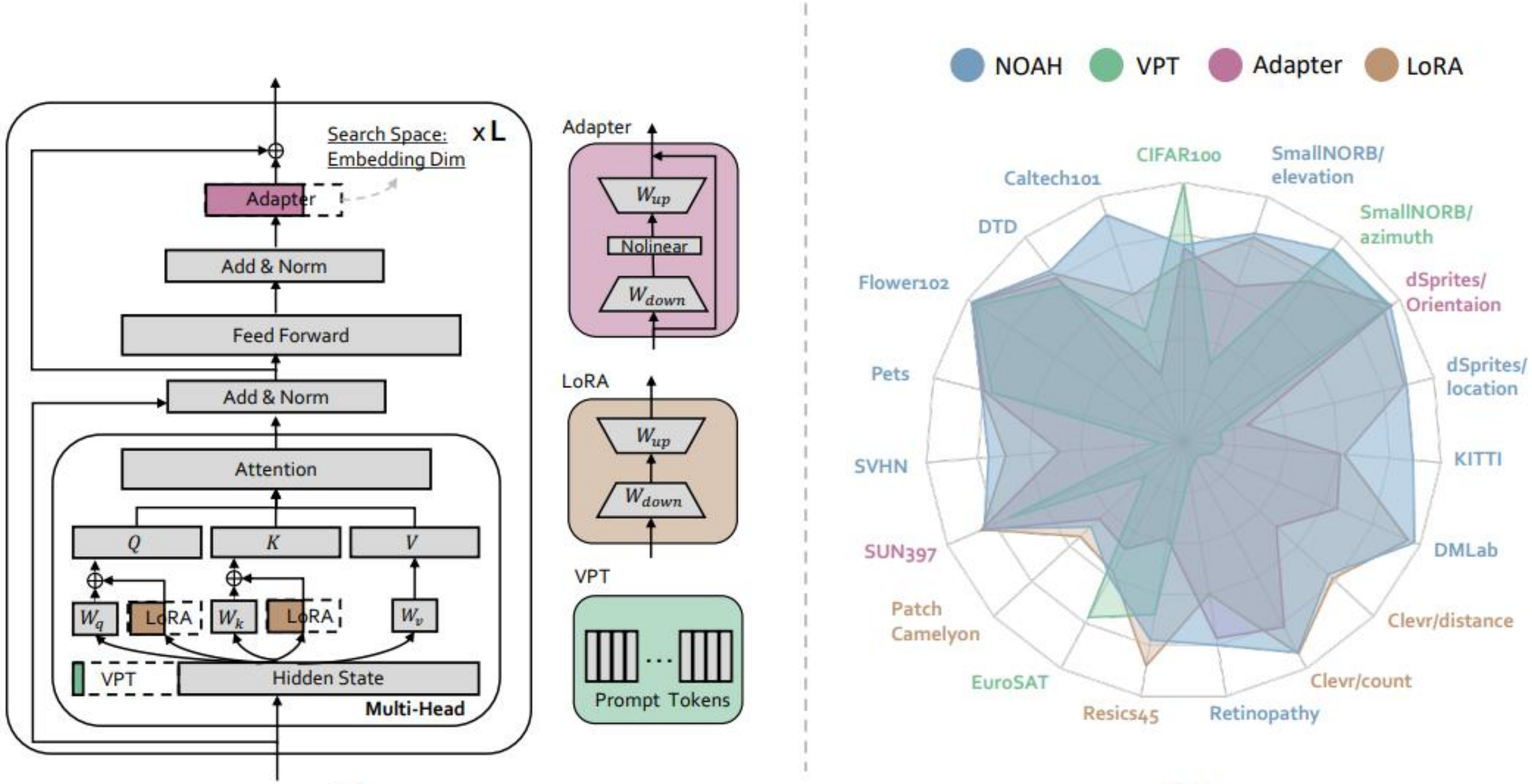
CoOp is a few-shot learner





Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

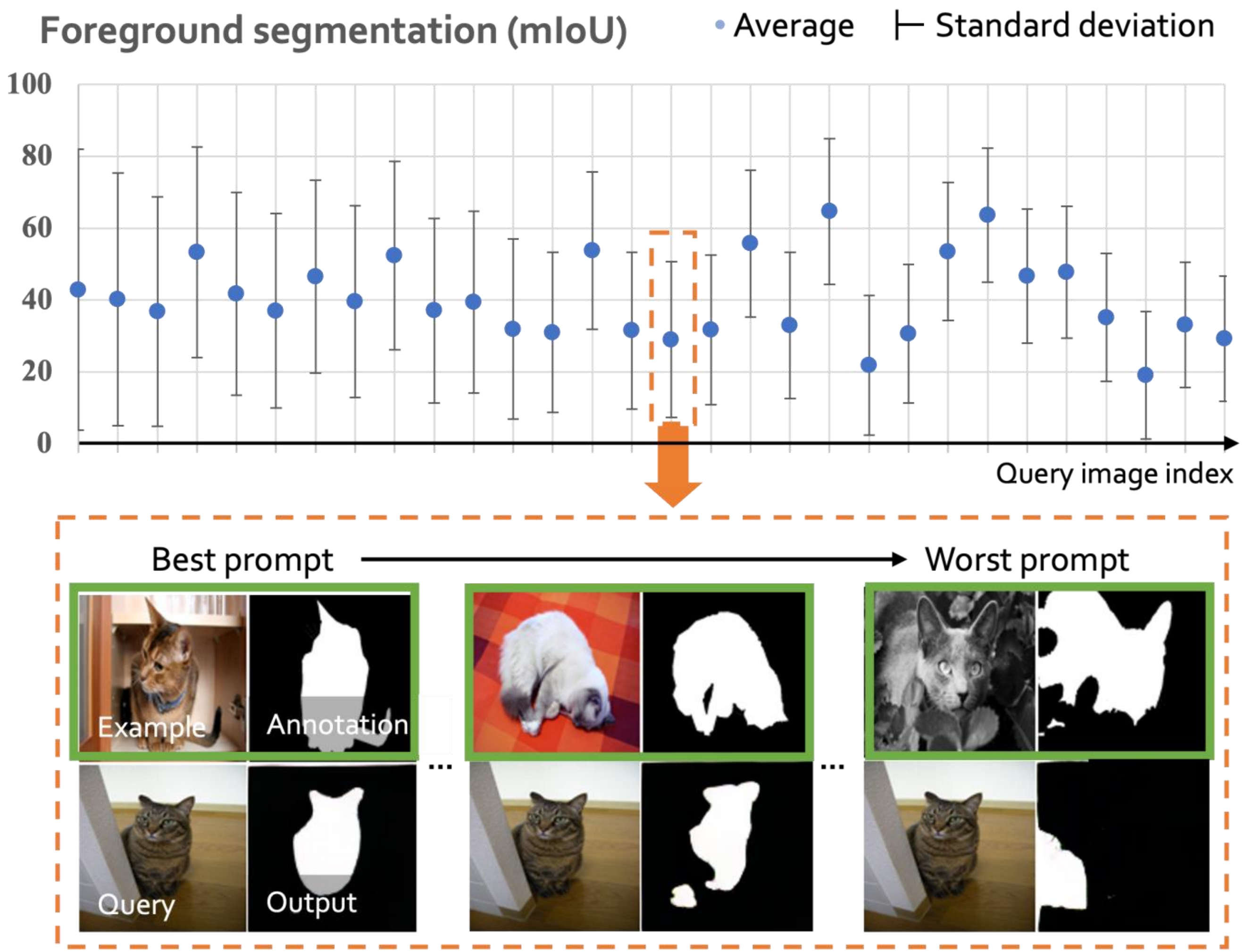
NOAH



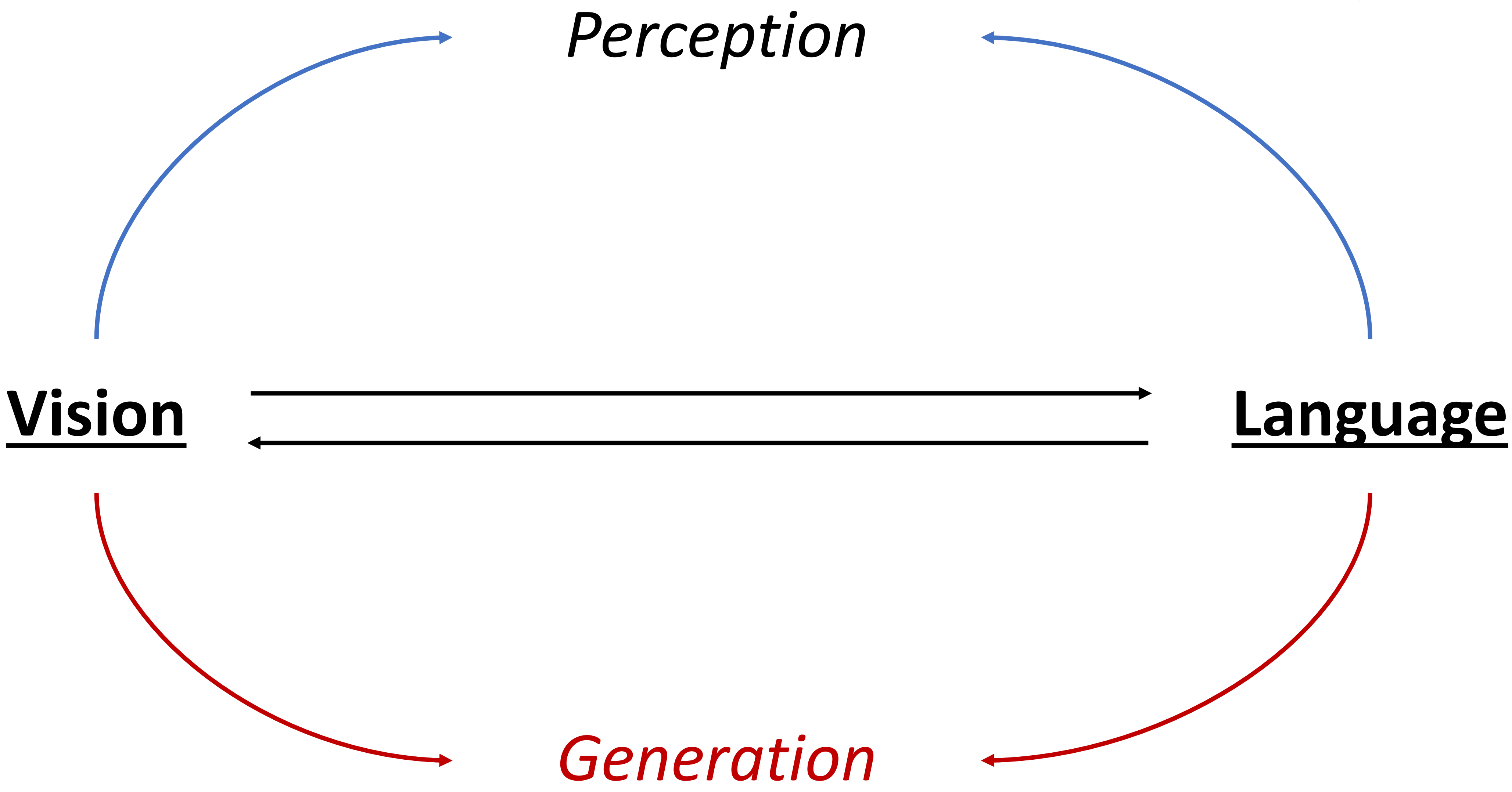
Search Space: Embedding Dim
x L
Adapter
Add & Norm
Feed Forward
Add & Norm
Attention
Q
K
V
W_q
LoRA
W_k
LoRA
W_v
VPT
Hidden State
Multi-Head
Adapter
W_{up}
Nolinear
W_{down}
LoRA
W_{up}
W_{down}
VPT
Prompt Tokens
NOAH
VPT
Adapter
LoRA
CIFAR100
SmallNORB/ elevation
SmallNORB/ azimuth
dSprites/ Orientaion
dSprites/ location
KITTI
DMLab
Clevr/distance
Clevr/count
Retinopathy
Resics45
EuroSAT
Patch Camelyon
SUN397
SVHN
Pets
Flower102
DTD
Caltech101

(a) (b)

Visual In-Context Learning



(a) Visual in-context learning is sensitive to prompt selection


Perception
Vision
Language
Generation

Content Generation Powered by Foundation Models

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Visual AIGC

Avatar

Foundation Model

Object

Scene

Visual AIGC

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Avatar

Foundation Model

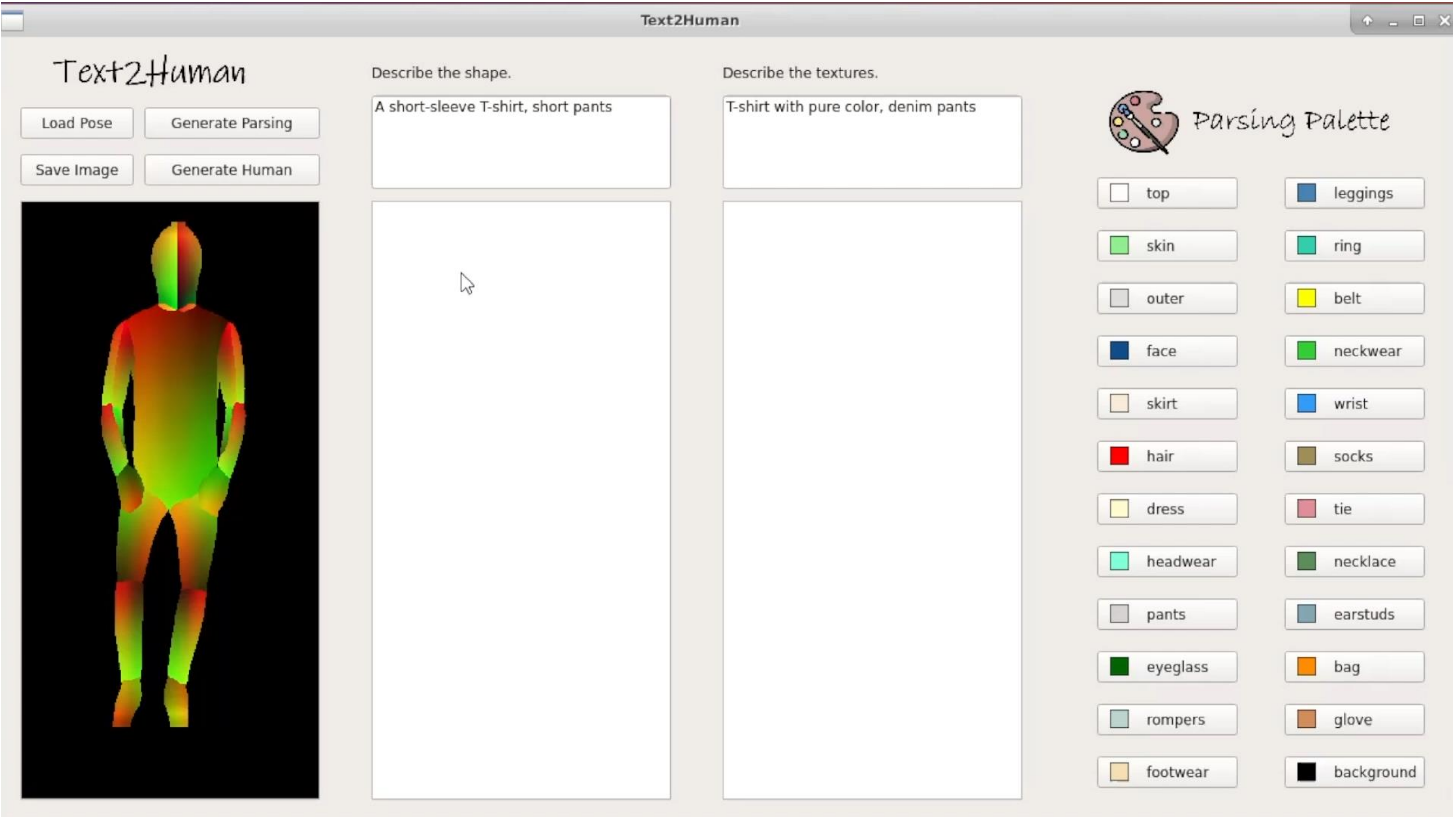
Object

Scene

StyleGAN-Human: 2D Human Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Text2Human: Text-to-2D Human



Text2Performer: Text-to-2D Human Video

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

The dress the person wears has medium sleeves and it is of short length. The texture of it is pure color.

The lady moves to the left.

She is turning right from the front to the side.

She is turning right from the side to the back.

She turns right from the back to the side.

She turns right from the side to the front.

She moves to the right.

EVA3D: 3D Human Generation

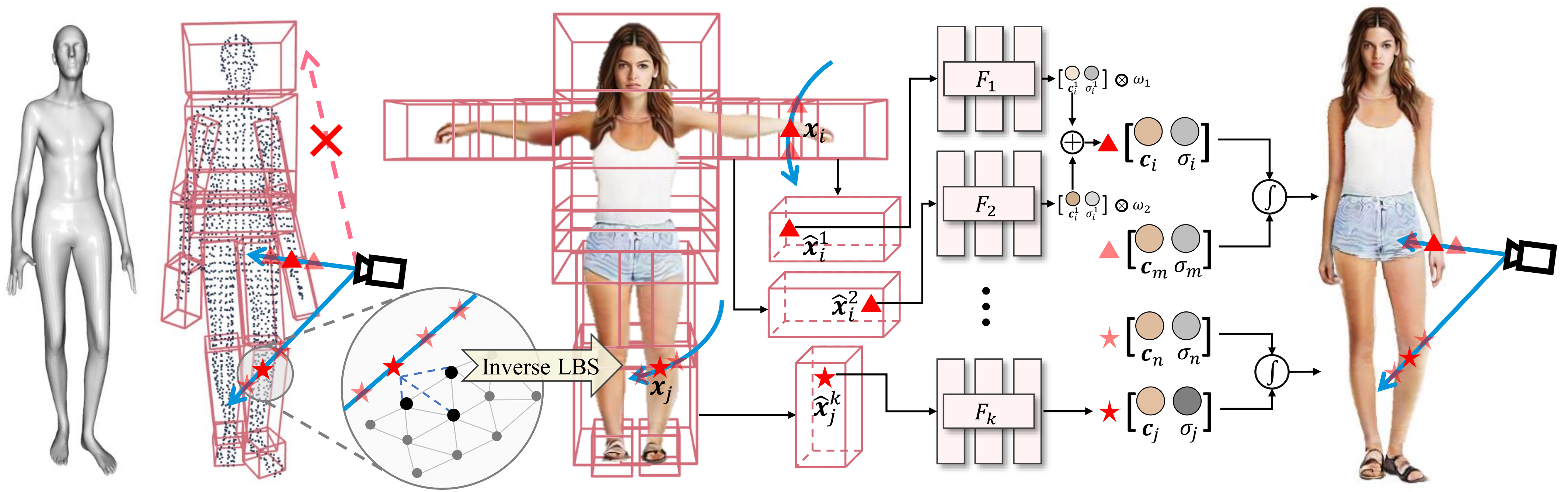
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

- Learn 3D generation from 2D image collections

Static ? ***Articulated***

EVA3D: 3D Human Generation

- Compositional Human NeRF



a) Sample Rays & Points at Observation Space SMPL(β, θ)

b) Inverse LBS to Canonical Space

c) Transform to Local Bounding Box & Query Subnetworks

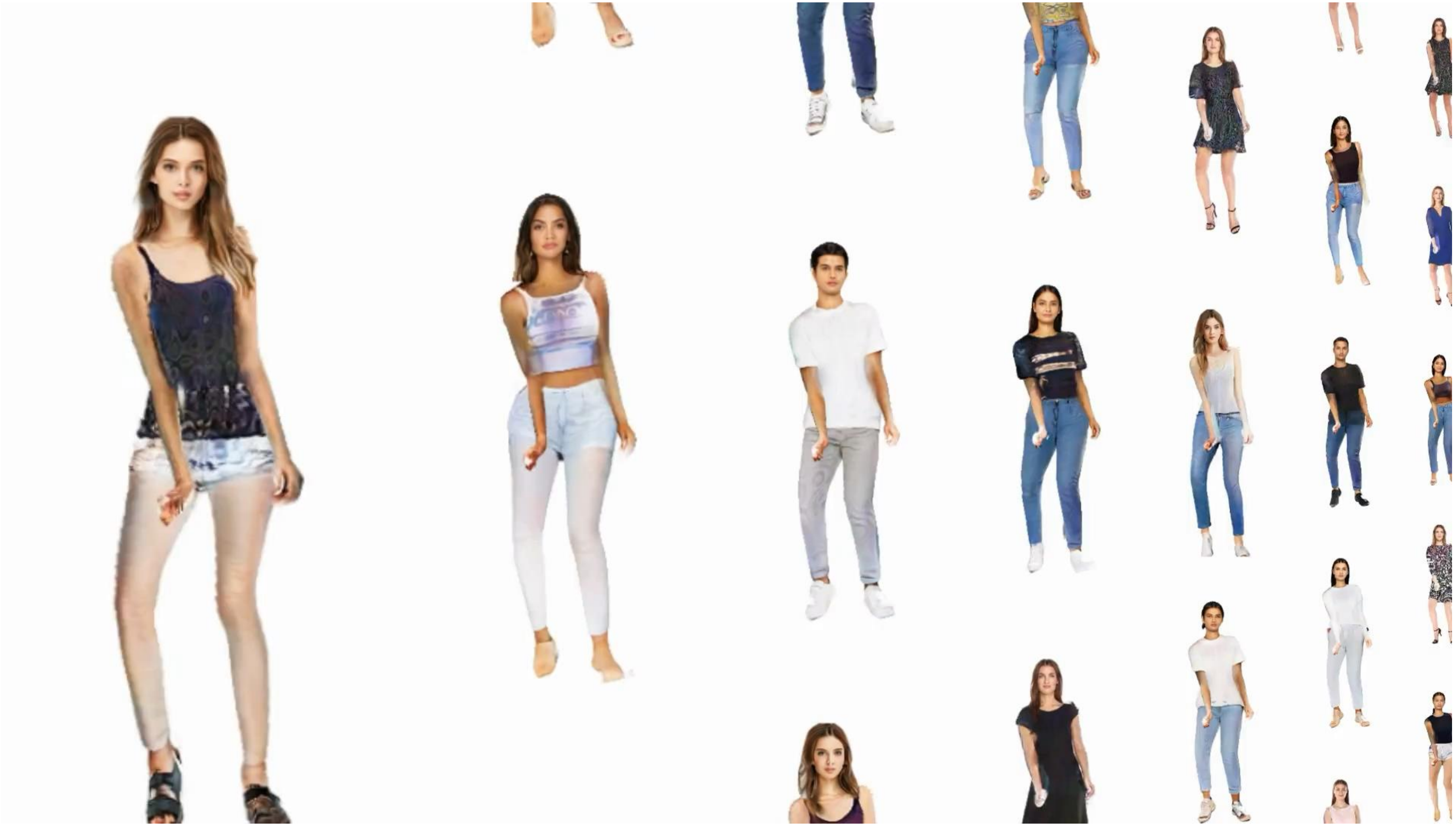
d) Integrate Queried Results Along Rays

EVA3D: 3D Human Generation

- Qualitative Results

EVA3D: 3D Human Generation

- Explicit Pose/ Shape Control



AvatarCLIP: Text-to-3D Avatar

I want to generate a tall and fat Iron Man that is running.

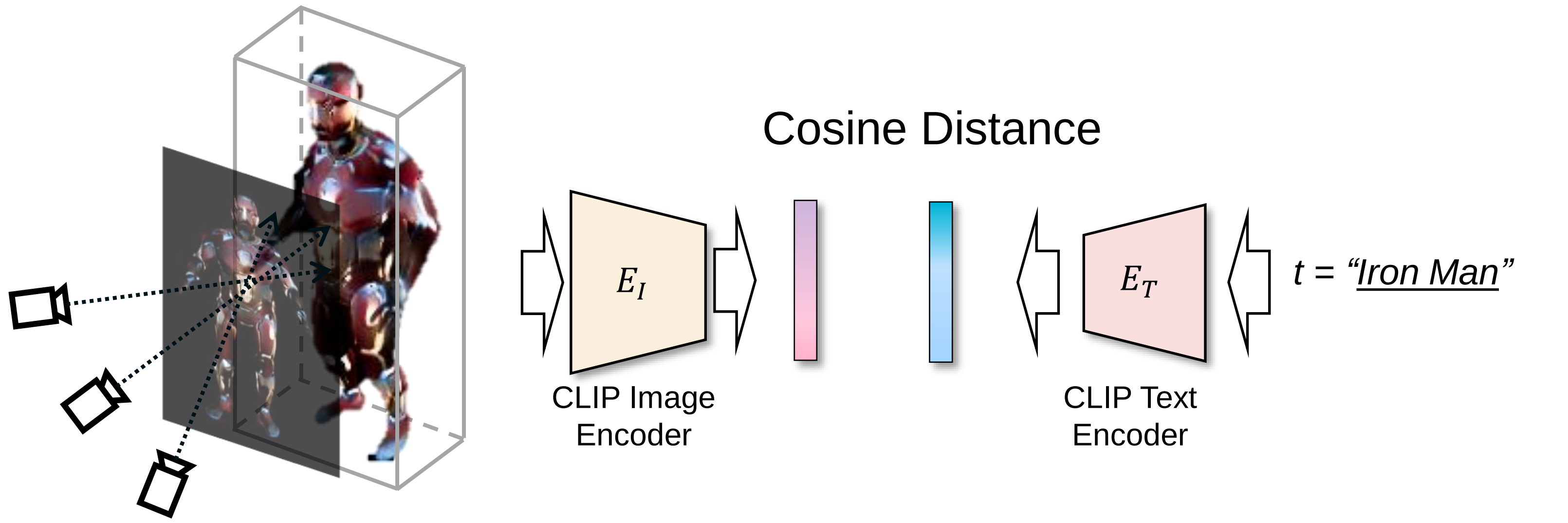
I would like to generate a skinny ninja that is raising arms.

I want to generate a tall and skinny female soldier that is arguing.

I want to generate an overweight sumo wrestler that is sitting.

TEXT-DRIVEN 3D GENERATION

CLIP + DIFFERENTIABLE RENDERING

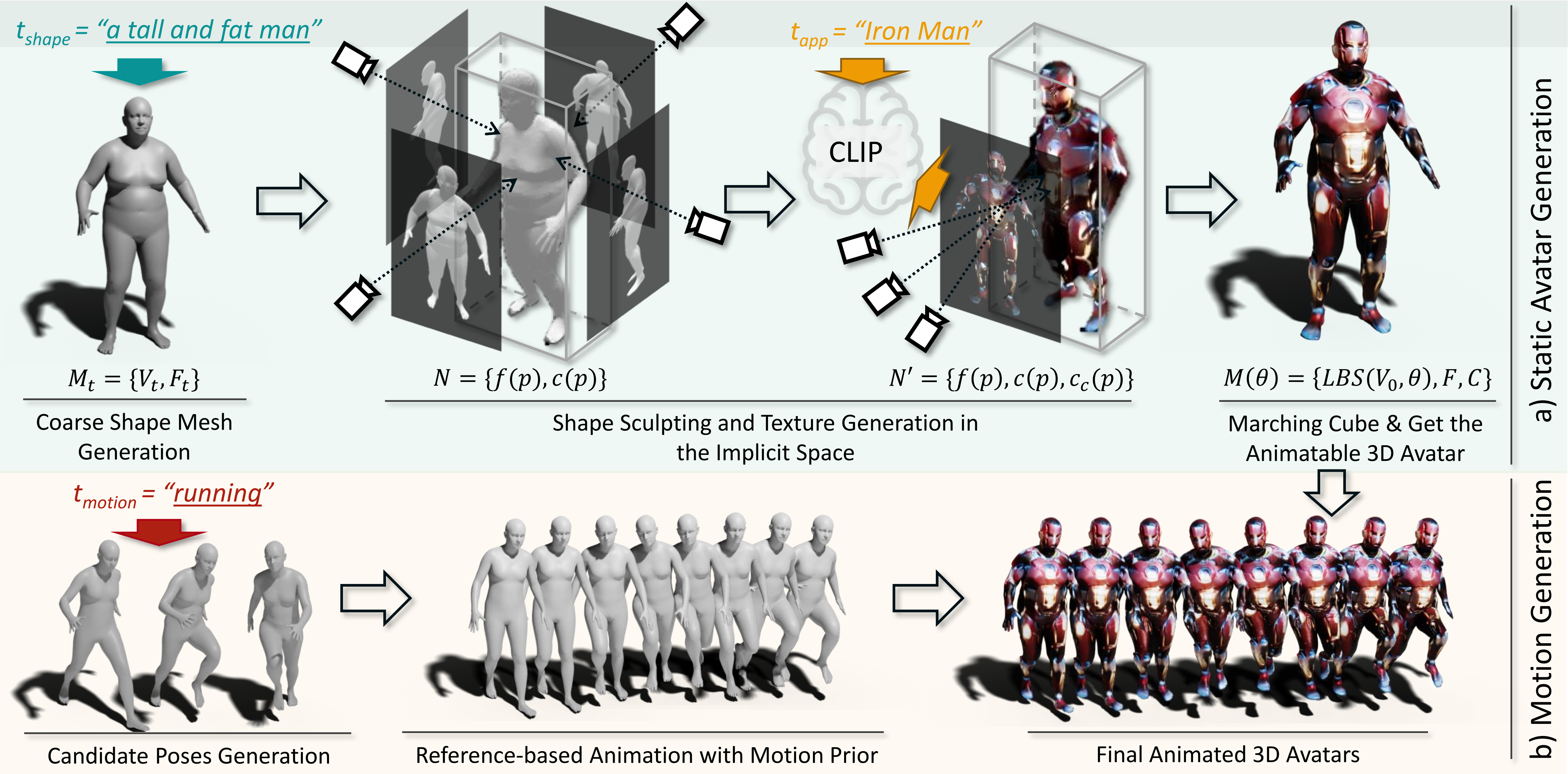


a) Differentiable Rendering

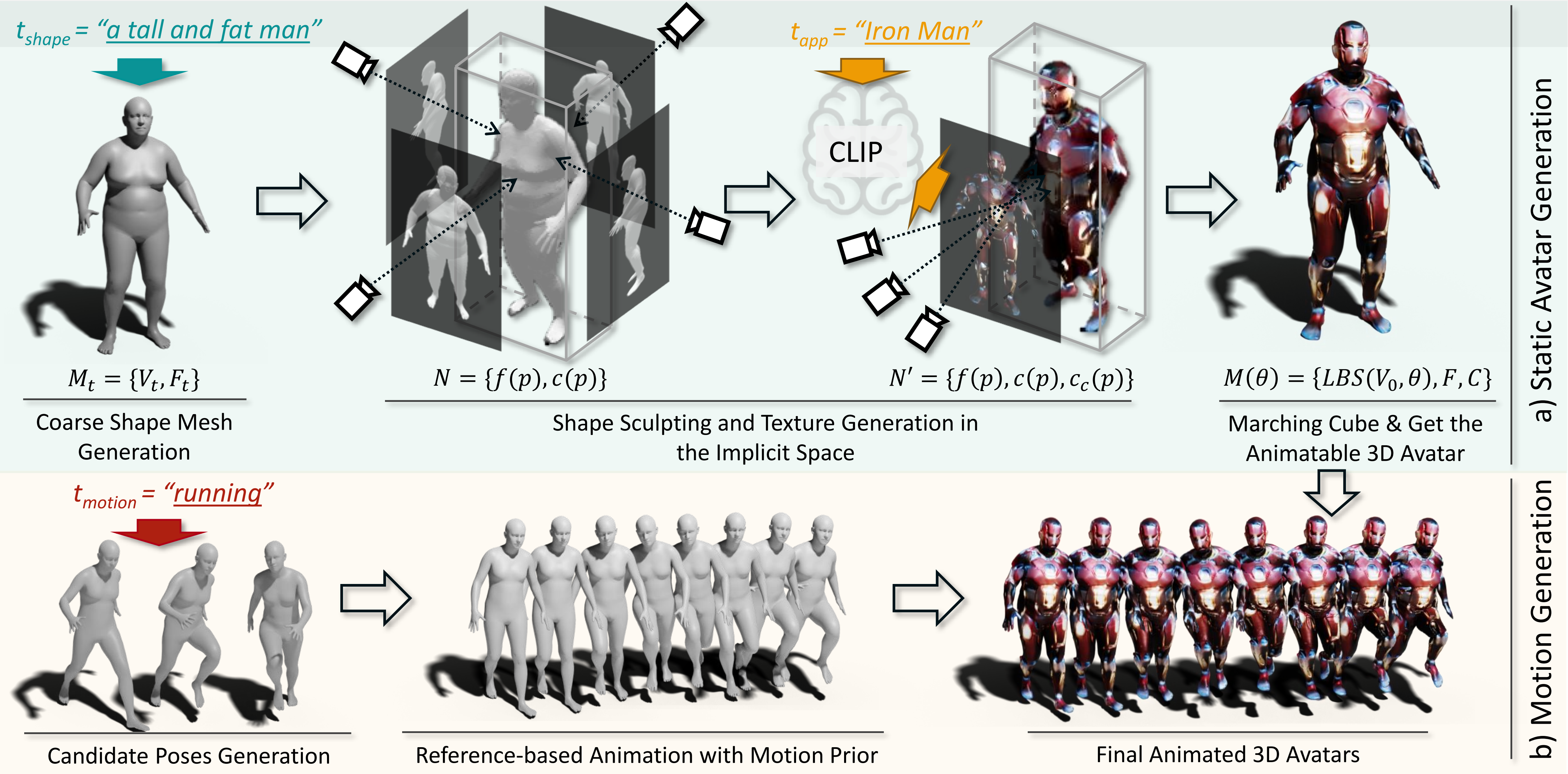
b) Optimization guided by CLIP

AVATARCLIP: DETAILED PIPELINE

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG



AVATARCLIP: DETAILED PIPELINE



CONTROLLING & CONCEPT MIXING ABILITIES

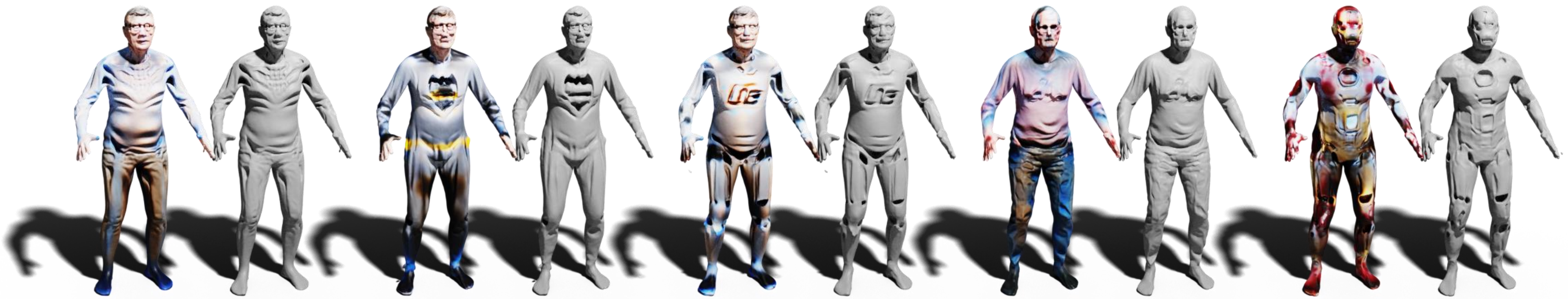
1. Superman
2. the face of Bill Gates

1. Iron Man
2. the face of Steve Jobs

Steve Jobs in White Shirt

Man in Jeans

Man in White Shirt



Alien Bill Gates

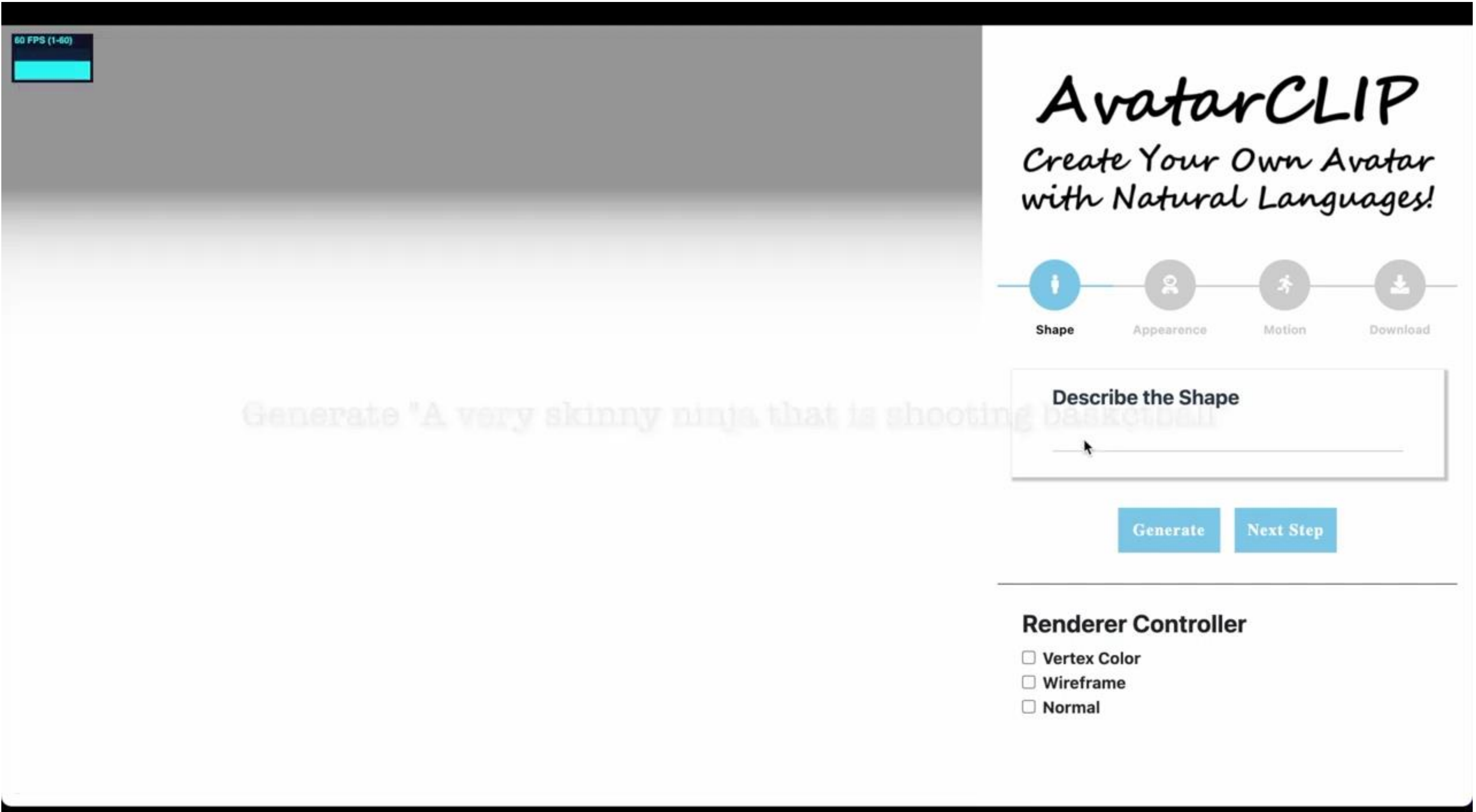
Bill Gates Wearing Batman Suit

Robot Bill Gates

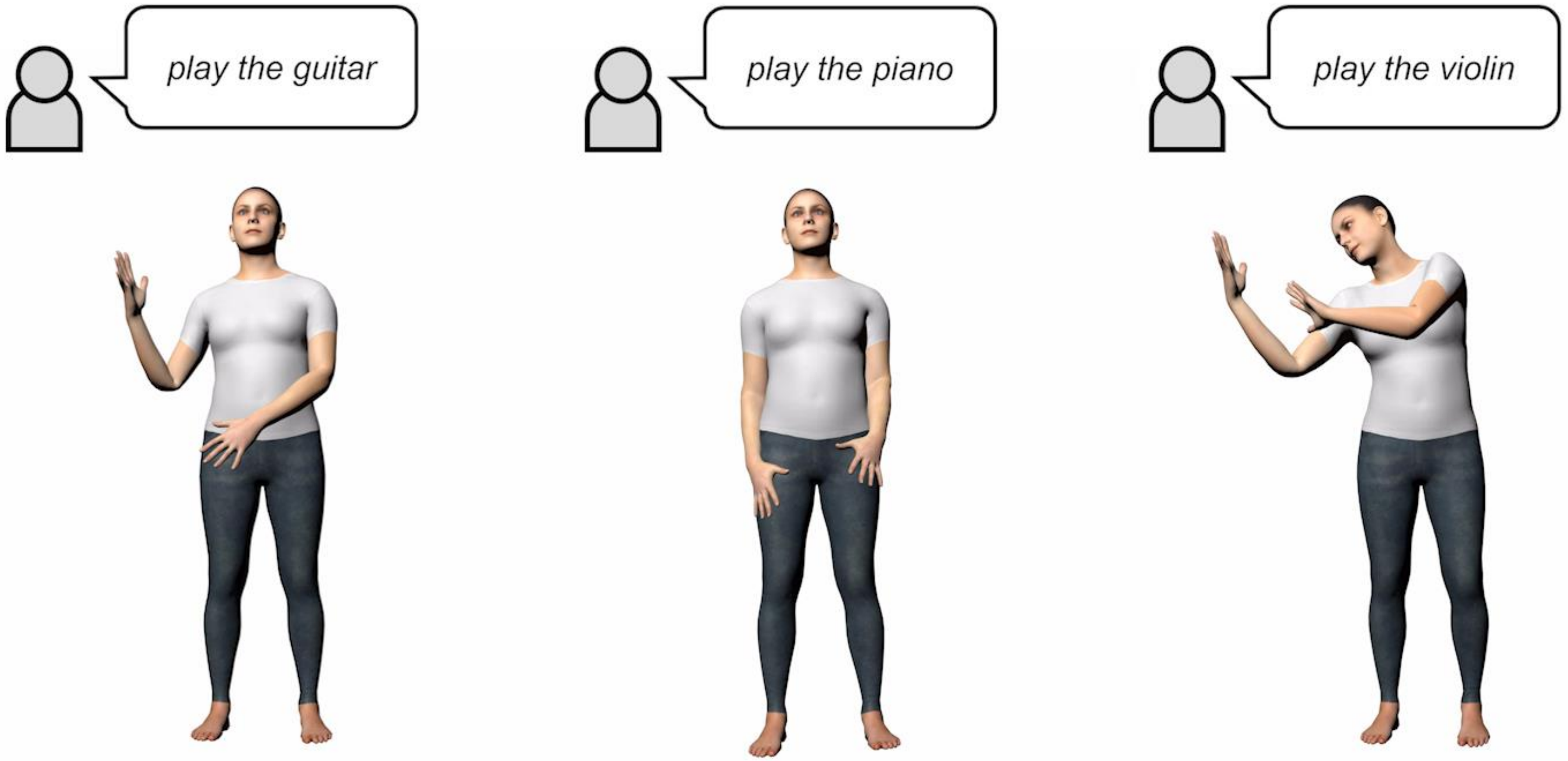
Zombie Steve Jobs

Zombie Iron Man

AvatarCLIP: Text-to-3D Avatar



MotionDiffuse: Text-to-3D Human Video



3D Animation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Video Games

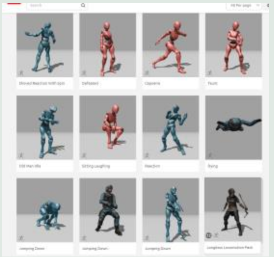
Films

VTuber

Motion Collection

Manual Editing

Motion Capture



Gallery

1. Expensive
2. Time-consuming
3. Not User-friendly

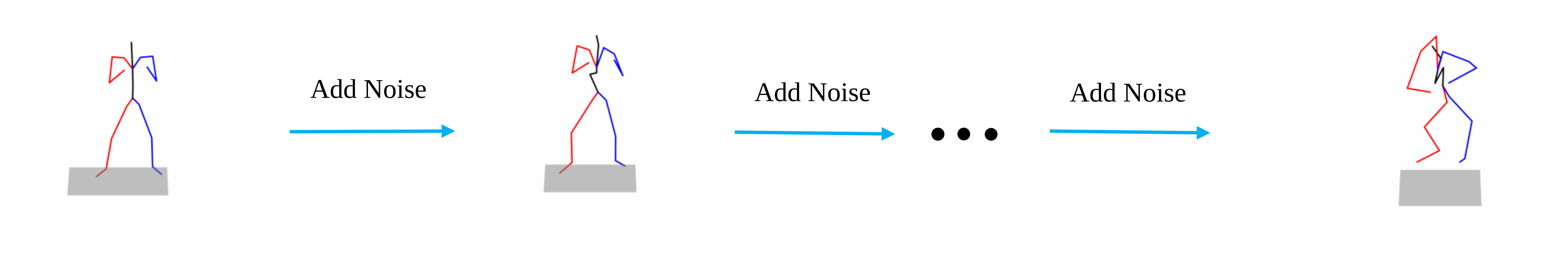
Human Mesh Recovery

Conditional Motion Generation

1. Cheap
2. Efficient
3. User-friendly

Motion Generation with Diffusion Model

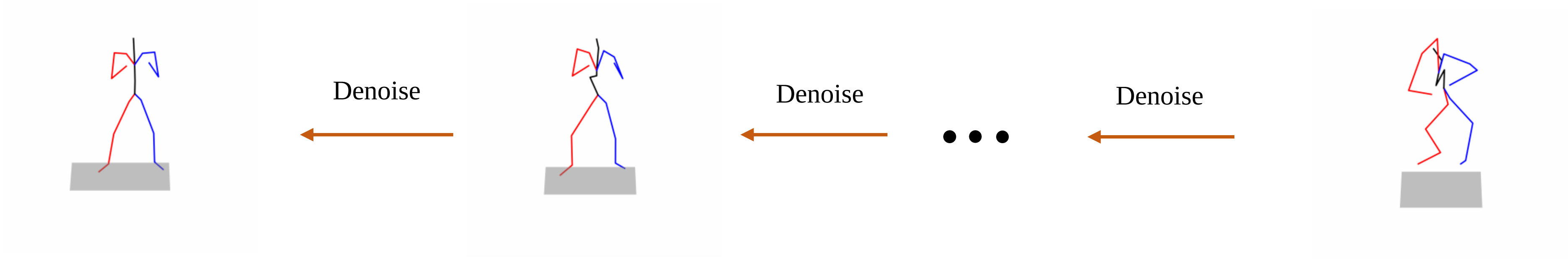
Diffusion Process



$\mathbf{x_0} \sim q(\mathbf{x_0})$

$p(\mathbf{x_T}) = \mathcal{N}(\mathbf{x_T}; \mathbf{0}, \mathbf{I})$

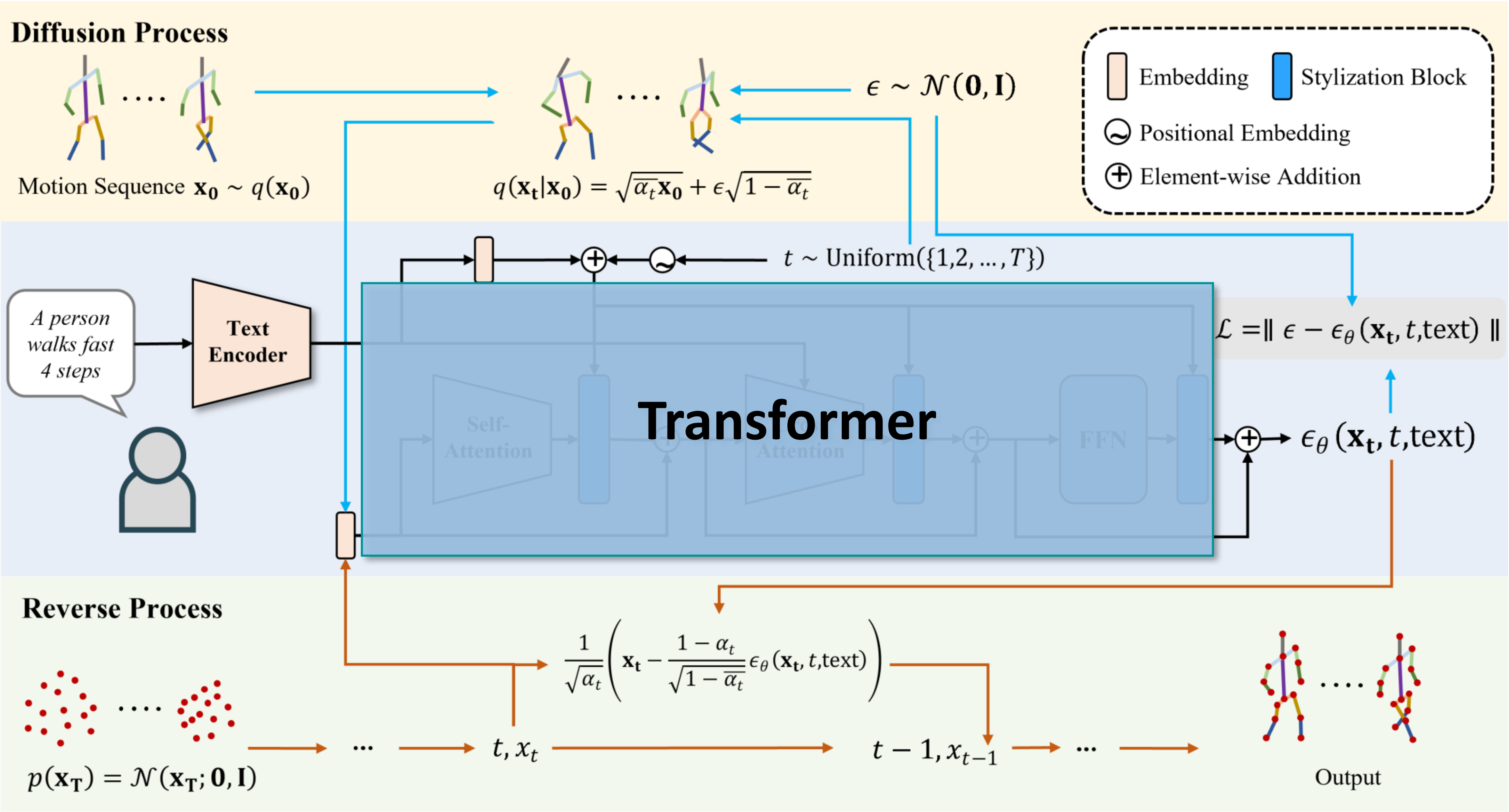
Reverse Process



$\mathbf{x_0} \sim q(\mathbf{x_0})$

$p(\mathbf{x_T}) = \mathcal{N}(\mathbf{x_T}; \mathbf{0}, \mathbf{I})$

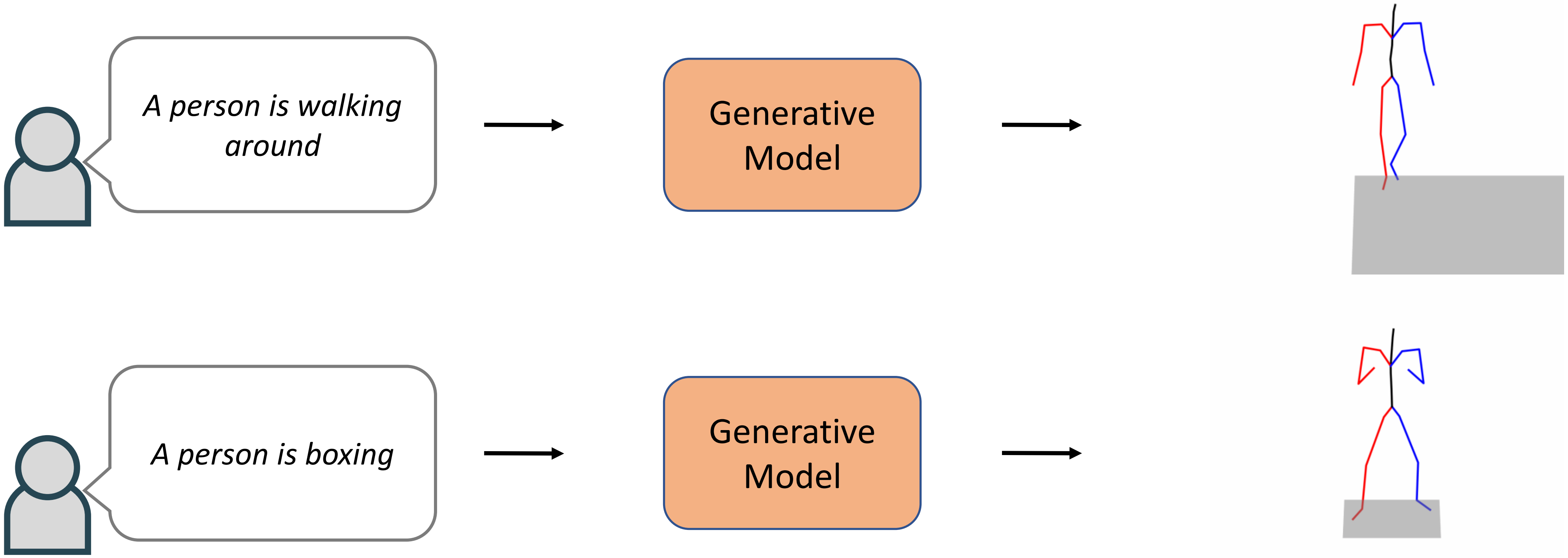
Framework



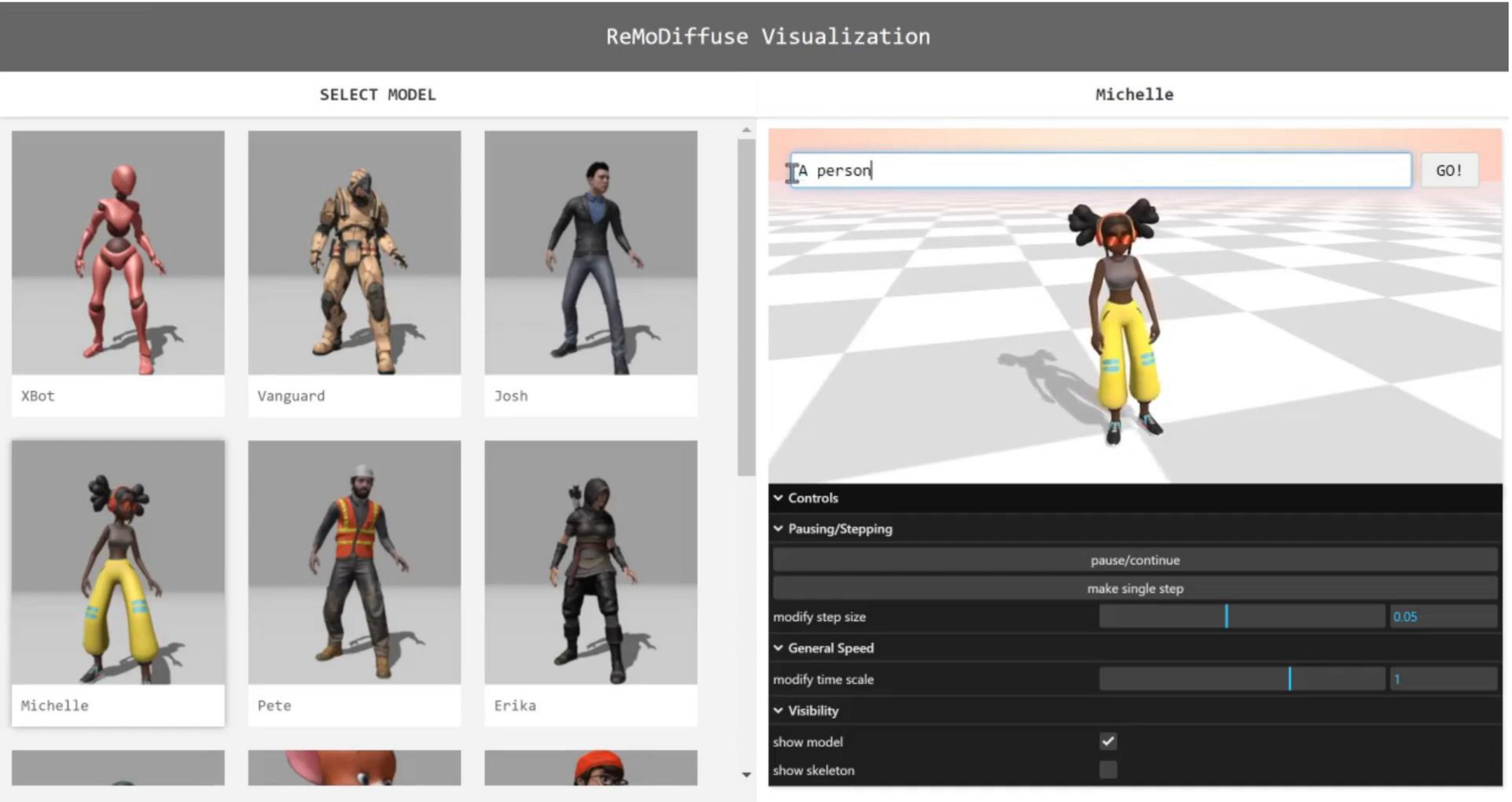
Challenge：

1. Variable length
2. Fusing timestep
3. Improve efficiency

Text-driven Motion Generation

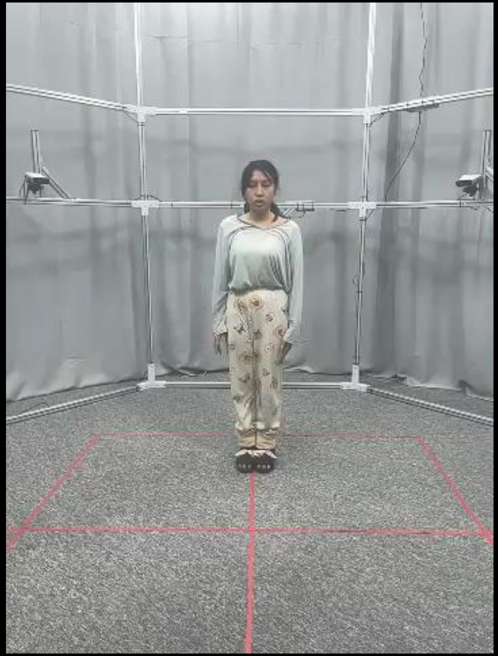


ReMoDiffuse: Text-to-3D Human Video



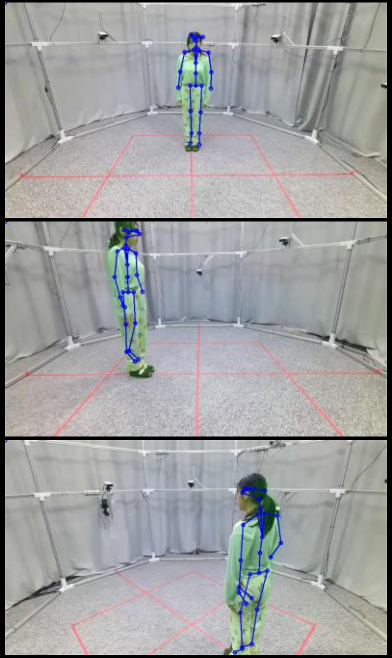
HuMMan Dataset

Artec Eva

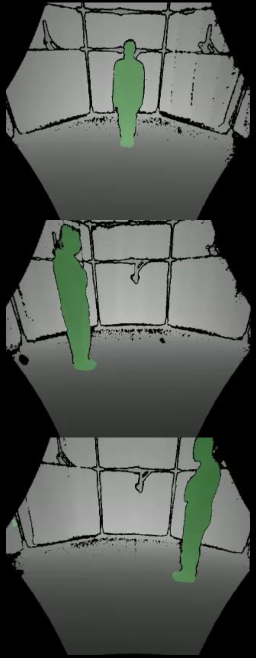


iPhone RGB

iPhone Depth



Kinect RGB



Kinect Depth

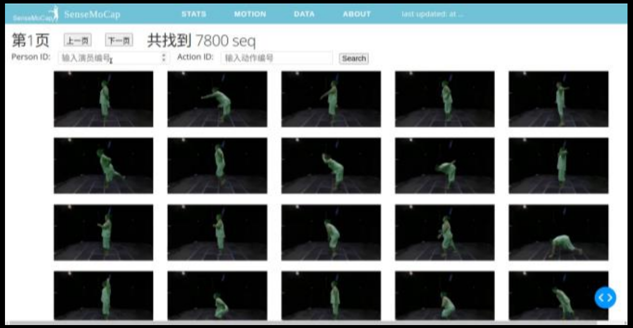
0.1mm
Accuracy

11
Cameras

1G
Data / Sec

6
Actor / Day

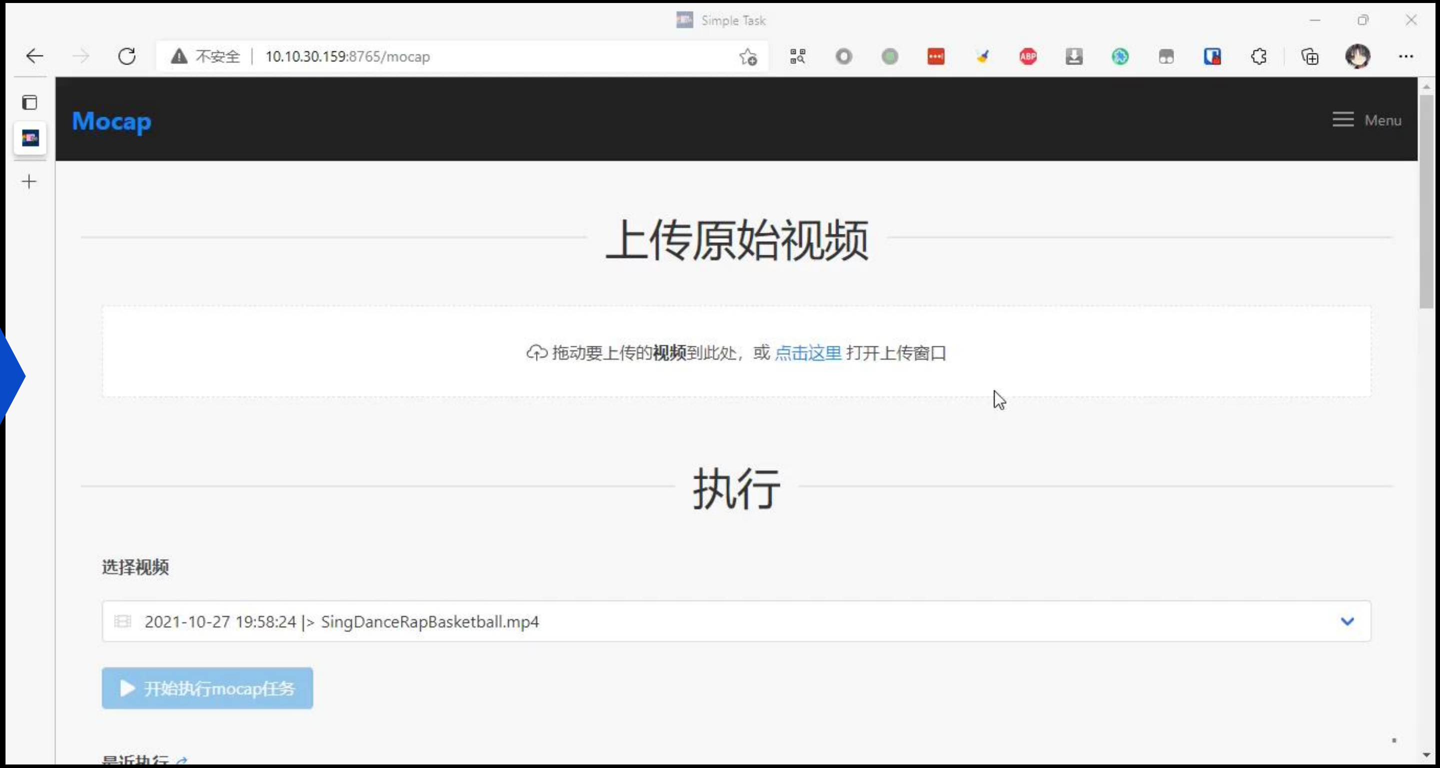
Search by Action



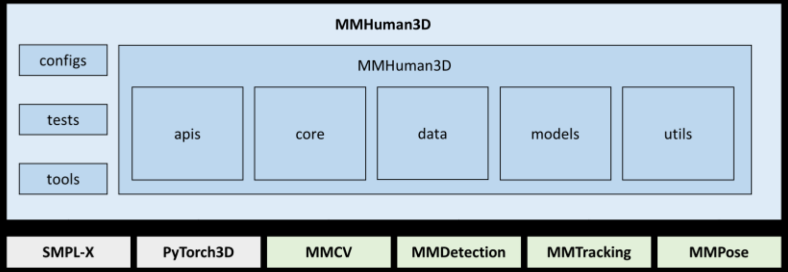
Search by Actor

MMHuman3D Software

Input Video



Online Motion Capture



3D Animation Production:

3 days -> 30 min

Visual AIGC

Avatar

Foundation Model

Object

Scene

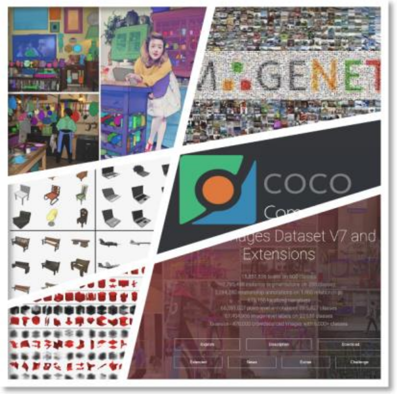
OmniObject3D: Text-to-3D Object

OmniObject3D is a **large-vocabulary** 3D dataset for **real-world scanned objects**.

- ✓ **6k** high-quality 3D models
- ✓ **190** categories
- ✓ **4** modalities: textured mesh, point cloud, real-captured video, synthetic multi-view images.
- ✓ Many down-stream tasks

Dataset	Year	Real	Full 3D	Video	Num Objs	Num Cats
ShapeNet	2015		√		51k	55
ModelNet	2014		√		12k	40
3D-Future	2020		√		16k	34
ABO	2021		√		8k	63
Toys4K	2021		√		4k	105
CO3D	2021	√		√	19k	50
DTU	2014	√	√		124	NA
GSO	2021	√	√		1k	17
AKB-48	2022	√	√		2k	48
Ours	2022	√	√	√	**6k**	**190**

Real-world 3D scans

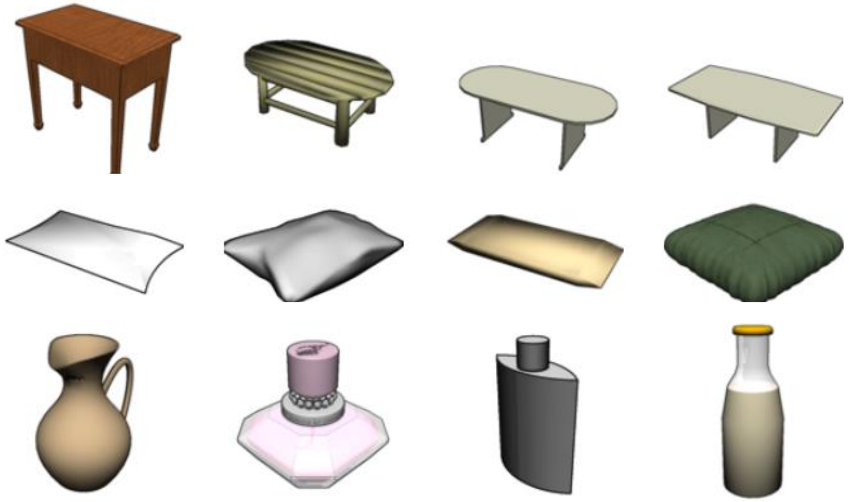


Background and motivation

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Synthetic data

ShapeNet
large in scale
low quality
not realistic



Multi-view images

CO3D
large in scale
No 3D GT

Real-world 3D scans

Google scanned objects
high quality
real-world scans
household objects

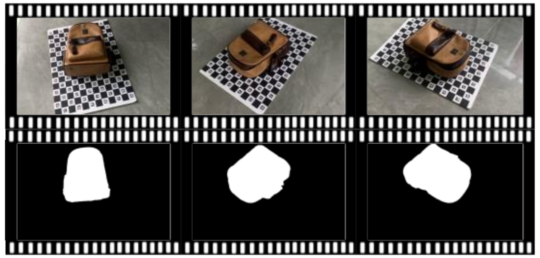
OmniObject3D

large-vocabulary
high quality
real-world scans

Overview

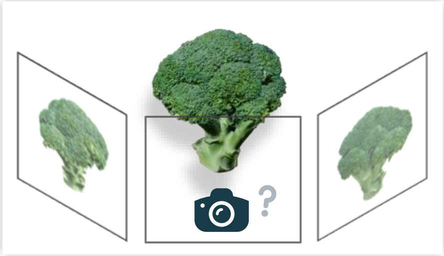
Textured meshes

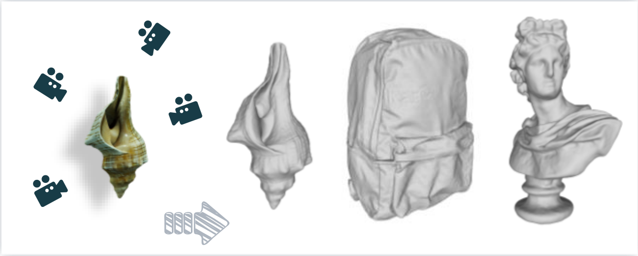
Point clouds Rendered images



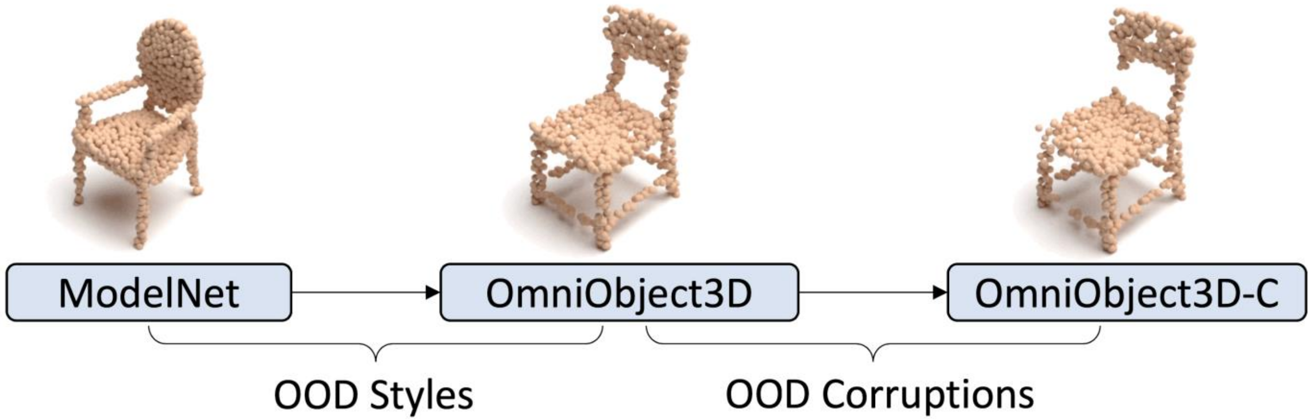
Real-captured videos

Perception | Novel View Synthesis | Surface Reconstruction | Generation





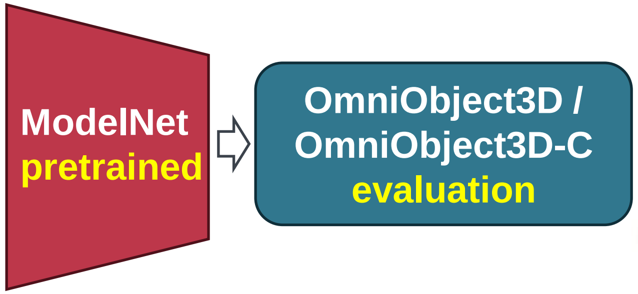
Robustness of point cloud classification



Differences between CAD models and real-scanned objects

Common corruptions

PointCloud-C (Ren et al. 2022)



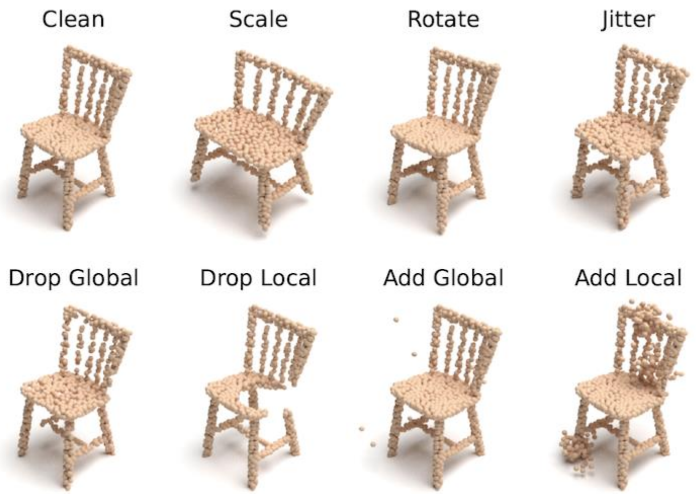


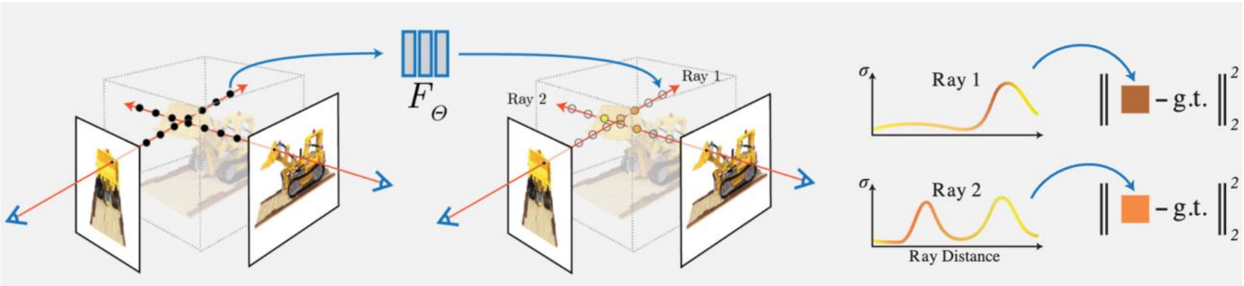
Table 2. **Point cloud perception robustness analysis on OmniObject3D with different architecture designs.** Models are trained on the ModelNet-40 dataset, with OA_{Clean} to be their overall accuracy on the standard ModelNet-40 test set. OA_{Style} on OmniObject3D evaluates the robustness to OOD styles. mCE on the corrupted OmniObject3D-C evaluates the robustness to OOD corruptions. Blue shadings indicate rankings. †: results on ModelNet-C [75]. Full results are presented in the supplementary materials.

	$mCE^{\dagger} \downarrow$	$OA_{Clean} \uparrow$	$OA_{Style} \uparrow$	$mCE \downarrow$
DGCNN [92]	1.000	0.926	0.448	1.000
PointNet [71]	1.422	0.907	0.466	0.969
PointNet++ [72]	1.072	0.930	0.407	1.066
RSCNN [51]	1.130	0.923	0.393	1.076
SimpleView [30]	1.047	**0.939**	0.476	0.990
GDANet [99]	<u>0.892</u>	0.934	<u>0.497</u>	**0.920**
PAConv [98]	1.104	0.936	0.403	1.073
CurveNet [97]	0.927	<u>0.938</u>	**0.500**	<u>0.929</u>
PCT [32]	0.925	0.930	0.459	0.940
RPC [75]	**0.863**	0.930	0.472	0.936

Novel view synthesis (two settings)

❑ *Single-scene optimization models*

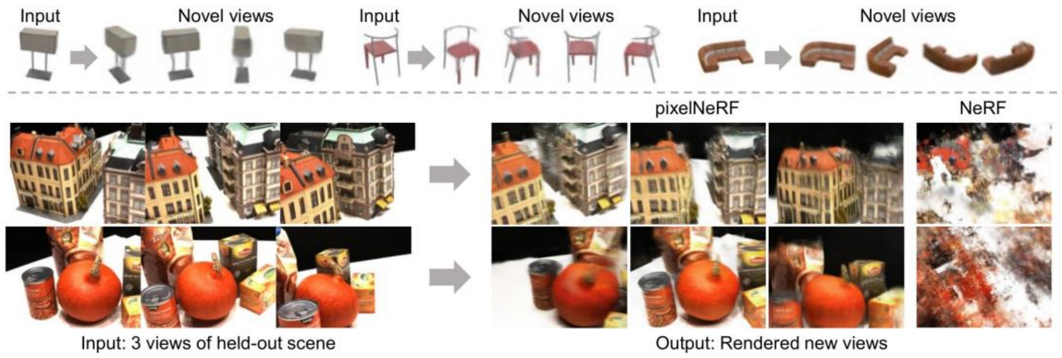
Multi-view images from **one scene** ⇨ **Train** ⇨ **The same scene Inference**



- NeRF (Mildenhall et al., 2021)
- Mip-NeRF (Barron et al., 2021)
- Plenoxels (Yu et al., 2021)

❑ *Generalizable models*

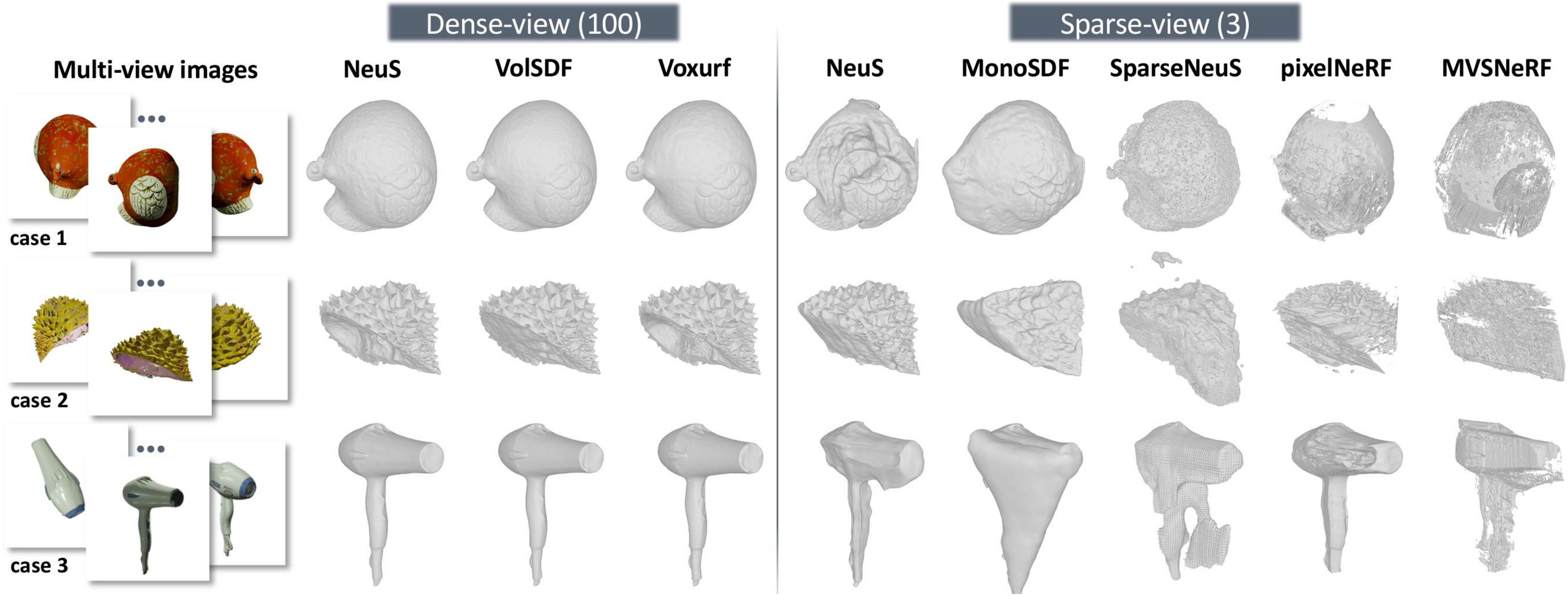
Multi-view images from scenes of **one category** or **across different categories.** ⇨ **Train** ⇨ **New scene with one or few views Inference**



- pixelNeRF (Yu et al., 2021)
- MVSNeRF (Chen et al., 2021)
- IBRNet (Wang et al., 2021)

Surface reconstruction (two settings)

❑ *Multi-view image surface reconstruction*



3D object generation

3D Object Generation

Interpolation across different categories

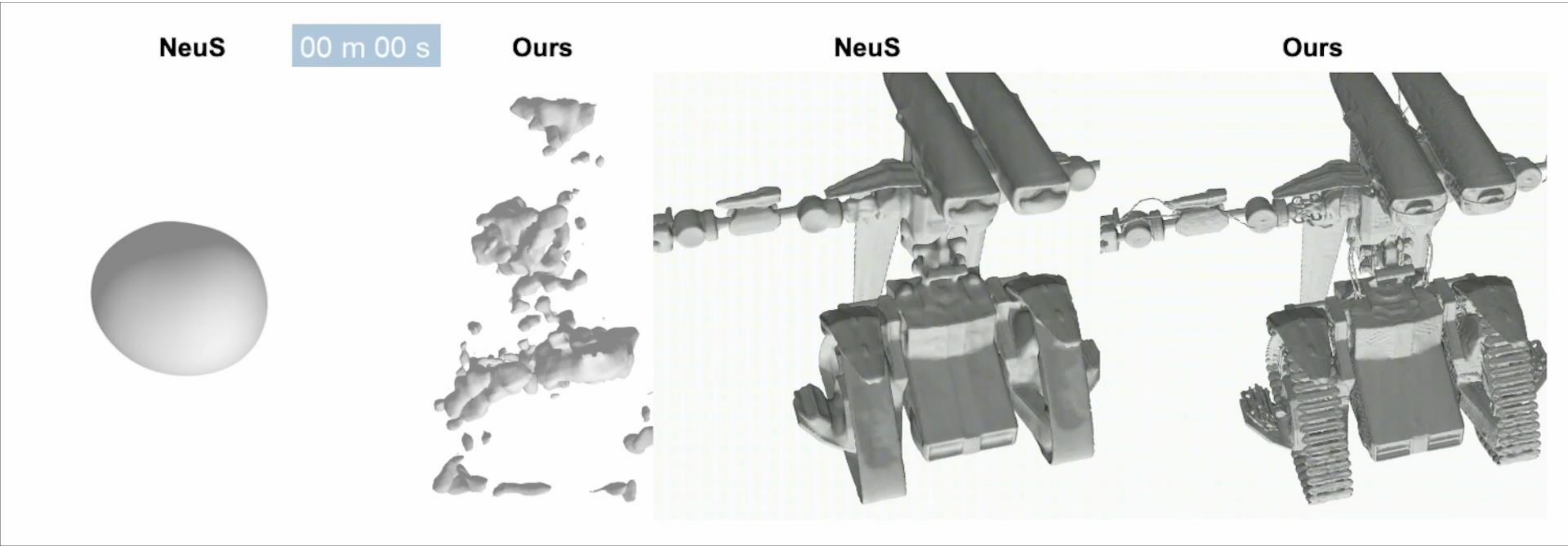
OmniObject3D: Text-to-3D Object

I want to generate a toy dinosaur.

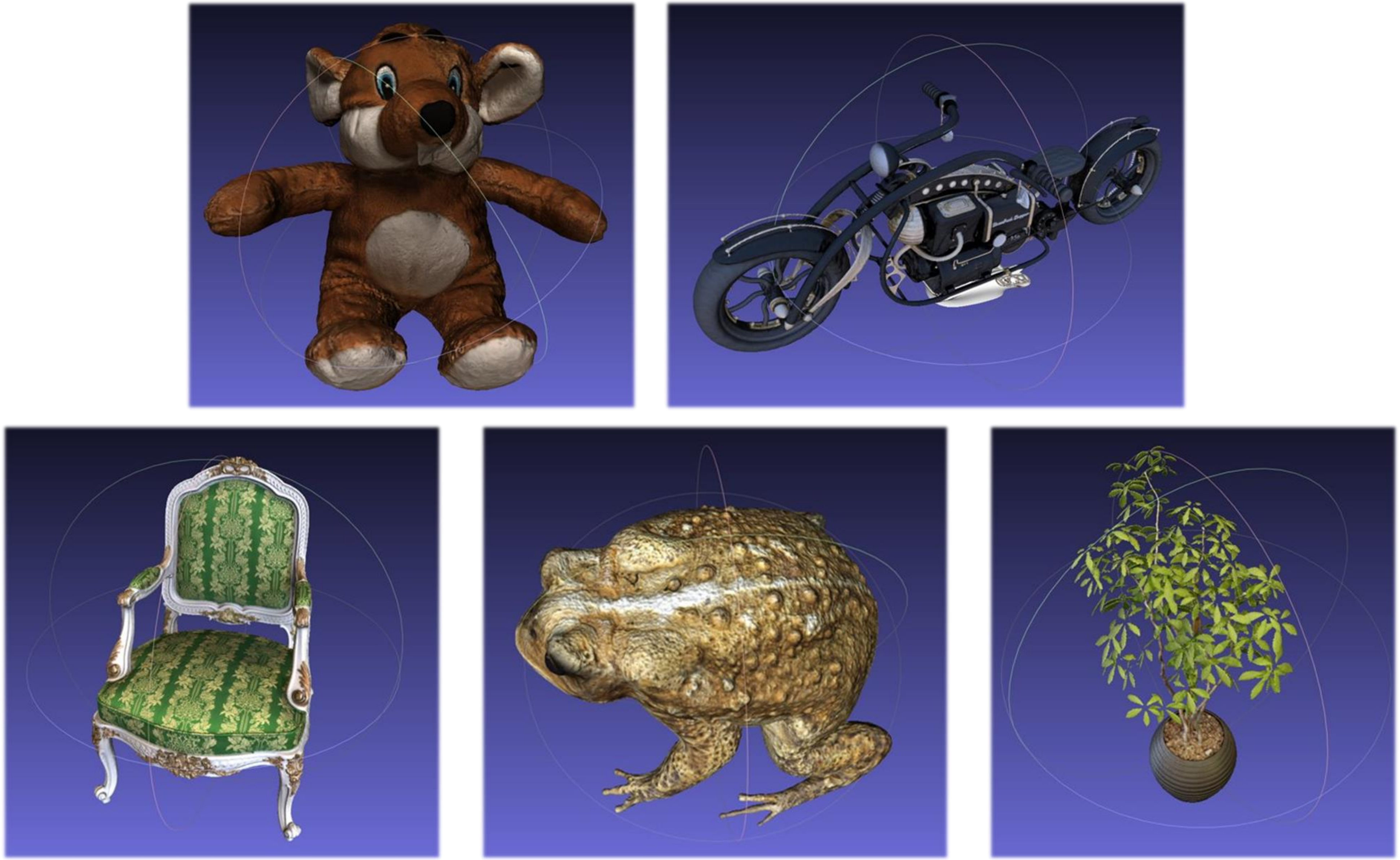
I want to generate a music box.

I want to generate a plaster statue.

Voxurf: Fast 3D Object Reconstruction



Voxurf: Fast 3D Object Reconstruction



Visual AIGC

Avatar

Foundation Model

Object

Scene

What about creating the environment?

The surrounding environment is also important to **an immersive VR experience.**

↓

Full field of view (360°) → Panorama

Realistic illuminations → HDR

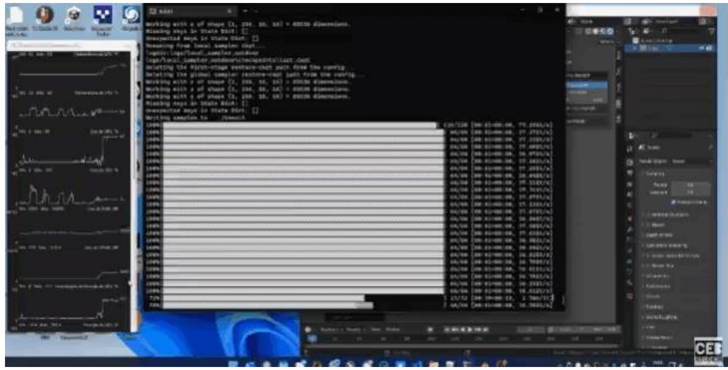
High-quality textures → 4K resolution

LDR – Low Dynamic Range, [0, 255]

HDR – High Dynamic Range, [0, +∞]

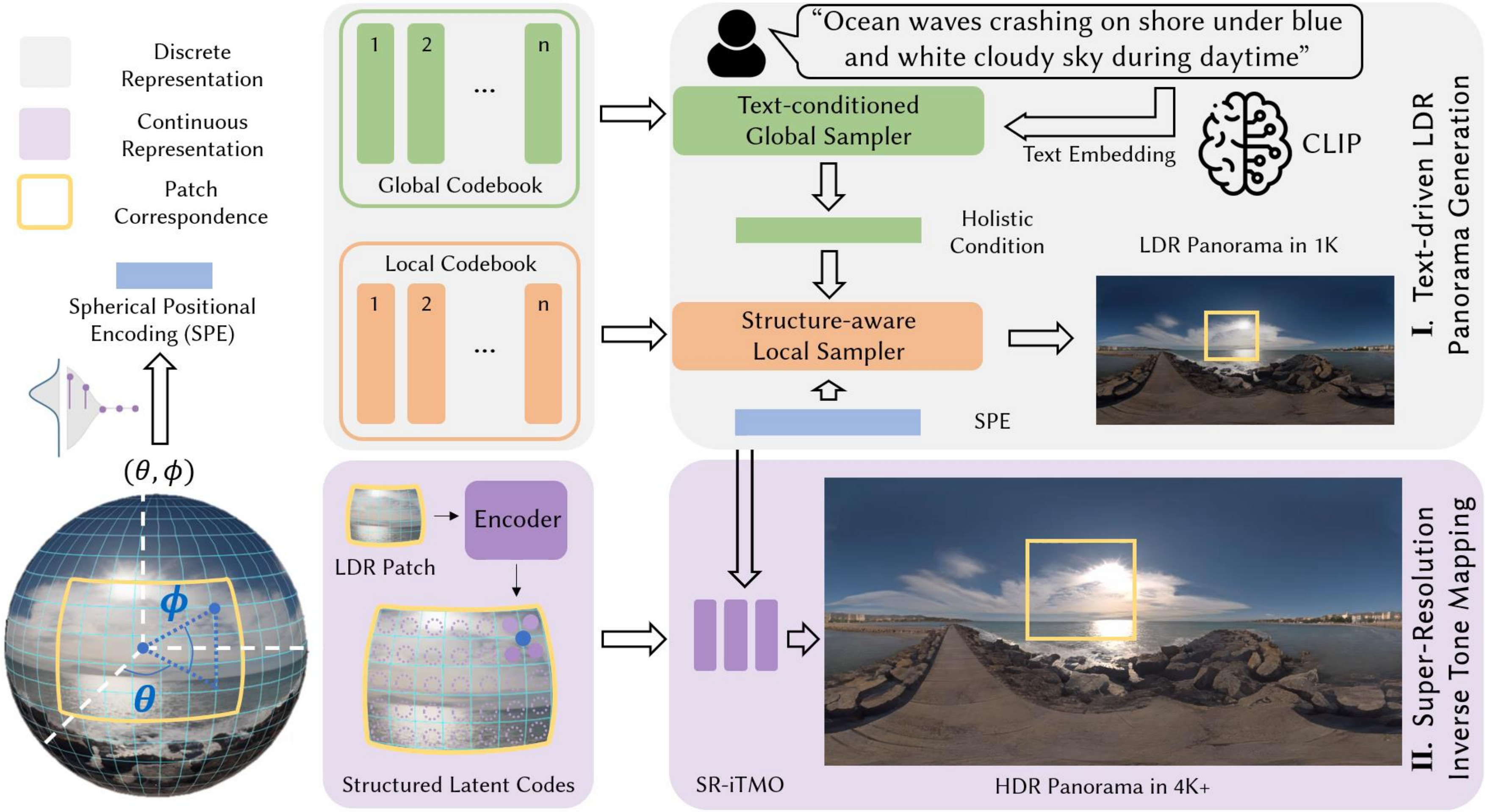
Text2Light: Text-to-3D Environment

"brown wooden dock on lake surrounded by green trees during daytime"



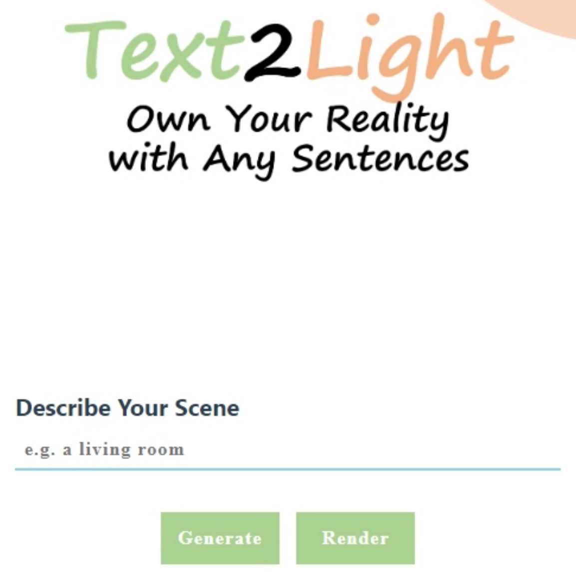
4K+ Resolution with High Dynamic Range

Text2Light
An Overview

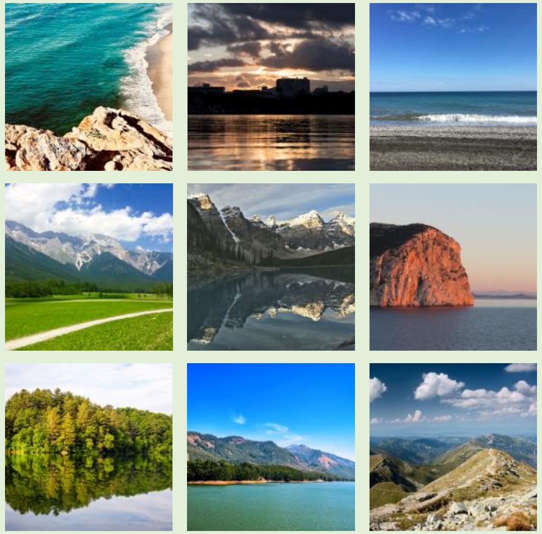


Suzanne Monkey: glossy Shader balls: glass, diffuse, glossy, mixture of diffuse and glossy

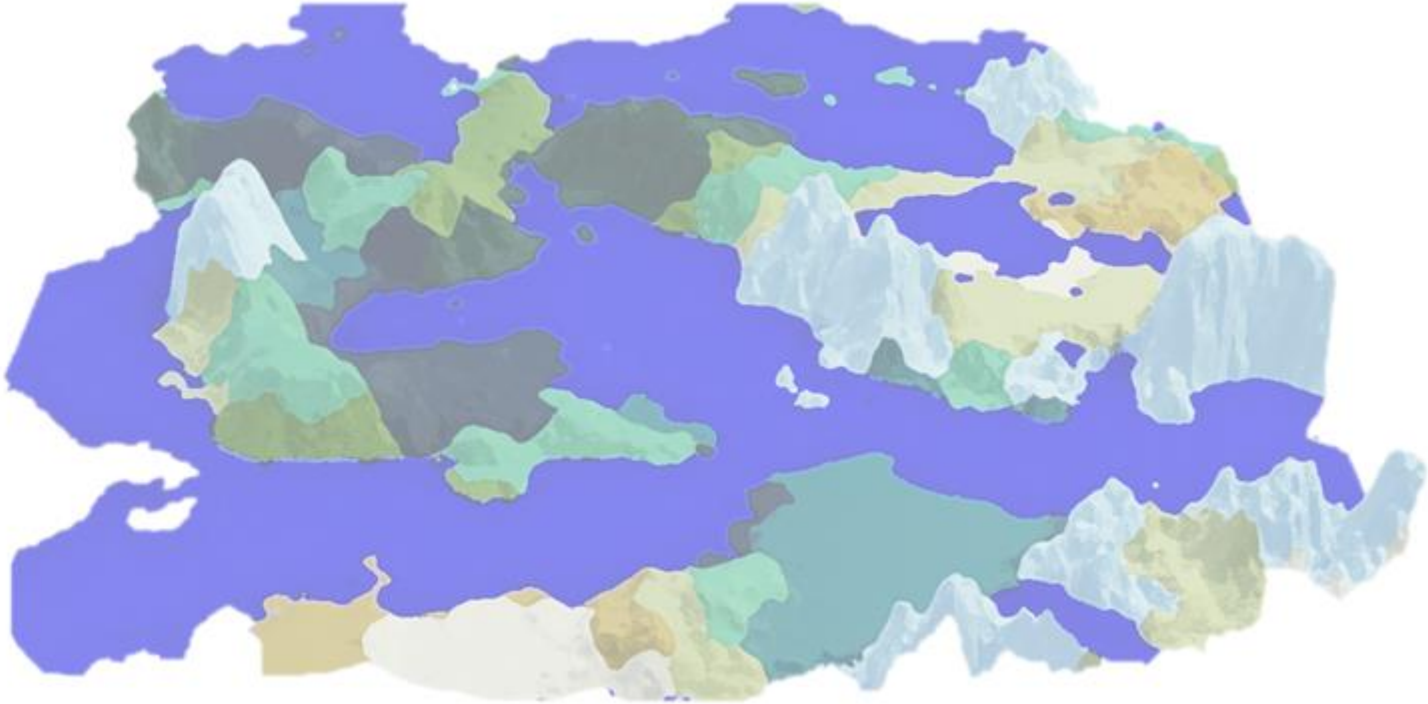
Text2Light: Text-to-3D Environment



SceneDreamer: Unbounded 3D Scene Generation

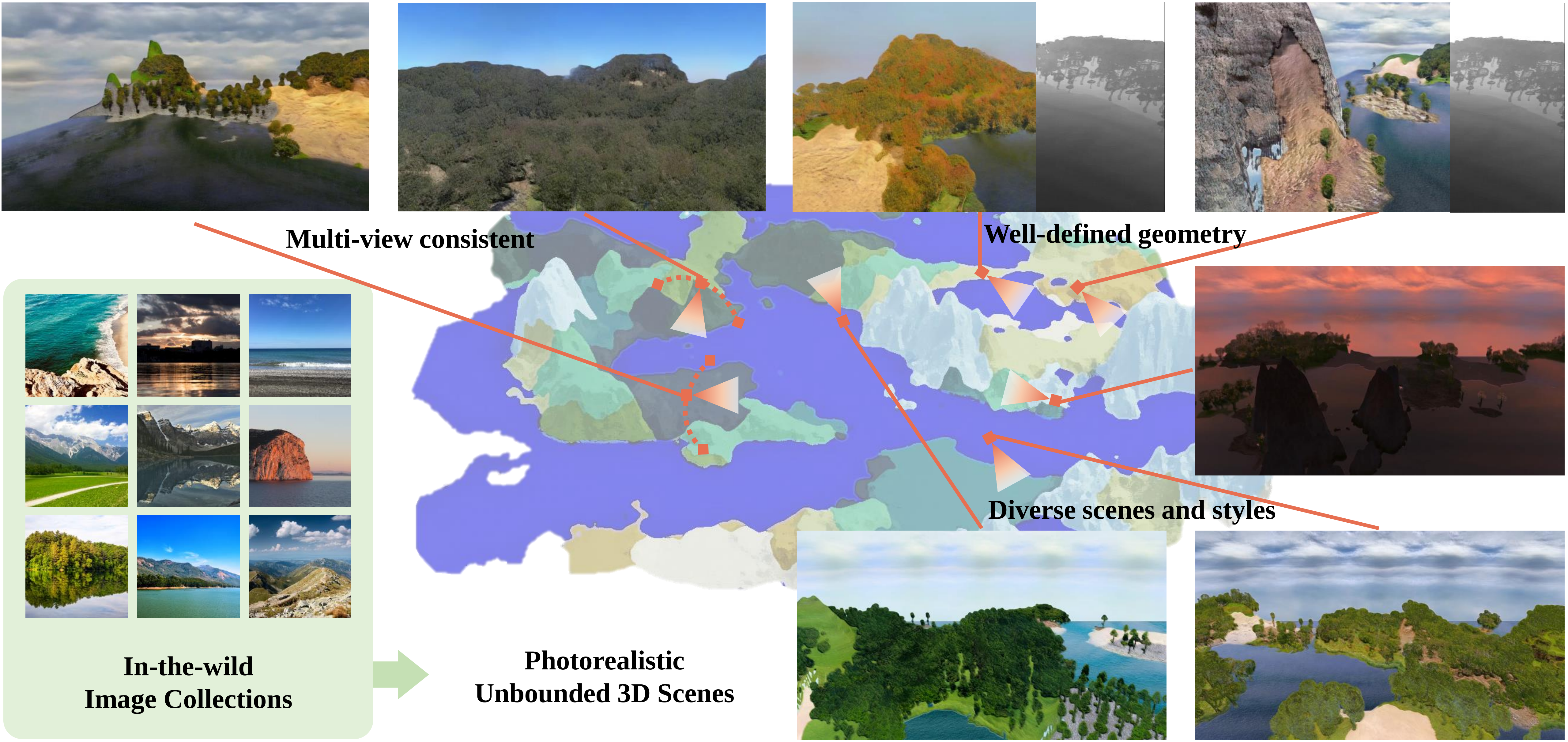


In-the-wild 2D
Image Collections



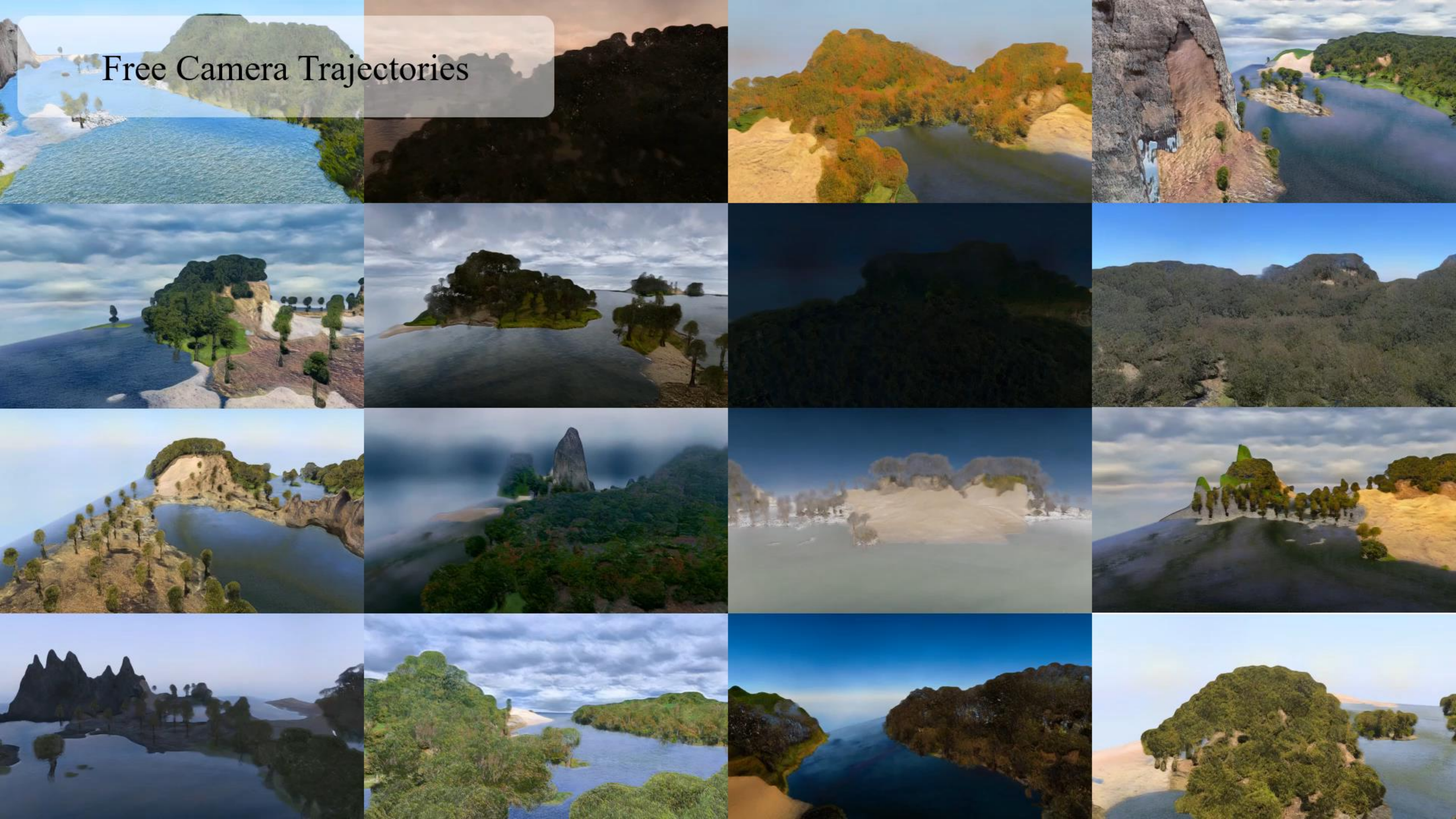
Photorealistic
Unbounded 3D Scenes

SceneDreamer: Unbounded 3D Scene Generation



Infinite 3D World!


Generate with Different Styles

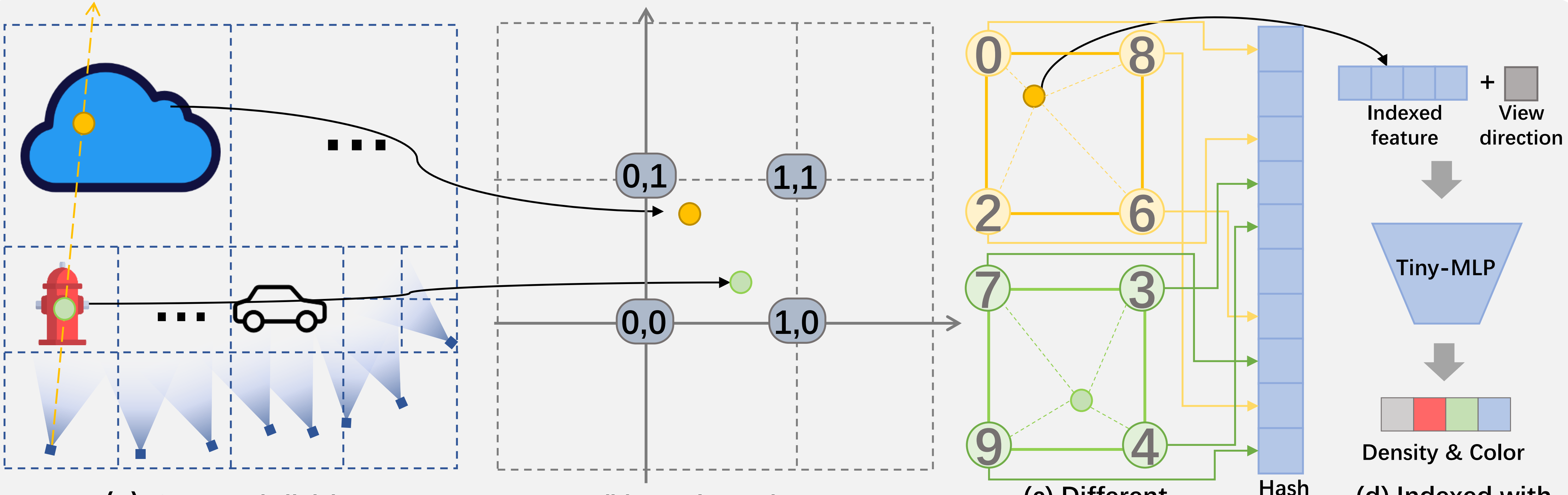

Free Camera Trajectories

F2NeRF: Mobile 3D Scene Reconstruction

What if the input camera trajectory is very irregular? – We call that a “free” trajectory

F2NeRF: Mobile 3D Scene Reconstruction

Adaptive warping method from input trajectories



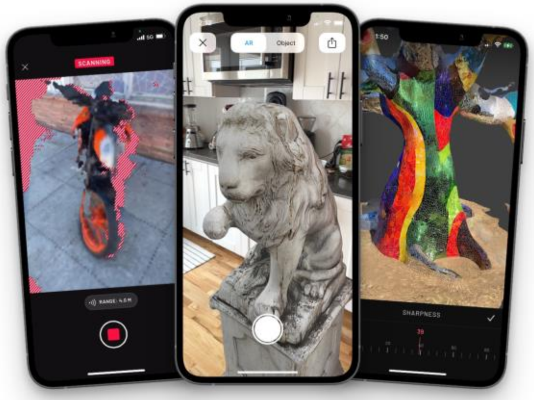
(a) Space subdivision

(b) Local warping

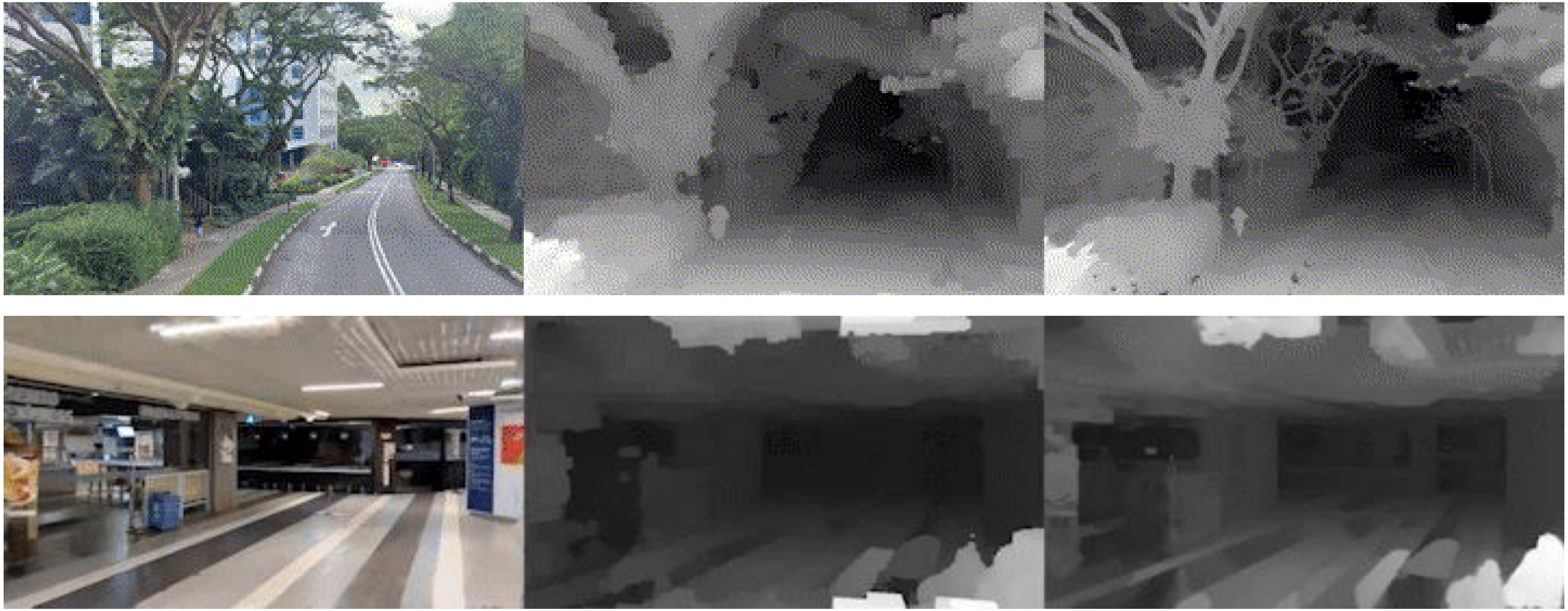
(c) Different Hash function

(d) Indexed with same Hash table

F2NeRF: Mobile 3D Scene Reconstruction



F2NeRF: Mobile 3D Scene Reconstruction



Visual AIGC

Avatar

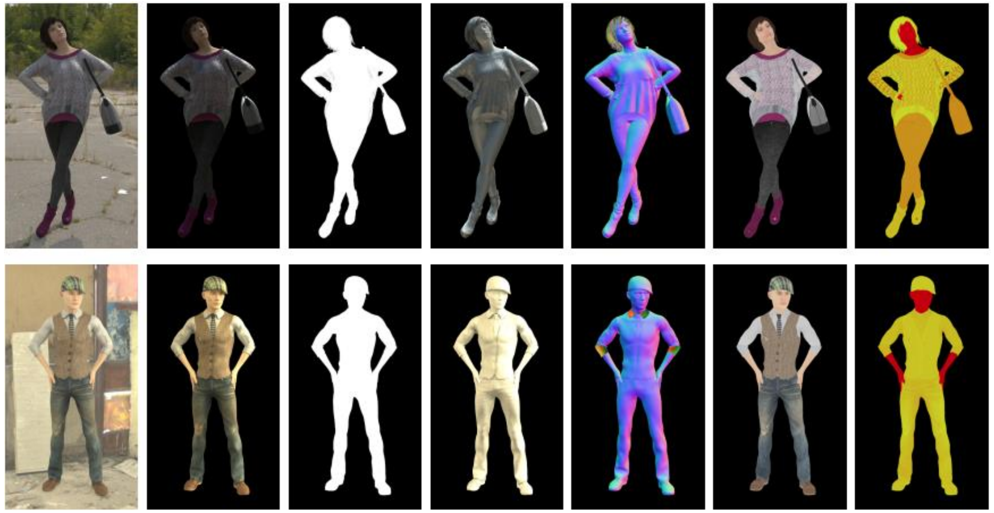
Foundation Model

Object

Scene

Relighting4D: Relightable 3D Human

Prior works



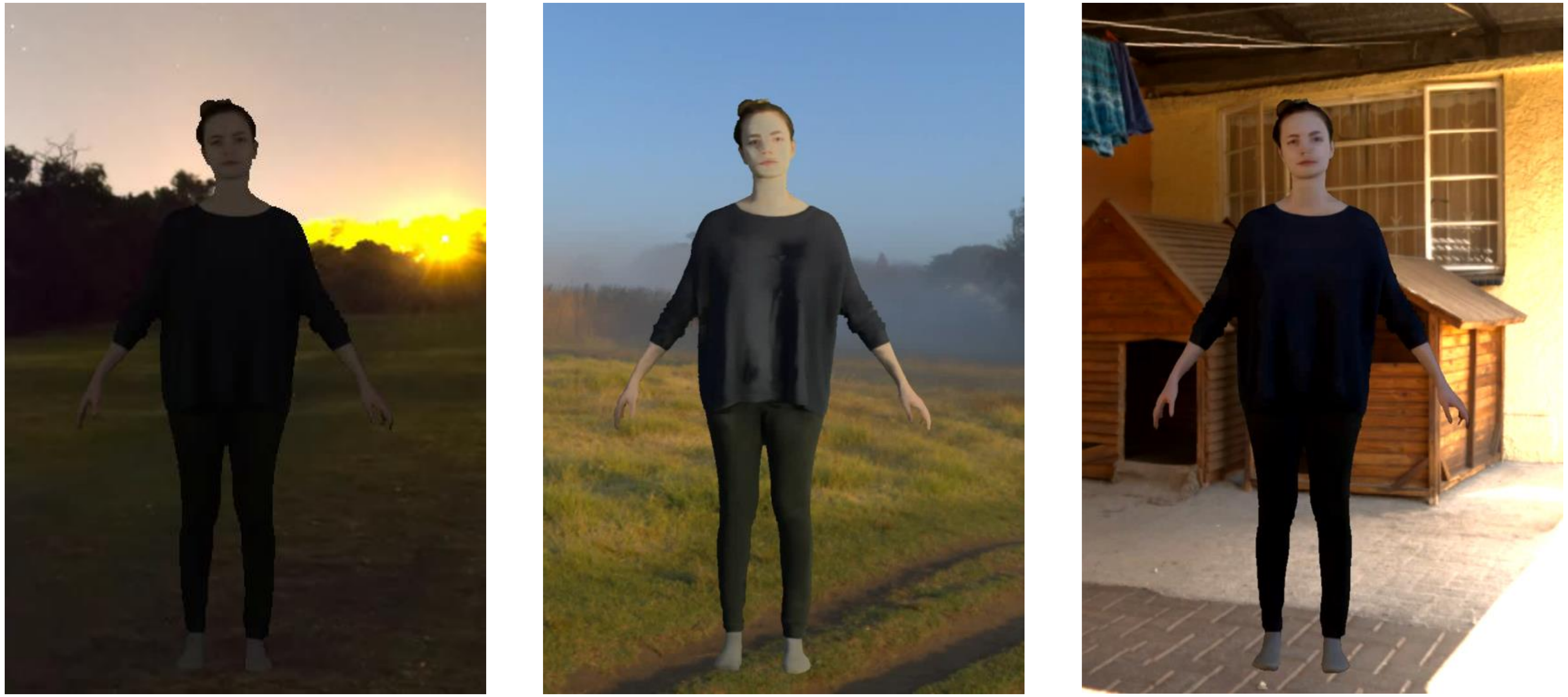
Synthetic dataset

Light Stage data

Relighting4D uses **only** videos to relight dynamic human actors from free viewpoints

Relighting4D: Relightable 3D Human

Video of human



Relight with different illuminations and free viewpoints

Visual AIGC

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Avatar

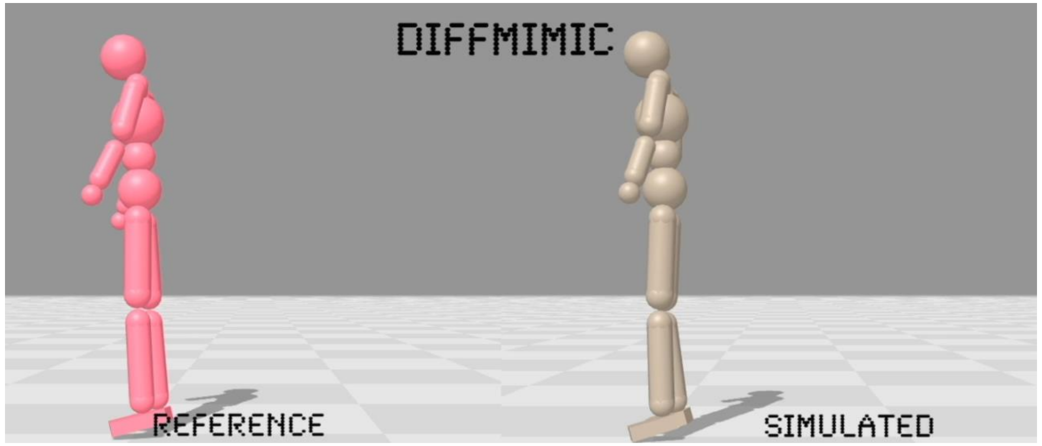
Foundation Model

Object

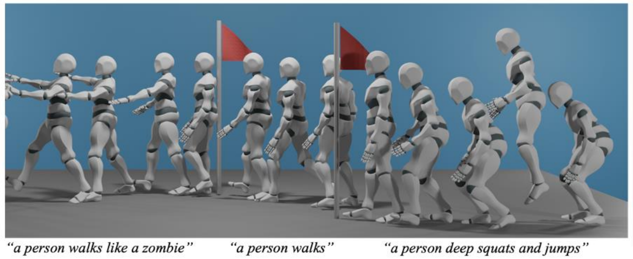
Scene

DiffMimic: Physically-Simulated Character

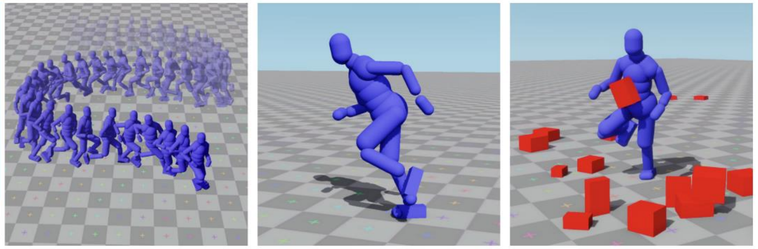
- Motion mimicking: let a **physically-simulated** character imitate a reference motion.



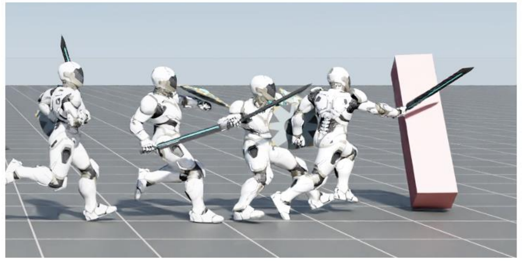
- A fundamental task for downstream animation applications.



Language-Conditioned Control



Responsive Control

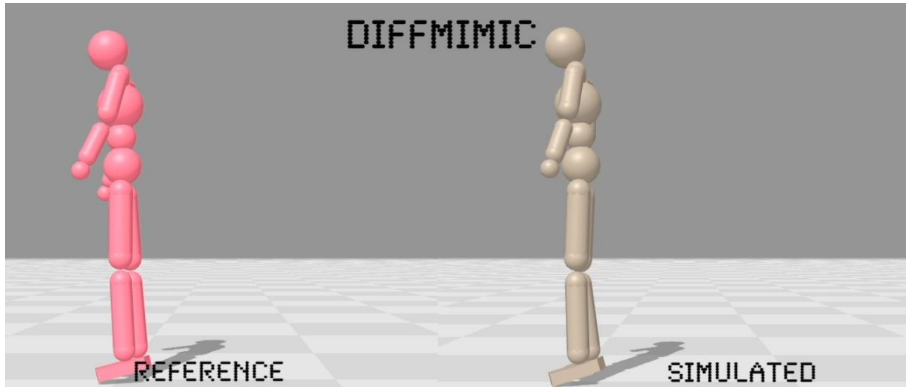


Skill Composition

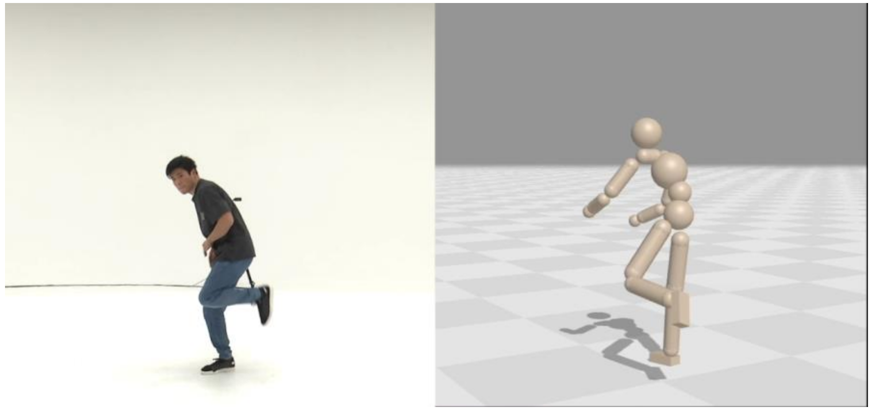
DiffMimic: Physically-Simulated Character

Motion	$T_{cycle}(s)$	DeepMimic	Spacetime Bound	Ours w/ RSI
Back-Flip	1.75	31.18	41.20 +32.1%	3.82 -87.7%
Cartwheel	2.72	30.45	17.35 -43.0%	4.72 -84.5%
Walk	1.25	23.80	4.08 -79.5%	1.55 -93.5%
Run	0.80	19.31	4.11 -78.7%	1.41 -92.7%
Jump	1.77	25.65	41.63 +77.8%	2.12 -91.7%
Dance	1.62	24.59	10.00 -59.3%	2.19 -91.1%

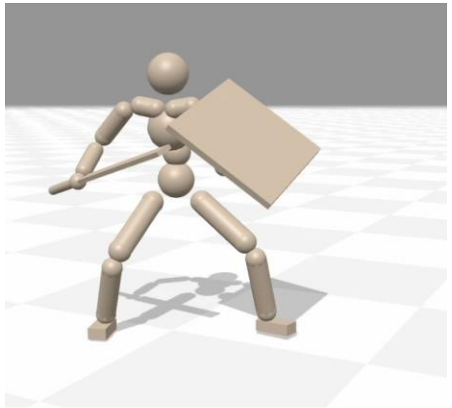
a) ~10x better sample efficiency compared to DeepMimic



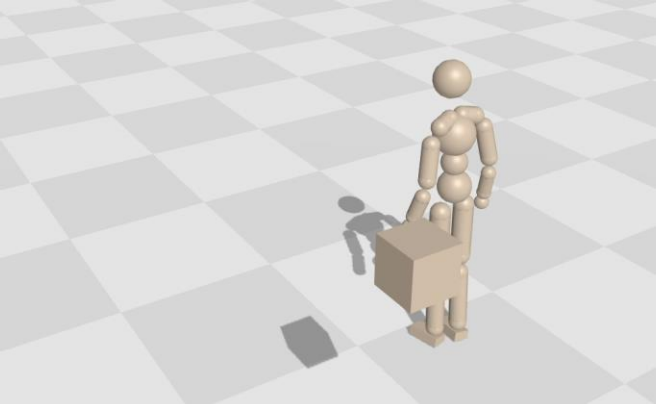
b) Learning backflip in 5 minutes



c) Scalable



d) General



e) Robust

Visual AIGC

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Avatar

Foundation Model

Object

Scene

ReVersion: Object Relation Generation

Input

Exemplar Images

“vegetable is contained inside paper bag”

Output

Relation Prompt

<R>

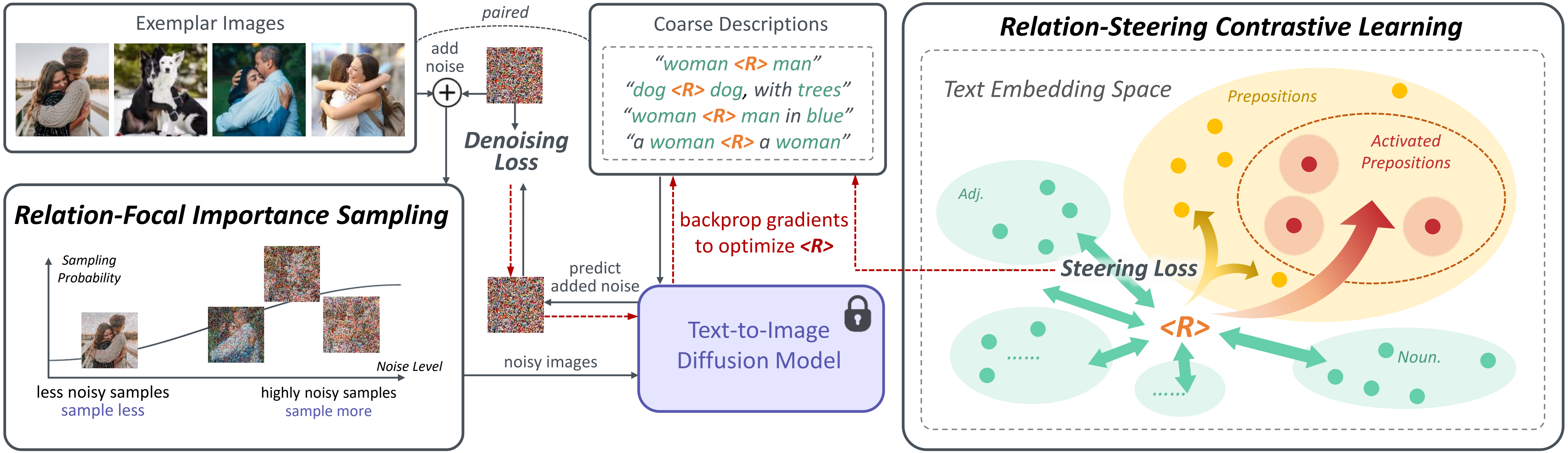
represent the co-existing relation in exemplar images

Application

Relation-Specific Text-to-Image Synthesis

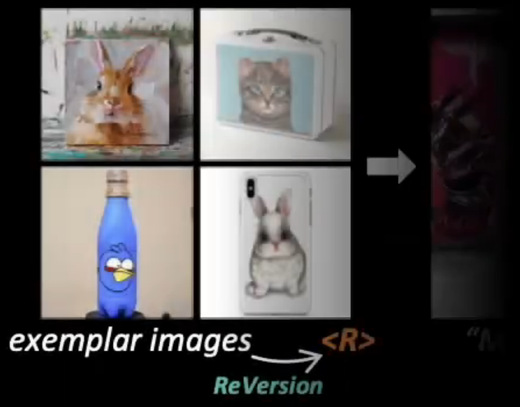
“Spiderman <R> paper bag”

ReVersion: Object Relation Generation



ReVersion: Object Relation Generation

Visual Results: ReVersion



Visual AIGC

Avatar

Thank You!

Object

Scene