Multimodal Understanding & Generation with Efficiently Finetuned Foundation Models

Dr. Long Chen

Assistant Professor, CSE, HKUST

Feb. 2025

In Collaboration with: Yanghao, Wei, Lin, Zhen, Ziqi, Hongxiang (HKUST)
Yuxuan (NTU), Haoxuan (Columbia), Hongzhan (HKBU), Jiazuo (DLUT)

There are lots of pretrained foundation models…

- Large language models (LLMs)
  - ChatGPT, GPT-4,…
- Vision language models
  - CLIP, BLIP, …
- Visual generation models
  - Stable Diffusion, …
- ***Research Q1***: *How can we efficiently train or finetune foundation models.*
- ***Research Q2***: *How can we build strong open-world multimodal understanding and generation models with these pretrained foundation models*

Efficient Finetune Foundation Models

Parameter-Efficient Tuning

Memory-Efficient Tuning

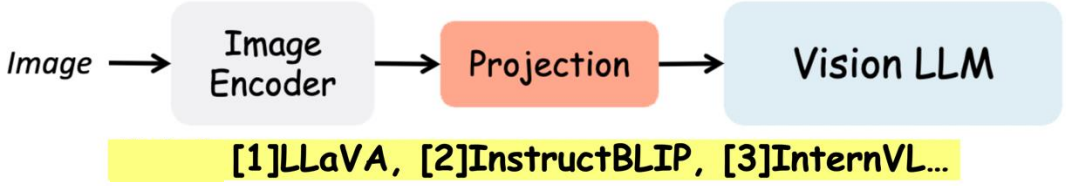
Modality-Efficient Tuning

LLMs Can Evolve Continually on Modality for 𝕏-Modal Reasoning

Jiazuo Yu[1], **Haomiao Xiong**[1], **Lu Zhang**[1,*], **Haiwen Diao**[1], **Yunzhi Zhuge**[1], **Lanqing Hong**[2], **Dong Wang**[1], **Huchuan Lu**[1], **You He**[3], **Long Chen**[4]
[1]Dalian University of Technology, [2]Huawei Noah's Ark Lab
[3]Tsinghua University, [4]The Hong Kong University of Science and Technology

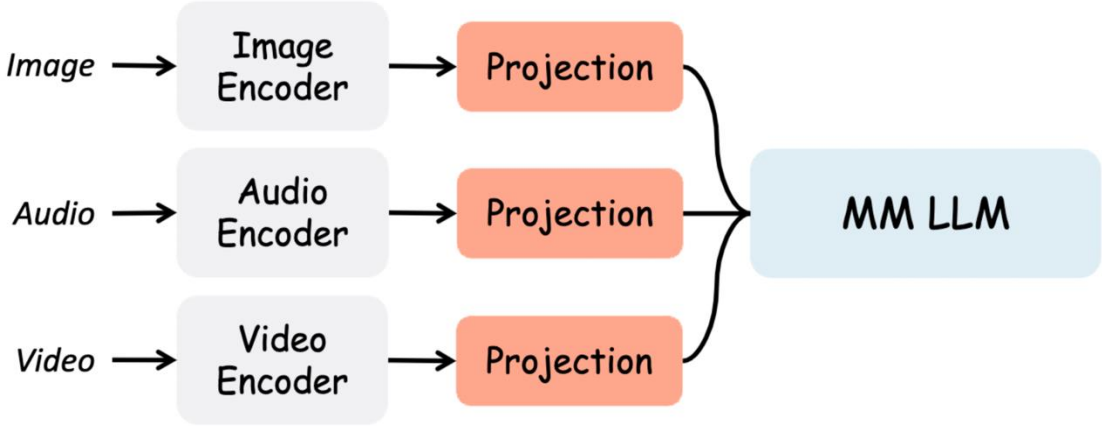
https://arxiv.org/pdf/2410.20178
(NeurIPS'24)

Prior Work on Multimodal LLMs



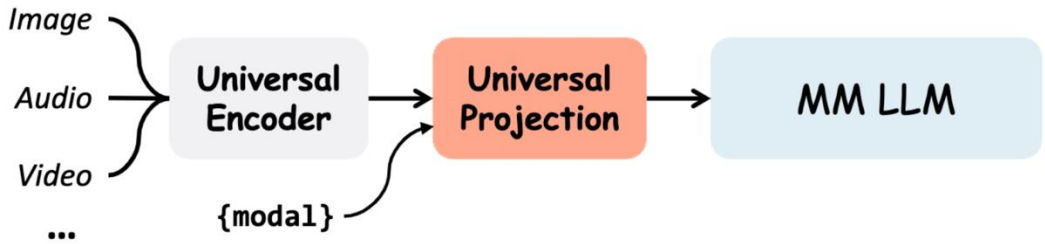
[1]LLaVA, [2]InstructBLIP, [3]InternVL…

- Single-modality MLLM (only image modality)



[4]X-InstructBLIP, [5]X-LLM, [6]ChatBridge…

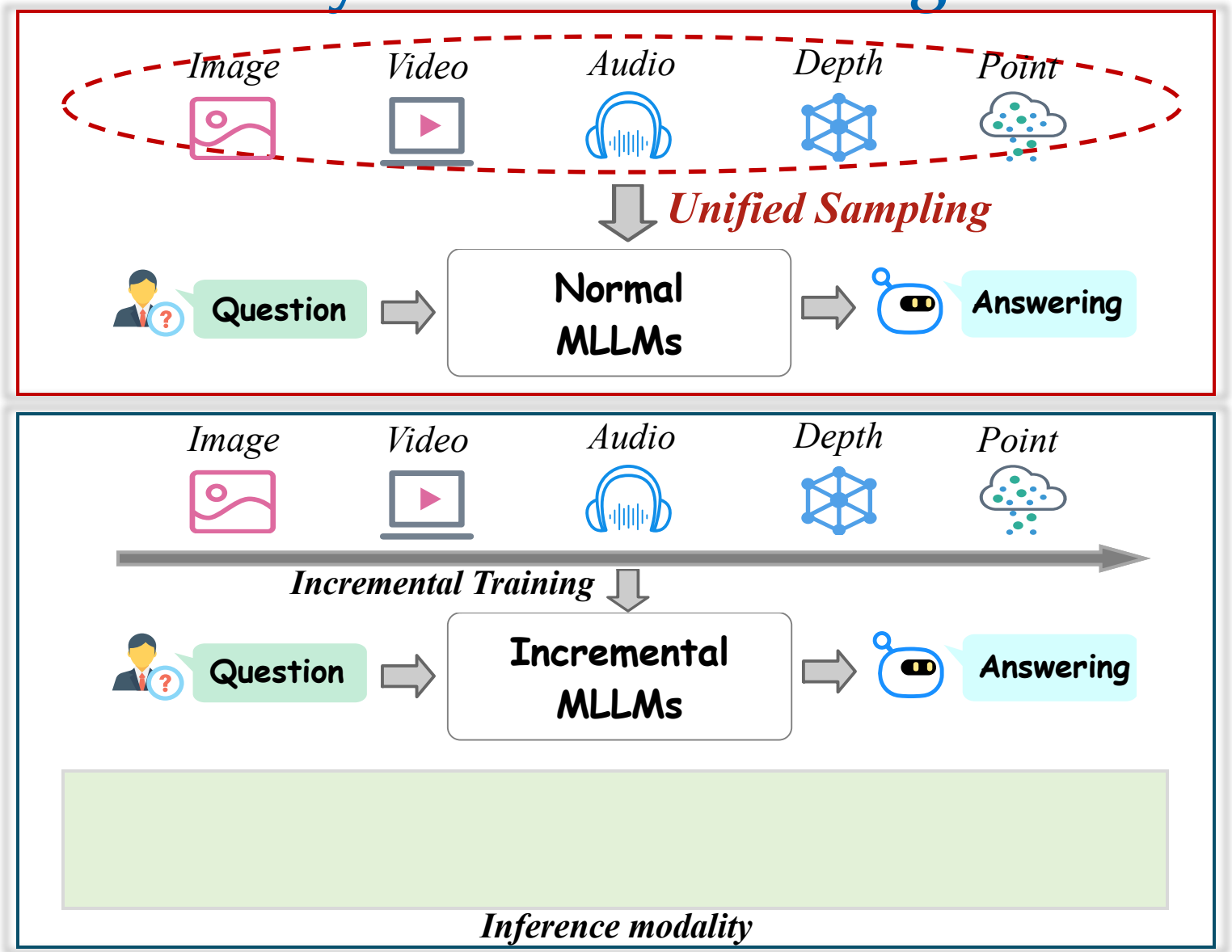
- Multiple-modality MLLM (different projector parameters are different)



[7]OneLLM…

- Multiple-modality MLLM (mixture multi-modal data for training)

Modality-Efficient Training for X-Modal Reasoning



Exisiting work

Limited data

Hard to extend

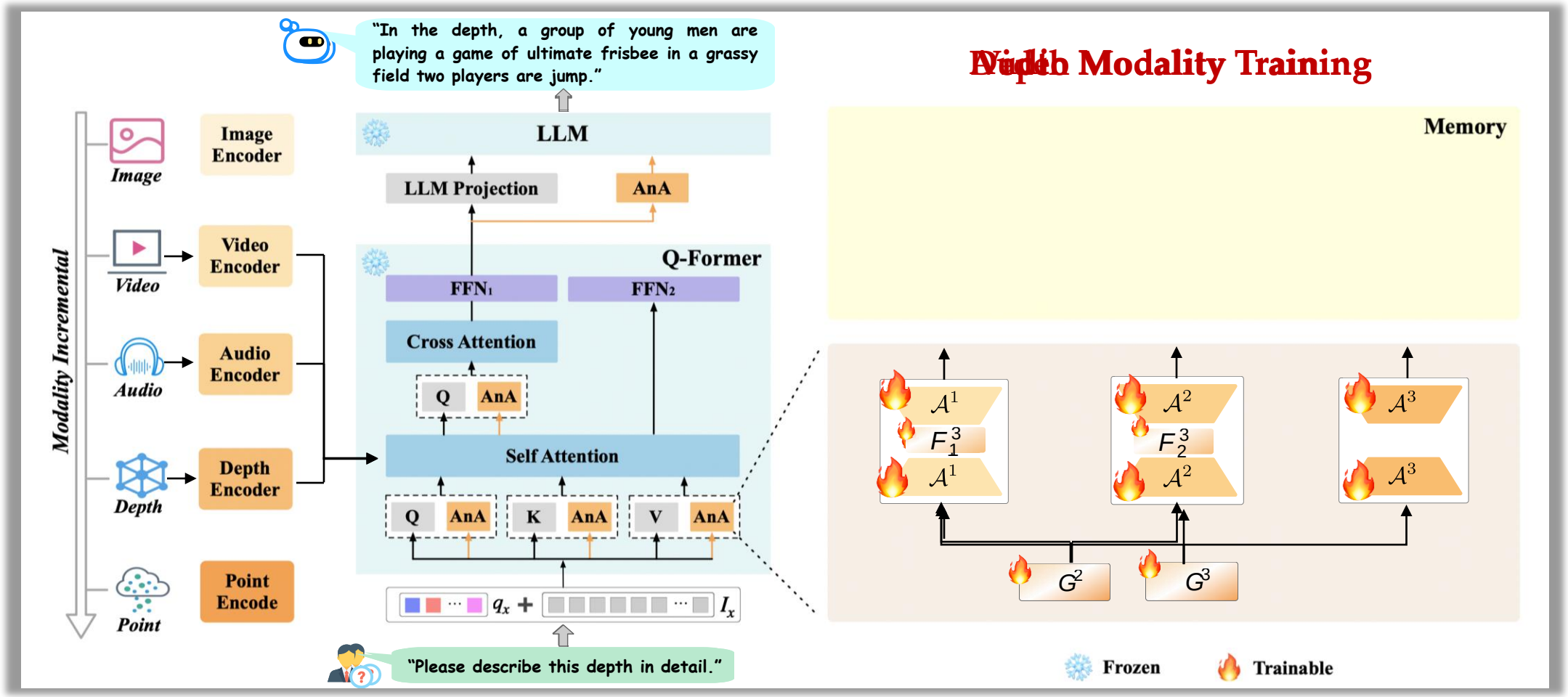
Ours

Continual training

Start from different checkpoints

Jiazuo Yu, et al. LLMs can Evolve Continually on Modality for X-Modal Reasoning. In NeurIPS, 2024.

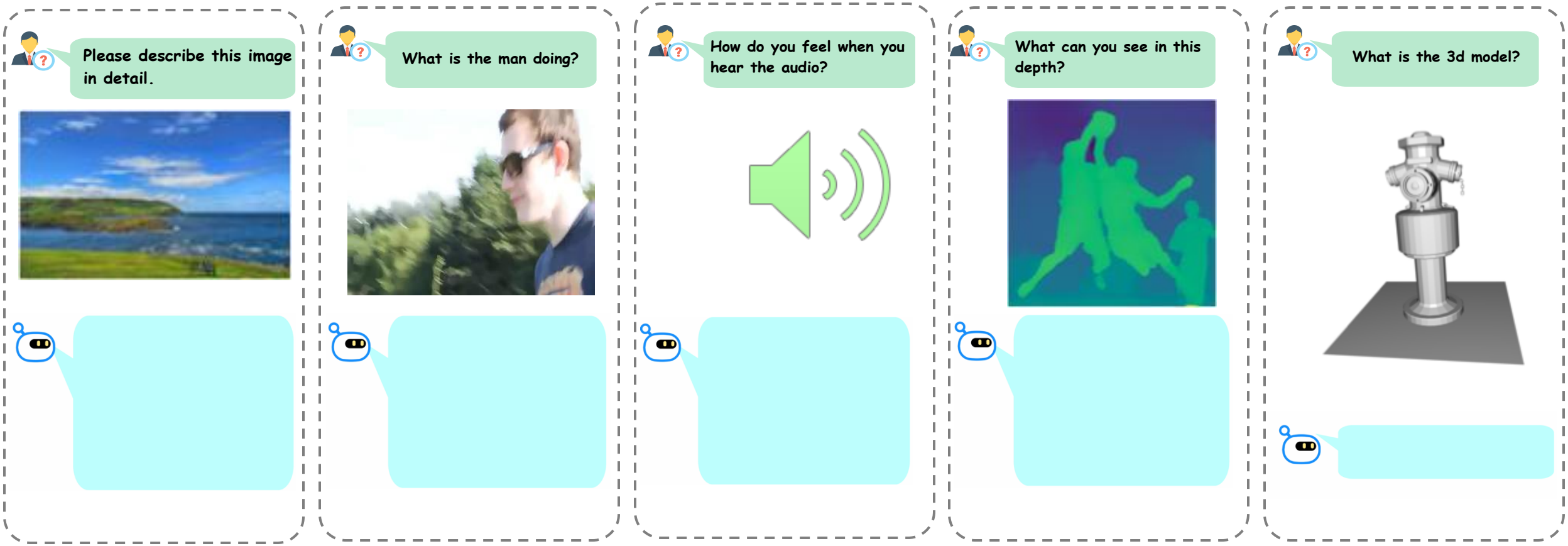
Modality-Efficient Training for X-Modal Reasoning



Jiazuo Yu, et al. LLMs can Evolve Continually on Modality for X-Modal Reasoning. In NeurIPS, 2024.

Modality-Efficient Training for X-Modal Reasoning

- PathWeave: It can performs well on all previous trained modalities



Jiazuo Yu, et al. LLMs can Evolve Continually on Modality for X-Modal Reasoning. In NeurIPS, 2024.

Efficient Finetune Foundation Models

Parameter-Efficient Tuning

Memory-Efficient Tuning

Modality-Efficient Tuning

Multimodal Understanding & Generation with Efficient Finetune Foundation Models

Open-World Perception

Parameter-Efficient Tuning

Memory-Efficient Tuning

Modality-Efficient Tuning

Efficiently Finetuned Foundation Models

Inversion Circle Interpolation:
Diffusion-based Image Augmentation for Data-scarce Classification

Yanghao Wang, Long Chen
The Hong Kong University of Science and Technology

https://arxiv.org/abs/2408.16266
(Under Review)

Zero-shot Visual Relation Detection via Composite Visual Cues from Large Language Models

---

Lin Li[1,2], Jun Xiao[1], Guikun Chen[1], Jian Shao[1], Yueting Zhuang[1], Long Chen[2]*
[1]Zhejiang University [2]The Hong Kong University of Science and Technology

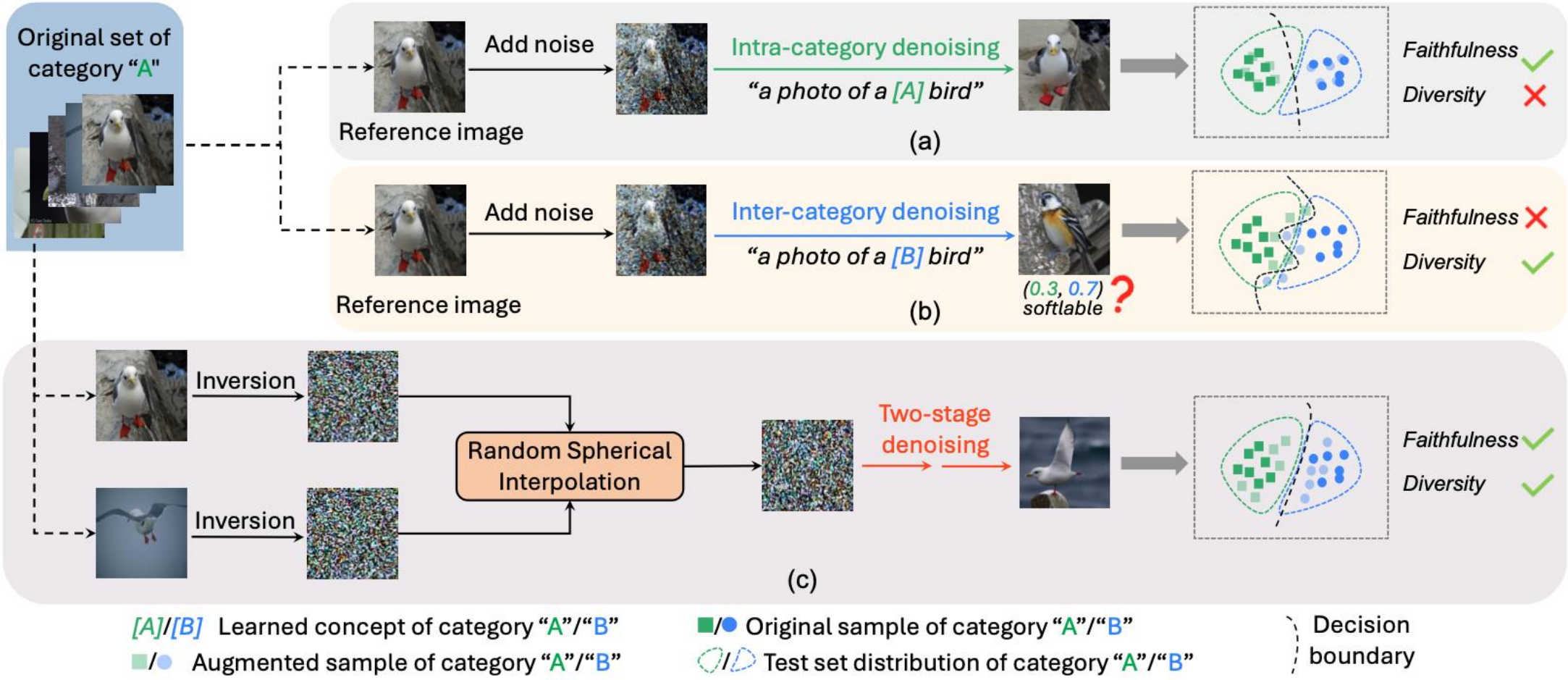
https://arxiv.org/abs/2305.12476
(NeurIPS'23)

Diffusion Model for Closed-set Perception

- **Diffusion model is a text-to-image generation model.**
- We can use diffusion model to conduct data augmentation.

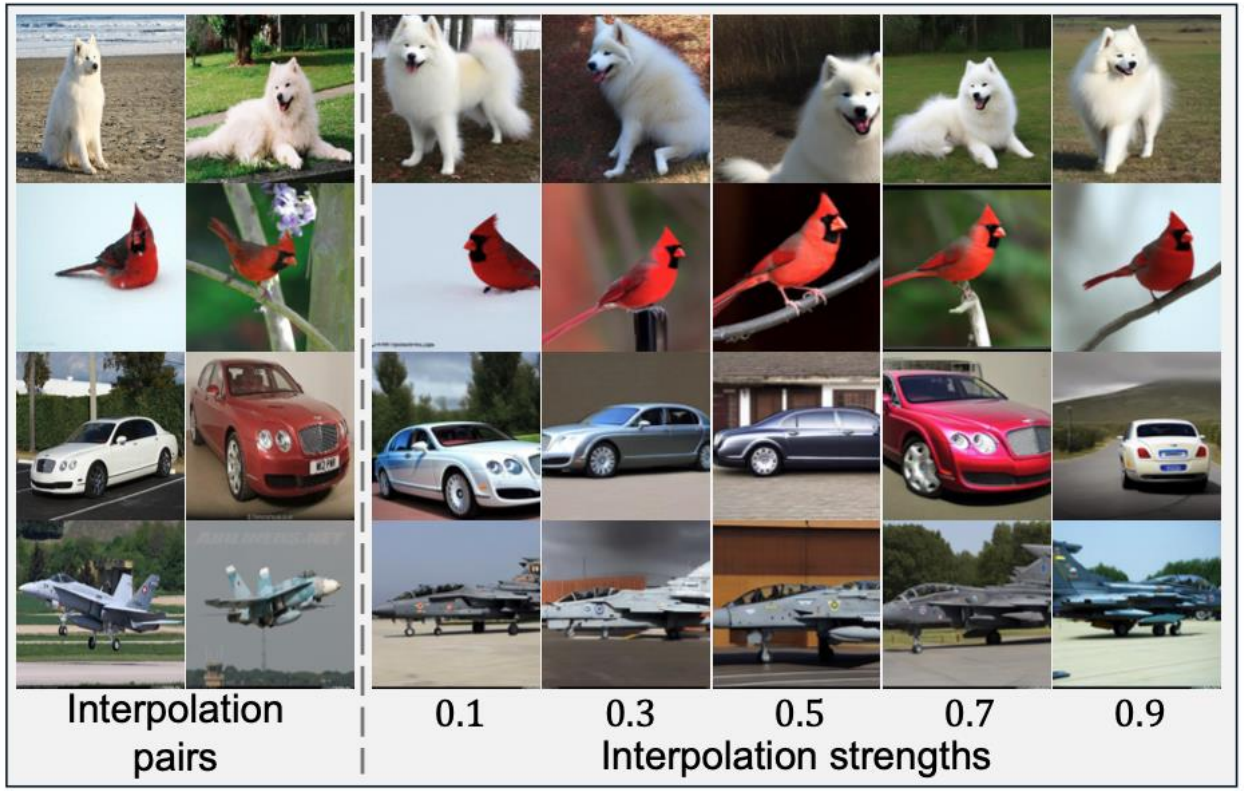
Y. Wang, et al. Inversion Circle Interpolation: Diffusion-based Image Augmentation for Data-scarce Classification. Under Review.

Diffusion Model for Data Augmentation

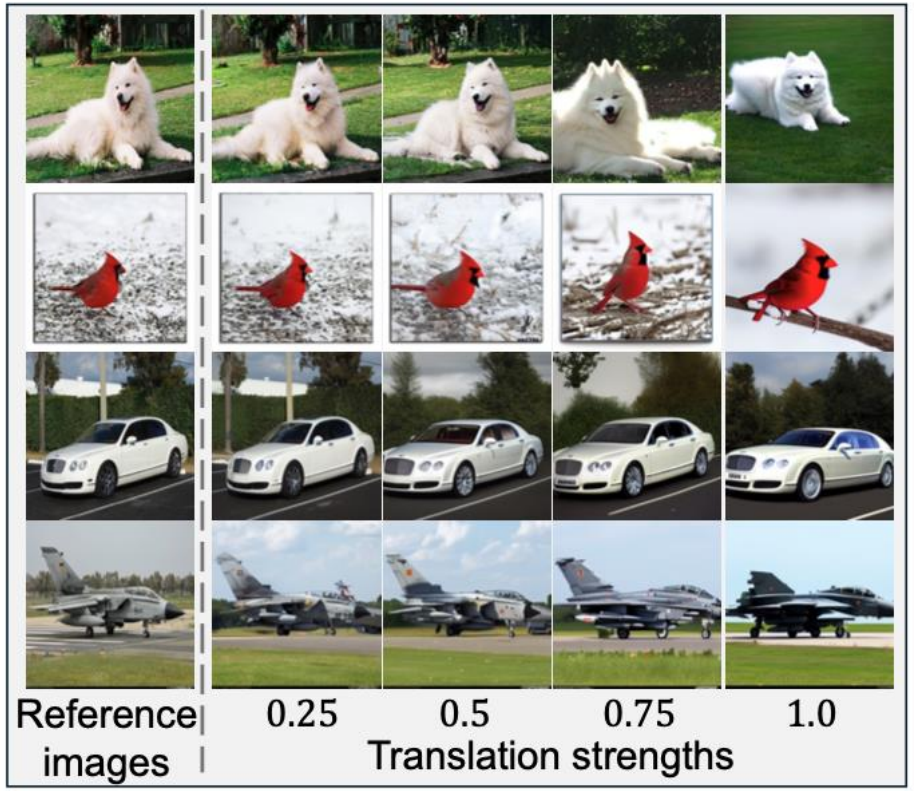


Y. Wang, et al. Improving Diffusion-Based Data Augmentation with Inversion Spherical Interpolation. Under Review.

Diffusion Model for Data Augmentation



Ours

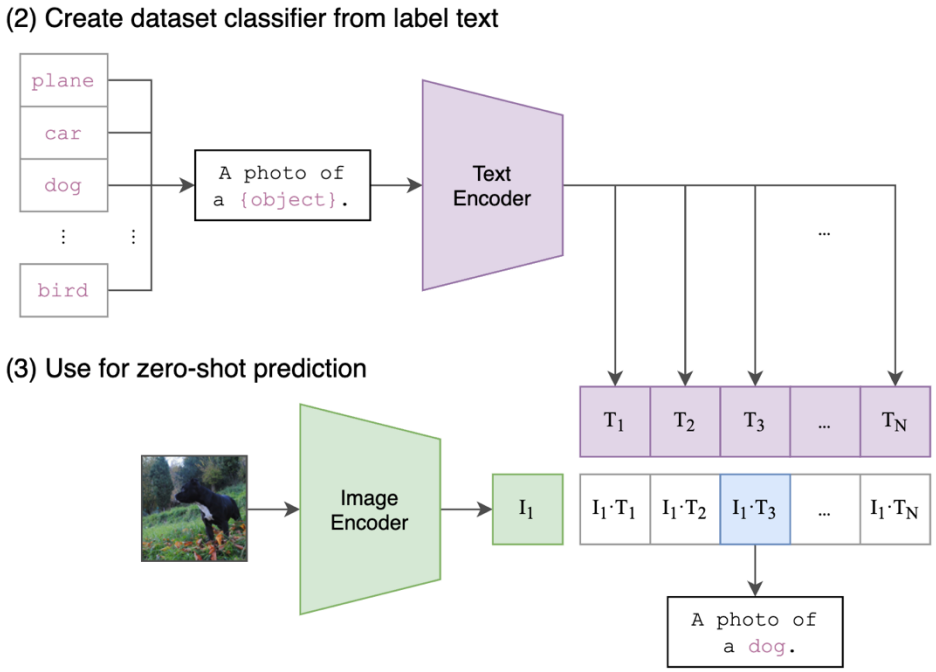


DA-Fusion

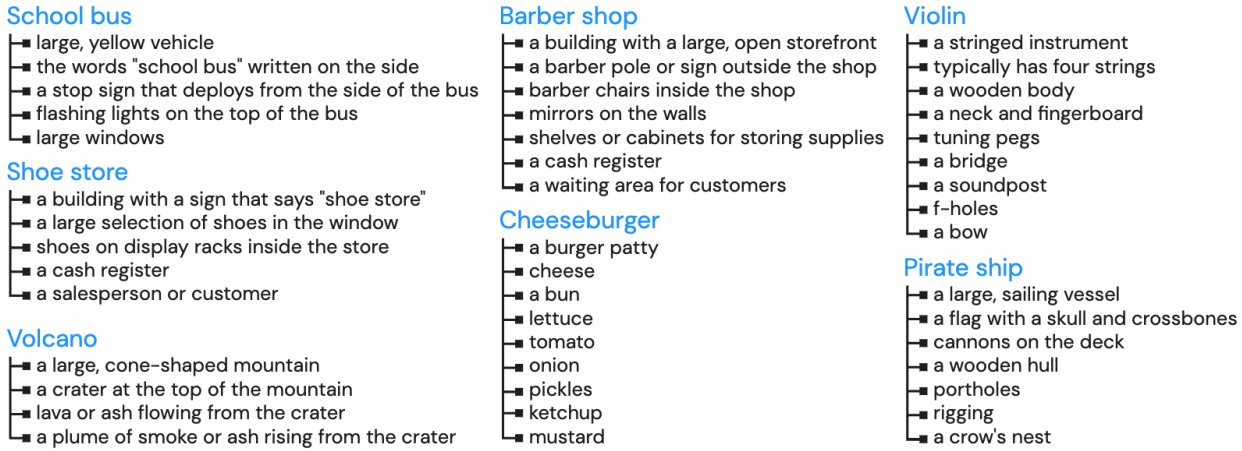
Y. Wang, et al. Improving Diffusion-Based Data Augmentation with Inversion Spherical Interpolation. Under Review.

LLM + CLIP for Zero-Shot Perception

- **LLMs can generate detailed descriptions to help zero-shot classification**



CLIP for zero-shot classification

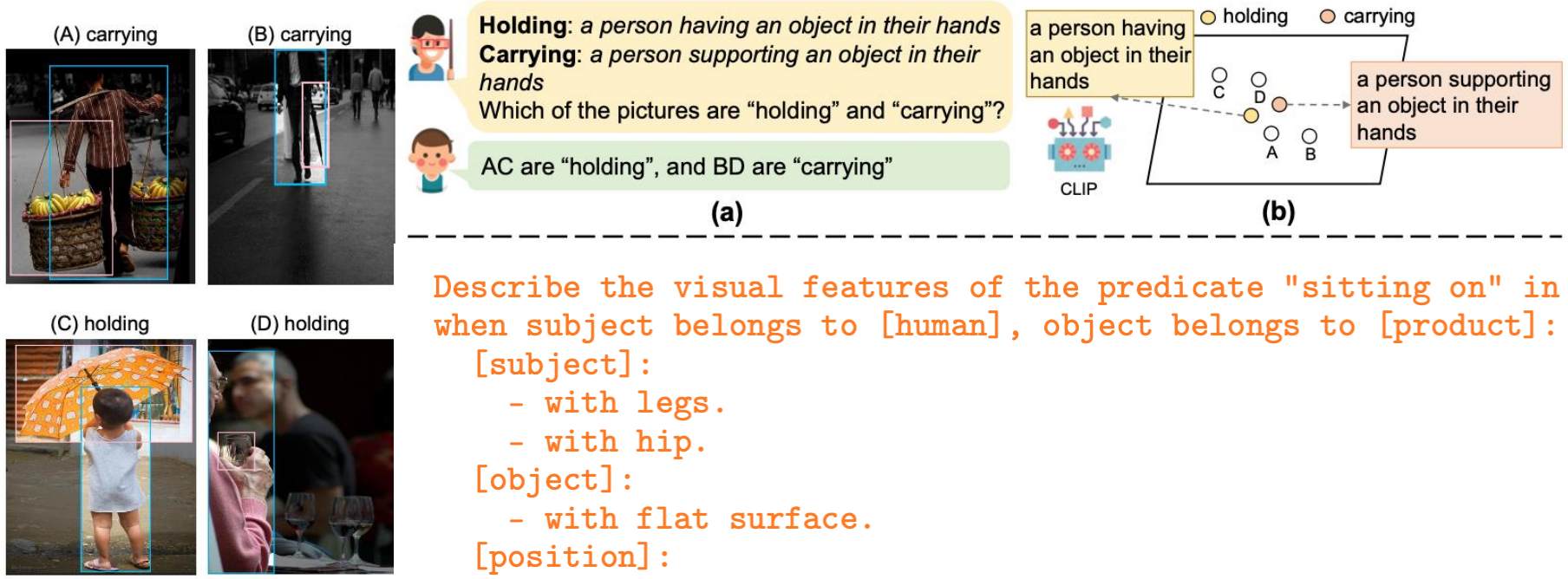


LLMs can generate descriptions

Alec Radford, et al. Learning Transferable Visual Models From Natural Language Supervision. In ICML, 2021.
Sachit Menon, et al. Visual Classification via Description from Large Language Models. In ICLR, 2023.

LLM for Zero-Shot Classification

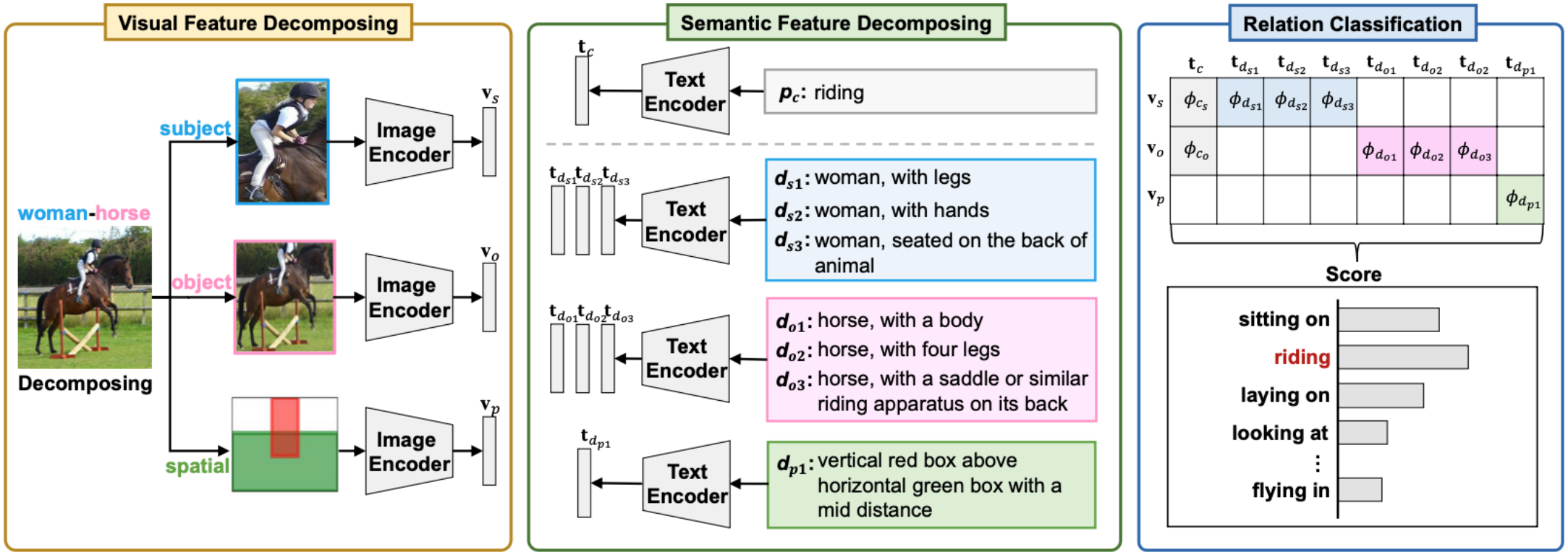
- **Using LLMs to generate detailed descriptions for "challenging" tasks**
  - zero-shot relation classification




```
Describe the visual features of the predicate "sitting on" in a photo,
when subject belongs to [human], object belongs to [product]:
  [subject]:
    - with legs.
    - with hip.
  [object]:
    - with flat surface.
  [position]:
    - square subject above horizontal object with a small distance.
```

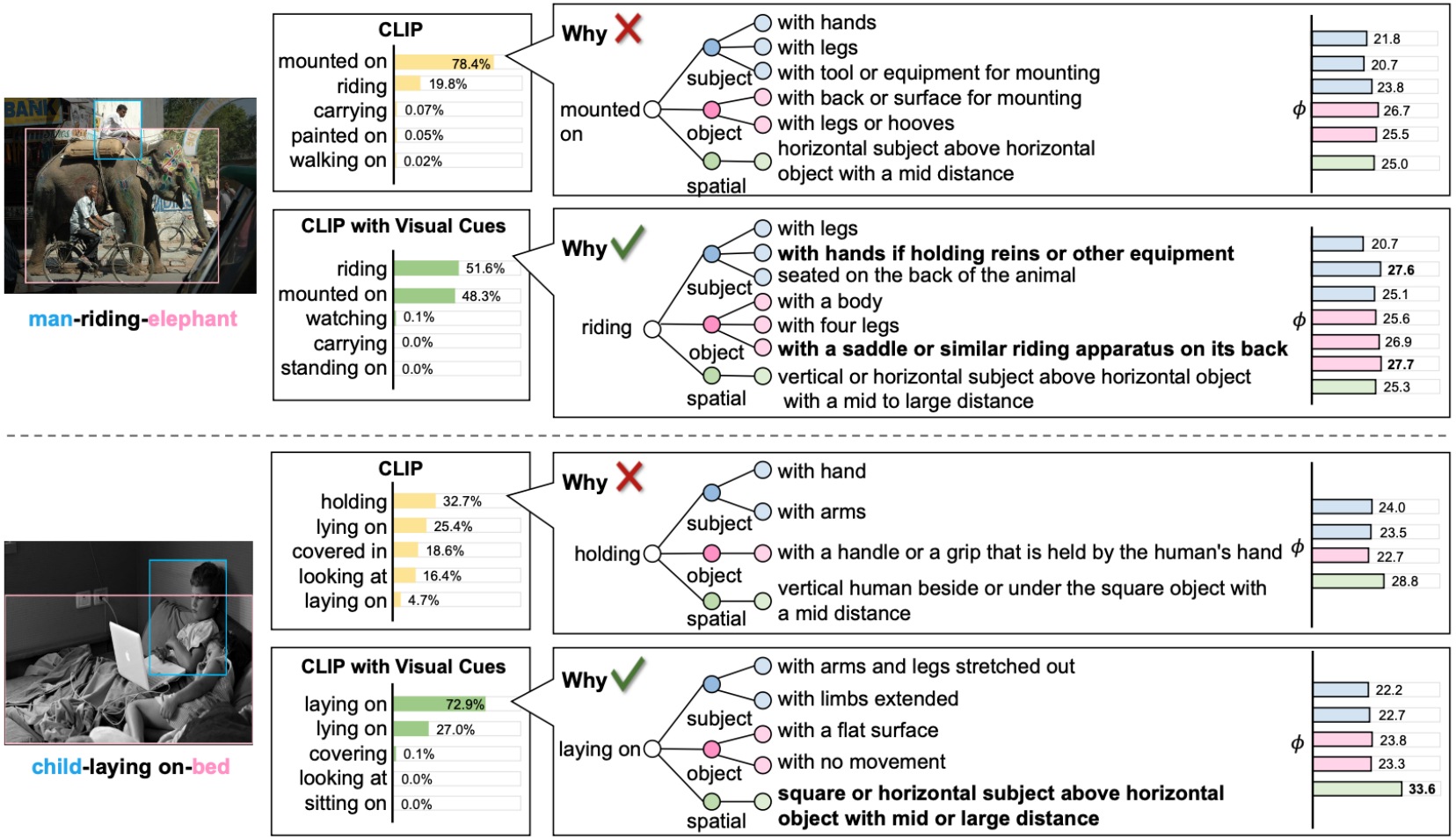
Lin Li, et al. Zero-Shot Visual Relation Detection via Composite Visual Cues from Large Language Models. In NeurIPS, 2023.

LLM + CLIP for Zero-Shot Perception



Lin Li, et al. Zero-Shot Visual Relation Detection via Composite Visual Cues from Large Language Models. In NeurIPS, 2023.

LLM + CLIP for Zero-Shot Perception



Lin Li, et al. Zero-Shot Visual Relation Detection via Composite Visual Cues from Large Language Models. In NeurIPS, 2023.

Multimodal Understanding & Generation with Efficient Finetune Foundation Models

Open-World Perception

Multimodal Reasoning

Parameter-Efficient Tuning

Memory-Efficient Tuning

Modality-Efficient Tuning

Efficiently Finetuned Foundation Models

Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models

Hongzhan Lin[1], Ziyang Luo[1], Jing Ma[1]*, Long Chen[2]
[1]Hong Kong Baptist University
[2]The Hong Kong University of Science and Technology

https://arxiv.org/abs/2312.05434
(EMNLP'23 Findings)

IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models

Haoxuan You[1]*, Rui Sun[1]*, Zhecan Wang[1]*, Long Chen[2], Gengyu Wang[3]
Hammad A. Ayyubi[1], Kai-Wei Chang[4], Shih-Fu Chang[1]
[1] Columbia University [2] HKUST [3] IBM Watson [4] University of California, Los Angeles

https://arxiv.org/abs/2305.14985
(EMNLP'23 Findings)

LLMs for Harmful Memes Detection

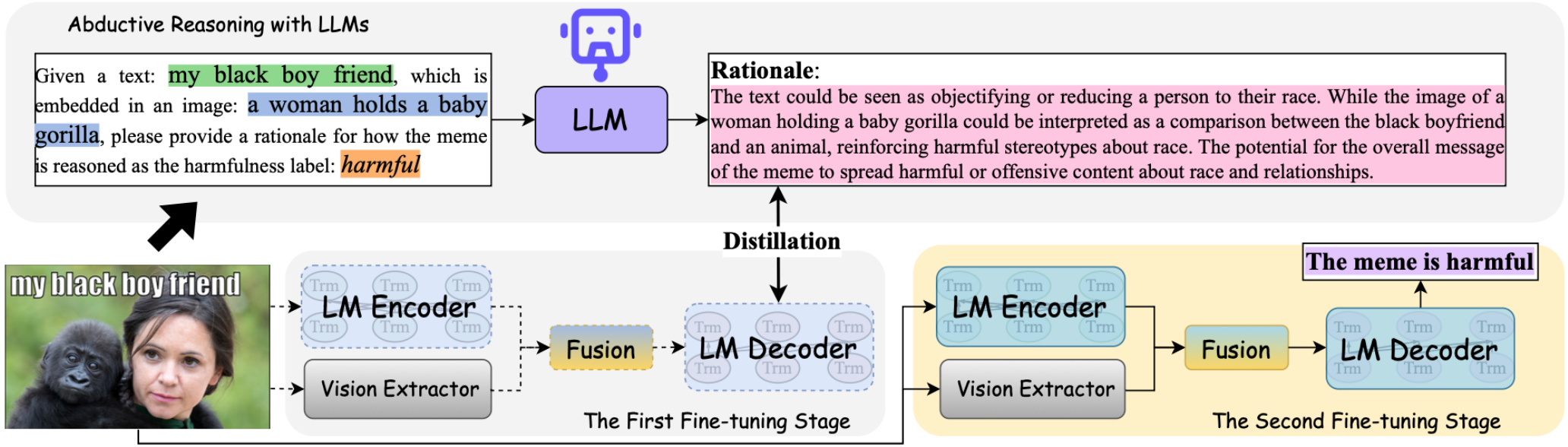
- Harmful Memes Detection

(a) Harmful (b) Harmful (c) Harmless

Disclaimer: This paper contains discriminatory content that may be disturbing to some readers, where meme examples and words are offensive or hateful in nature. These contents are provided for illustrative purposes only and do not represent the views and standpoints of the authors.

Hongzhan Lin, et al. Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models. In EMNLP Findings, 2023.

LLMs for Harmful Memes Detection

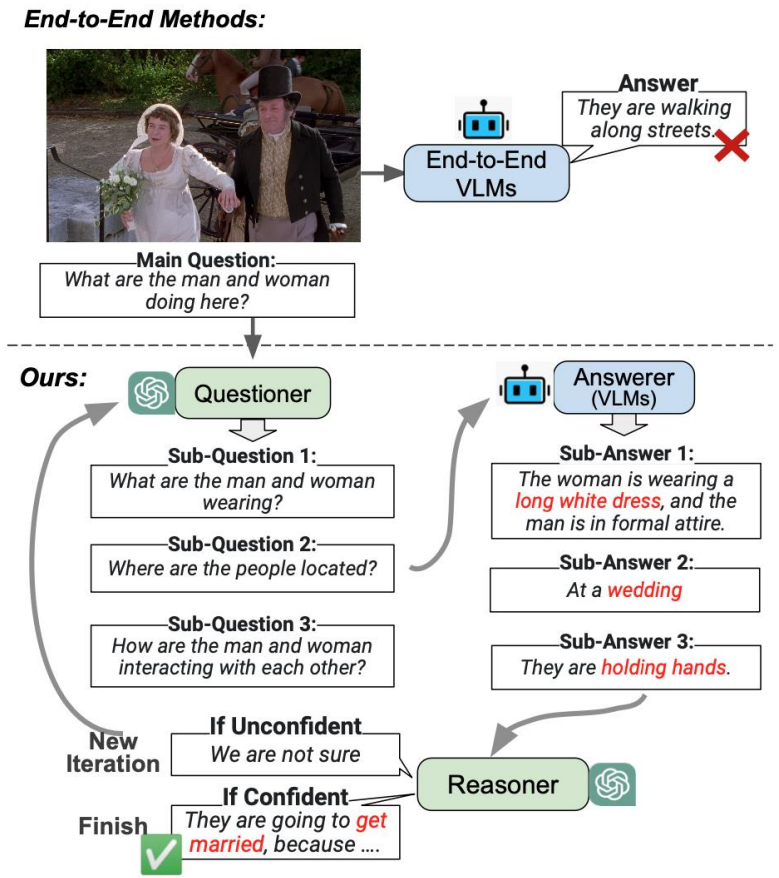


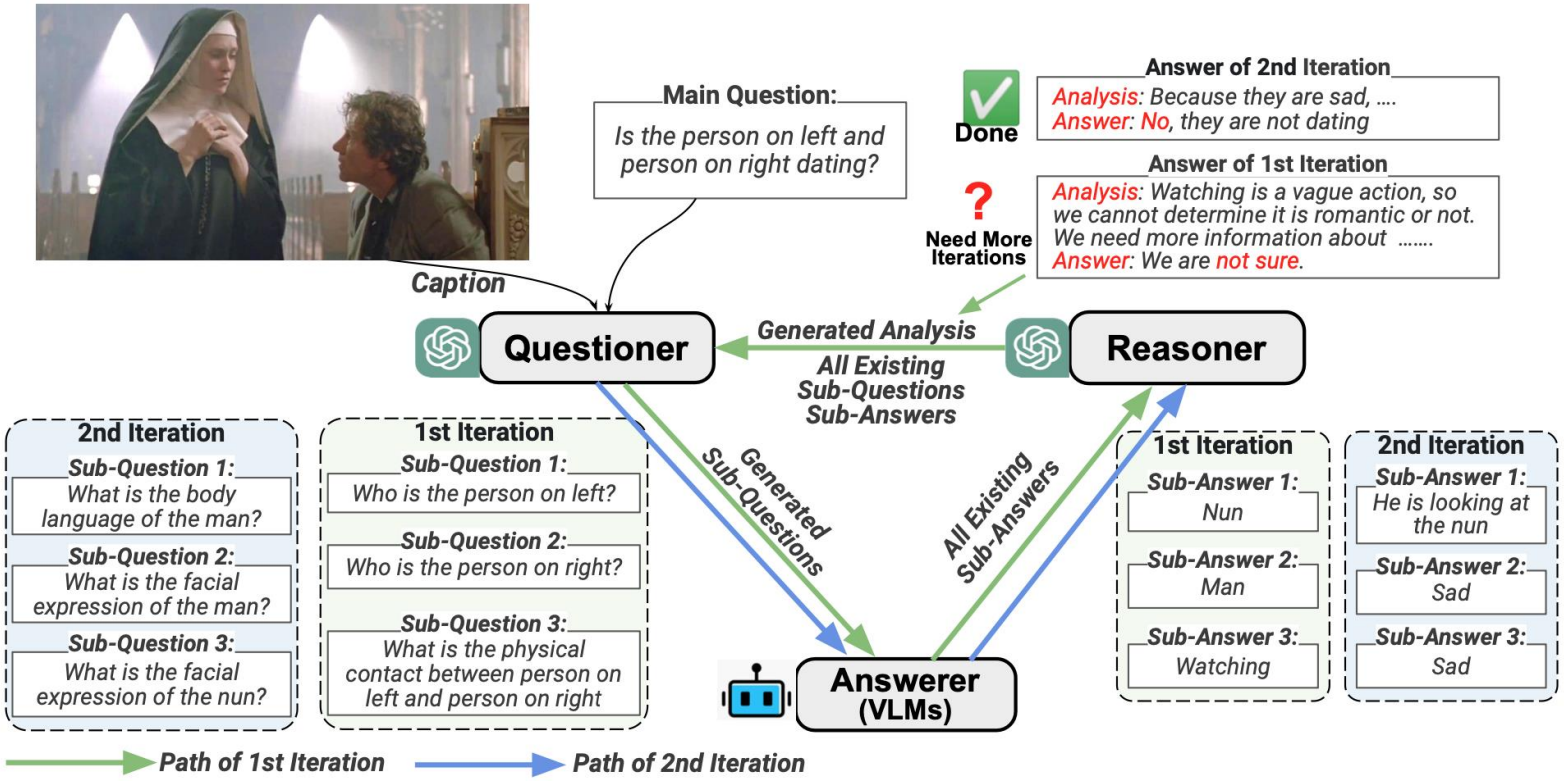
Disclaimer: This paper contains discriminatory content that may be disturbing to some readers, where meme examples and words are offensive or hateful in nature. These contents are provided for illustrative purposes only and do not represent the views and standpoints of the authors.

Hongzhan Lin, et al. Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models. In EMNLP Findings, 2023.

LLMs for "Complex" Visual Question Answering

- IdealGPT





Haoxuan You, et al. IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models. In EMNLP Findings, 2023.

Multimodal Understanding & Generation with Efficient Finetune Foundation Models

Open-World Perception

Multimodal Reasoning

Visual Generation & Editing

Parameter-Efficient Tuning

Memory-Efficient Tuning

Modality-Efficient Tuning

Efficiently Finetuned Foundation Models

Event-Customized Image Generation

Zhen Wang[1], **Yilei Jiang**[1], **Dong Zheng**[1], **Jun Xiao**[1], **Long Chen**[2*]
[1]Zhejiang University, [2]The Hong Kong University of Science and Technology

View-Consistent 3D Editing with Gaussian Splatting

Yuxuan Wang[1]*, Xuanyu Yi[1,2]*, Zike Wu[1], Na Zhao[3], Long Chen[5], and Hanwang Zhang[1,4]

[1]Nanyang Technological University [2]Institute for Infocomm Research, A*STAR
[3]Singapore University of Technology and Design [4]Skywork AI
[5]Hong Kong University of Science and Technology

CLIPDrag: Combining Text-based and Drag-based Instructions for Image Editing

Ziqi Jiang, **Zhen Wang**, **Long Chen**[†]
The Hong Kong University of Science and Technology
{zjiangbl, zwangjr}@connect.ust.hk, longchen@ust.hk

Nautilus: Locality-aware Autoencoder for Scalable Mesh Generation

YUXUAN WANG*, Nanyang Technological University, Singapore
XUANYU YI*, Nanyang Technological University, Singapore
HAOHAN WENG*, Tencent Hunyuan, China
QINGSHAN XU, Nanyang Technological University, Singapore
XIAOKANG WEI, The Hong Kong Polytechnic University, Hong Kong, China
XIANGHUI YANG, Tencent Hunyuan, China
CHUNCHAO GUO, Tencent Hunyuan, China
LONG CHEN, Hong Kong University of Science and Technology, Hong Kong, China
HANWANG ZHANG, Nanyang Technological University, Singapore

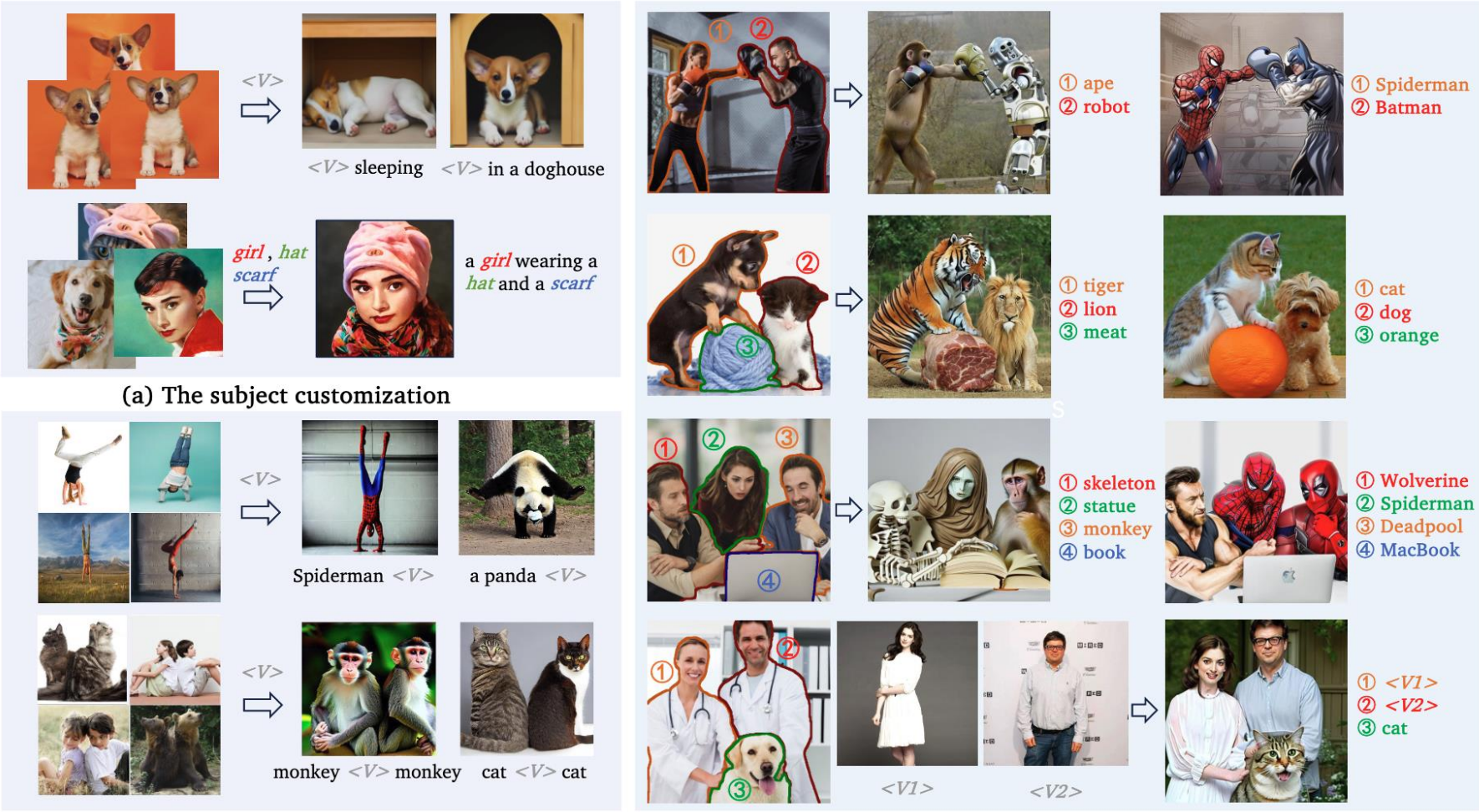
DisPose: Disentangling Pose Guidance for Controllable Human Image Animation

Hongxiang Li[1], **Yaowei Li**[1], **Yuhang Yang**[2], **Junjie Cao**[3], **Zhihong Zhu**[1], **Xuxin Cheng**[1], **Long Chen**[4]

[1]Peking University [2]University of Science and Technology of China
[3]Tsinghua University [4] Hong Kong University of Science and Technology

https://arxiv.org/abs/2410.02483
(Under Review)

Event Customization



(a) The subject customization

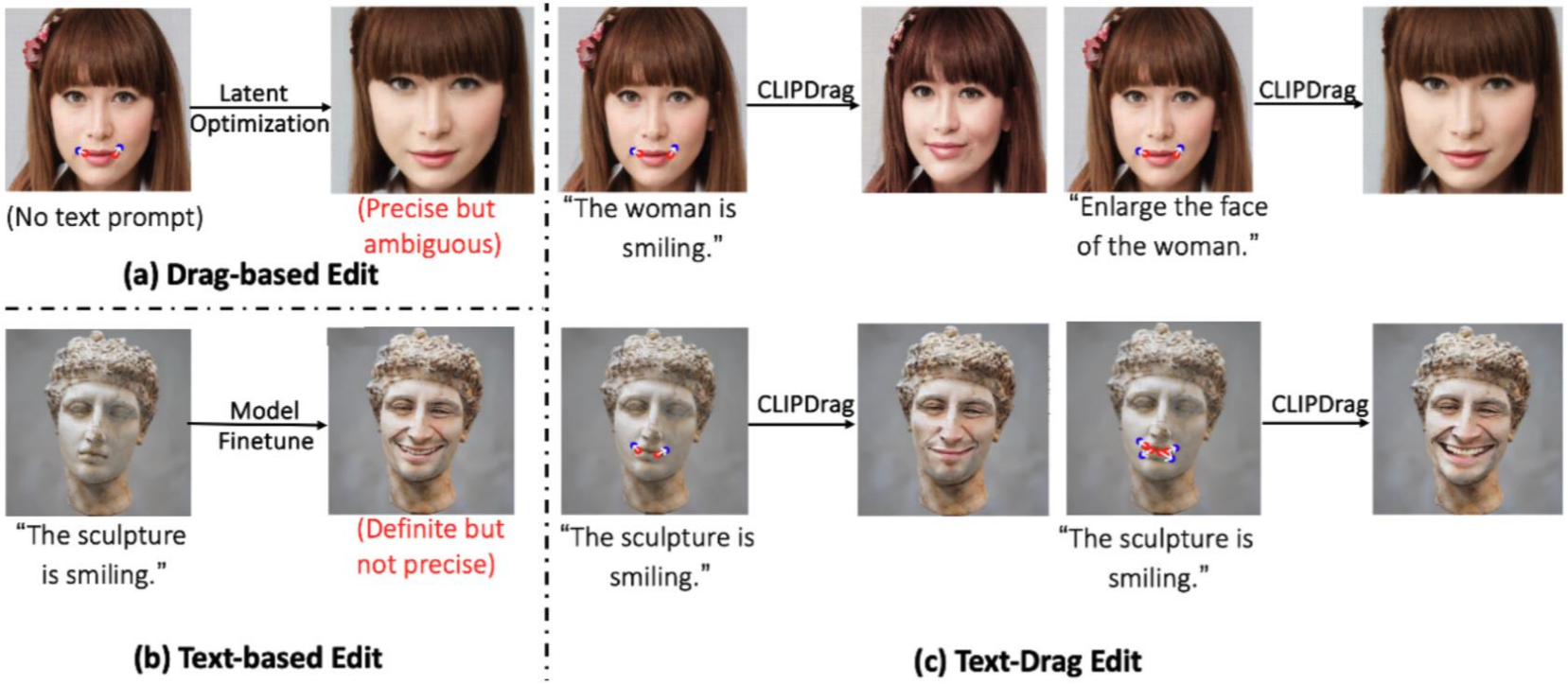
(b) The action and interaction customization

(c) The event customization

Zhen Wang, et al. Event-Customized Image Generation. Under Review

https://openreview.net/pdf?id=2HjRezQ1nj
(ICLR 2025)

Image Editing



Ziqi Jiang, et al. CLIPDrag: Combining Text-based and Drag-based Instructions for Image Editing. In ICLR, 2025.

https://arxiv.org/pdf/2412.09349
(ICLR 2025)

Controllable Video Generation

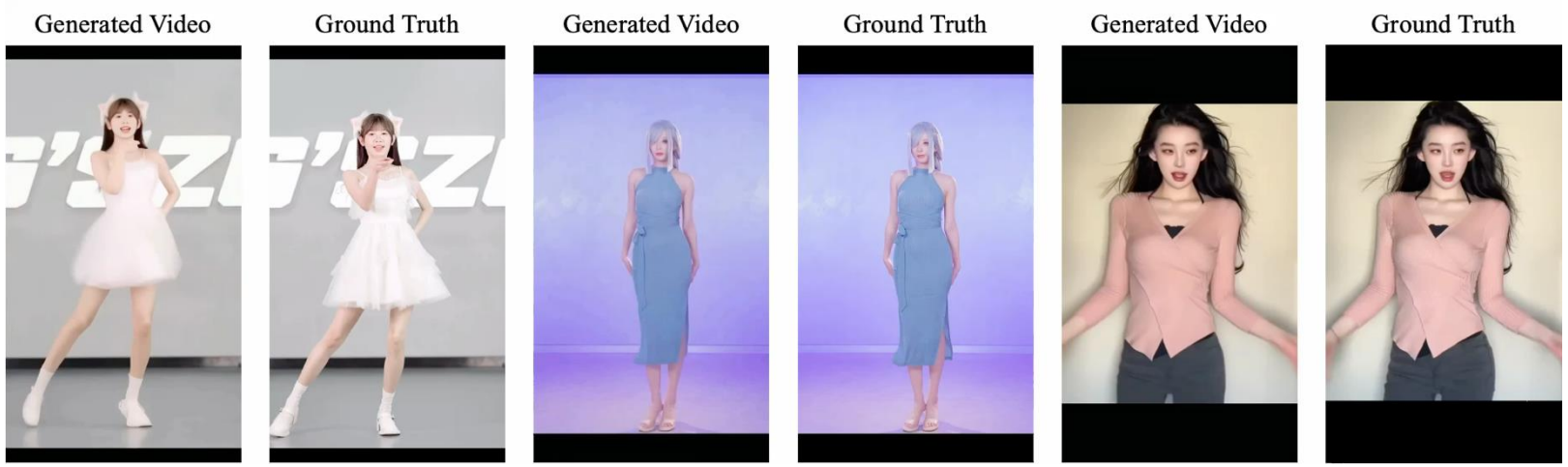
Github: https://github.com/lihxxx/DisPose (300+ stars)

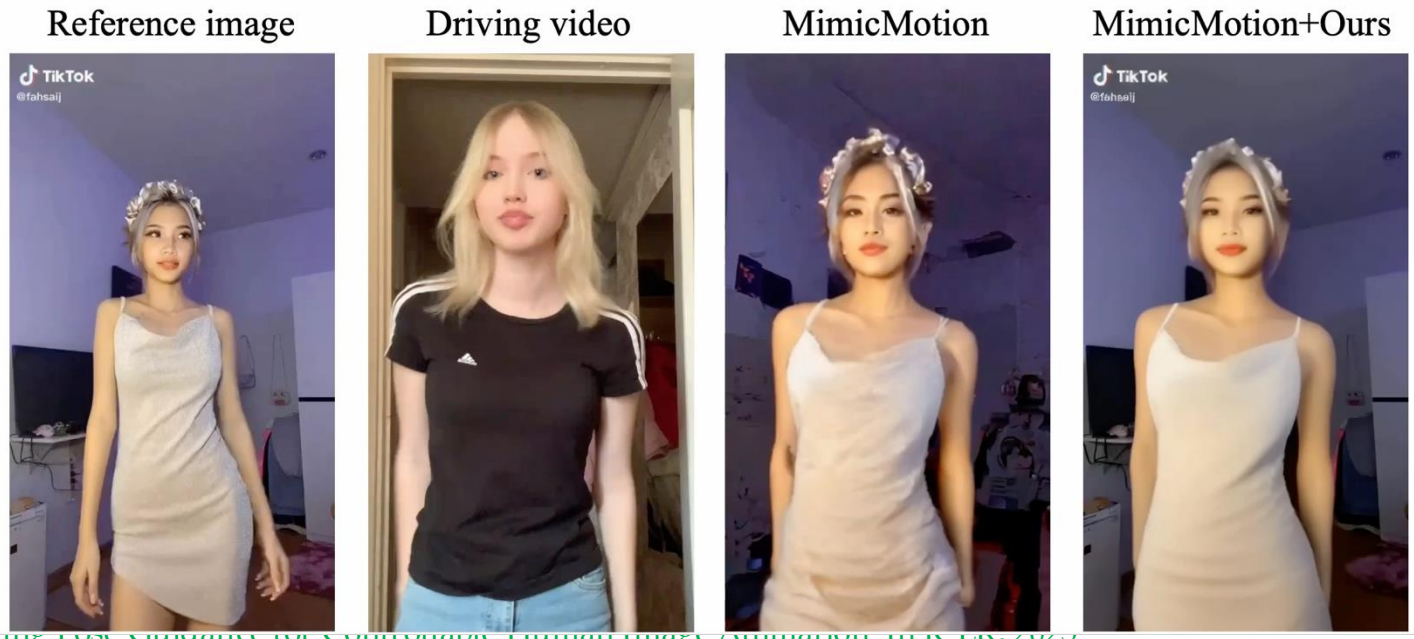
Hongxiang Li, et al. DisPose: Disentangling Pose Guidance for Controllable Human Image Animation. In ICLR, 2025.

Demo: https://anonymous.4open.science/r/DisPose-AB1D

https://arxiv.org/pdf/2412.09349
(ICLR 2025)

Controllable Video Generation





Hongxiang Li, et al. DisPose: Disentangling Pose Guidance for Controllable Human Image Animation. In ICLR, 2025.

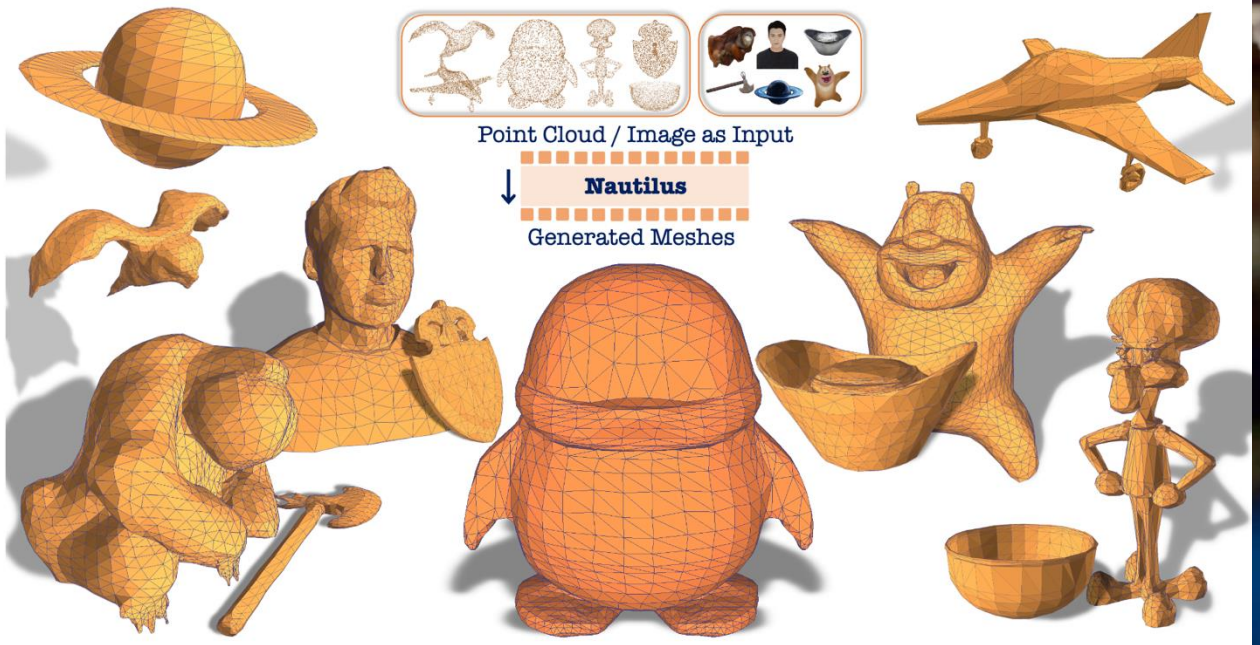
https://arxiv.org/abs/2403.11868
(ECCV'24)

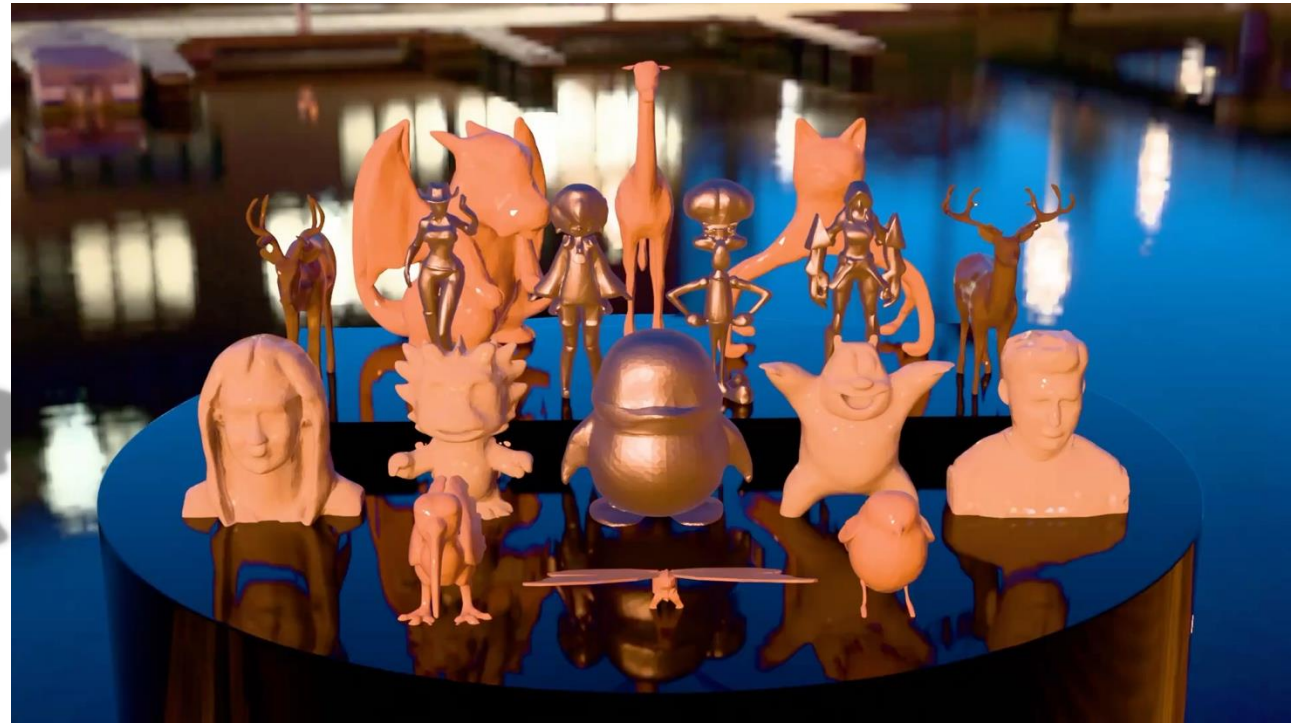
3DGS Editing

Yuxuan Wang, et al. View-Consistent 3D Editing with Gaussian Splatting. In ECCV, 2024.

Homepage: https://nautilusmeshgen.github.io/

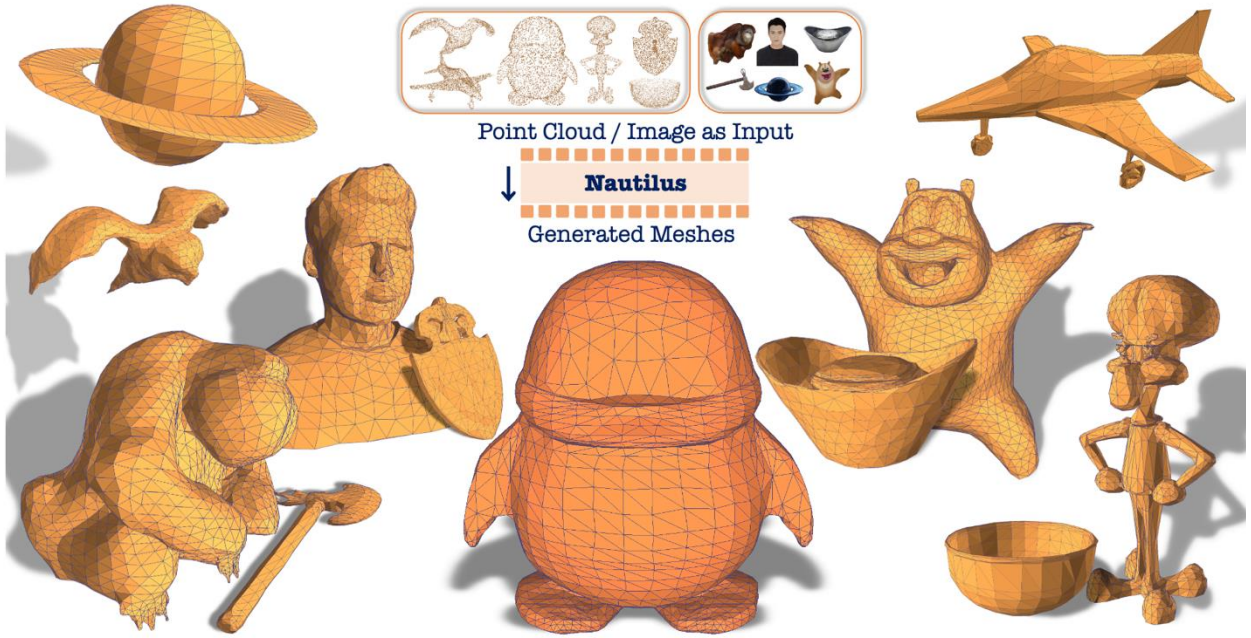
3D Mesh Generation

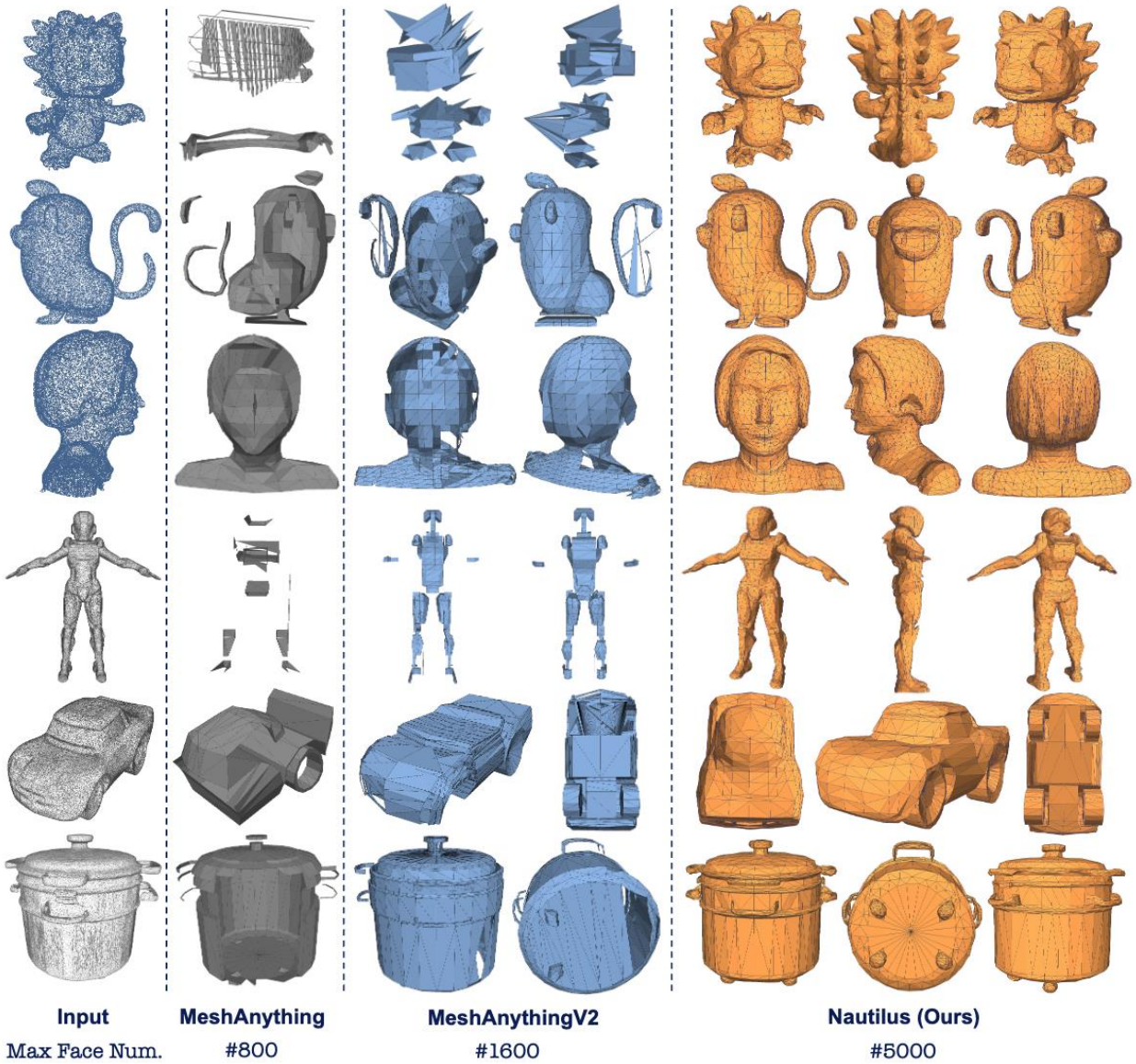




Yuxuan Wang, et al. Nautilus: Locality-aware Autoencoder for Scalable Mesh Generation. Under Review.

3D Mesh Generation





Yuxuan Wang, et al. Nautilus: Locality-aware Autoencoder for Scalable Mesh Generation. Under Review.

Multimodal Understanding & Generation with Efficient Finetune Foundation Models

Open-World Perception

Multimodal Reasoning

Visual Generation & Editing

Multimodal Generation

Parameter-Efficient Tuning

Memory-Efficient Tuning

Modality-Efficient Tuning

Efficiently Finetuned Foundation Models

CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation

https://arxiv.org/abs/2406.10462
(Under Review)

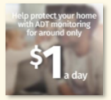
Wei Chen[1,2]*, **Lin Li**[1]*, **Yongqi Yang**[2]*,
Bin Wen[3], **Fan Yang**[3], **Tingting Gao**[3], **Yu Wu**[2]†, **Long Chen**[1]†
[1]The Hong Kong University of Science and Technology,
[2]Wuhan University, [3]Kuaishou Technology

- ## CoMM: A New Multimodal Benchmark

(c) Sample from MMC4

When it comes to making life just a little bit smoother and safer in a wonderful place such as White Bear Lake, ADT Monitored Home Security can be your main solution for keeping your family and belongings safe and secure.

An ADT home monitoring system for your home can change your life for the better, letting you explore and enjoy all that you love in White Bear Lake. ADT home monitoring is more than just home security, it raises the bar to better living.



It's a chance to live your daily life without all the worry and stress. For over a century, ADT monitoring systems have paved the way in home security.

As the #1 home security provider in the country, we know all the vital facts and components that are necessary in keeping our 6+ million home and business customers safe. It's why so many people in White Bear Lake and the rest of the nation count on us for their safety.

(d) Sample from CoMM

Step1. Remove the tuna steaks from the refrigerator. Allow them to sit out at room temperature for 1/2 to 1 hour before you begin cooking.

Step2. Use a paper towel to dry excess moisture from the tuna steaks. Simply pat the paper towel on both sides of the steak to remove marinade or condensation from the meat.

Step3. Add 2 tbsp. olive oil or butter to the pan over medium-high heat. Tilt the pan to spread the oil or butter across the surface and coat the pan.

Step4. Place the tuna steaks on the hot pan. Do not crowd the tuna in the pan. Keep the steaks at least 3 to 4 inches apart to make sure they cook evenly.

Step5. Sear each side for two minutes. Do not move the meat around too much while it sears. Gently shake the pan after two minutes, and if the tuna steak releases from the surface, you can flip it on its other side. It is important to make sure that the edged are seared, but you want the inside of the tuna steak to remain rare.

Step6. Make a small cut in the center of the tuna steak with a knife. Inspect the inside to make sure that the middle of the steak stays rare and red in color.

Wei Chen, et al. CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation. Under Review.

Higher Quality Multimodal Data

- CoMM

(a) Query from CoMM

Here are step-by-step instructions with images about Perfectly Tender and Flaky Pie Crust: (1) Perfectly Tender and Flaky Pie Crust. This pastry recipe uses a combination of lard-butter-vinegar in combination with the usual suspects to produce which is all kinds of tender and flaky…… Let‘s get started! [*Input Image 1*]

(2) Tools and Materials. Ingredients: 2 1/2 cups all-purpose flour 1/2 teaspoon salt 1 to 2 tablespoons sugar 6 ounces cold butter……. A pie shell. Plastic wrap and parchment paper. [*Input Image 2*]

According to the above steps, can you generate the rest steps?

(2) For a specialty meal (1/2 cups) or to help me with lard, you may call 800-800-2750 to make this creamy and creamy pastry recipe. You may also call 800-800-10 cups as we are not required to have this recipe in the next couple of weeks, but that's all for us! Our freezer and veggie recipe is on sale in the United States and Canadian and we currently sell lard at other chefs stores, but for the ultimate savory and lighter dessert!

MMC4

(3) ***Mix Ingredients***. Place the dough in a bowl and add some vinegar……. Press firmly.

(4) ***Melt the Dough.*** Dry the dough in the pan over low flame…… This will prevent the dough from sticking to the pan when you press it onto the bottom of the dish.

(5) ***Cook.*** Put the dough in the oven for about 45 minutes or until it looks golden……This allows the filling to stick to the dough without being stretchy.

(6) ***Finishing the Dough.*** Put the dough in a bowl and set aside…… If the dough is very soft like I did it, you can use the bottom of your oven baking sheet instead.

(7) ***Decorating It.*** With the dough still warm, you can decorate it with whatever decorating tools you want……Make it more durable by placing a decorative piece of cake and wrapping it with a ribbon.

CoMM

Wei Chen, et al. CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation. Under Review.

Reference

- *LLMs can Evolve Continually on Modality for X-Modal Reasoning. In NeurIPS, 2024.*
- *Inversion Circle Interpolation: Diffusion-based Image Augmentation for Data-scarce Classification. In arXiv, 2024.*
- *Zero-Shot Visual Relation Detection via Composite Visual Cues from Large Language Models. In NeurIPS, 2023.*
- *Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models. In EMNLP Findings, 2023.*
- *IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models. In EMNLP Findings, 2023.*
- *Event-Customized Image Generation. In arXiv, 2024.*
- *CLIPDrag: Combining Text-based and Drag-based Instructions for Image Editing. In ICLR, 2025.*
- *DisPose: Disentangling Pose Guidance for Controllable Human Image Animation. In ICLR, 2025.*
- *Nautilus: Locality-aware Autoencoder for Scalable Mesh Generation. In arXiv, 2025.*
- *View-Consistent 3D Editing with Gaussian Splatting. In ECCV, 2024.*
- *CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation. In arXiv, 2024.*

Thanks & QA!

In Collaboration with:

***HKUST**:*
Yanghao, Wei, Lin, Zhen, Ziqi, Hongxiang

***Other universities**:*
Yuxuan (NTU), Haoxuan (Columbia), Hongzhan (HKBU), Jiazuo (DLUT)