MLNLP 2024

Multimodal Large Language Model Session

Towards AGI: from Unified MLLM to Multimodal Generalist

探索从统一的多模态大模型Generalist到AGI之路

Hao Fei (费豪)

Postdoc at National University of Singapore

Associate Researcher at Skywork AI, Singapore

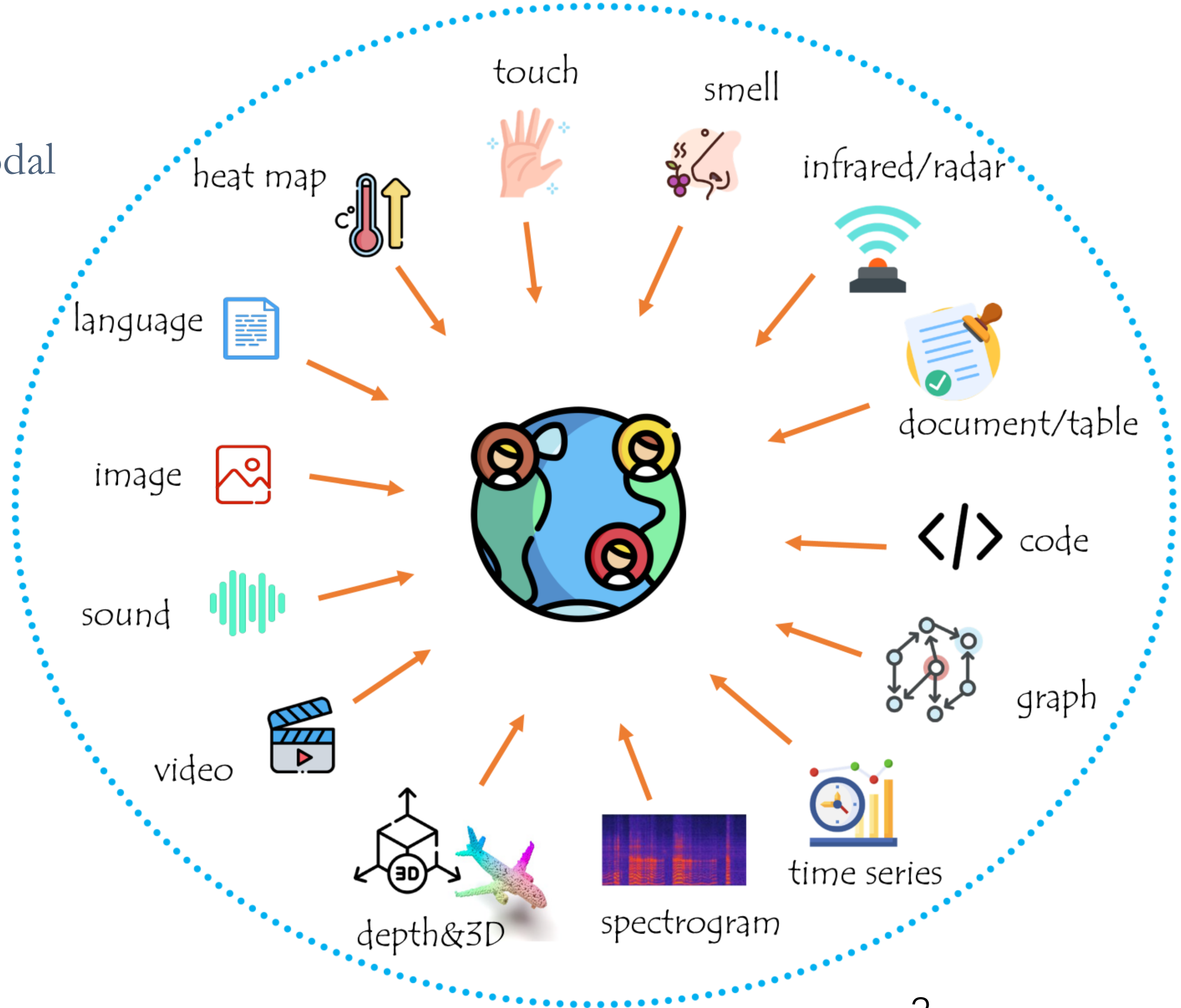
http://haofei.vip/

Sep 1st, 2024

Content

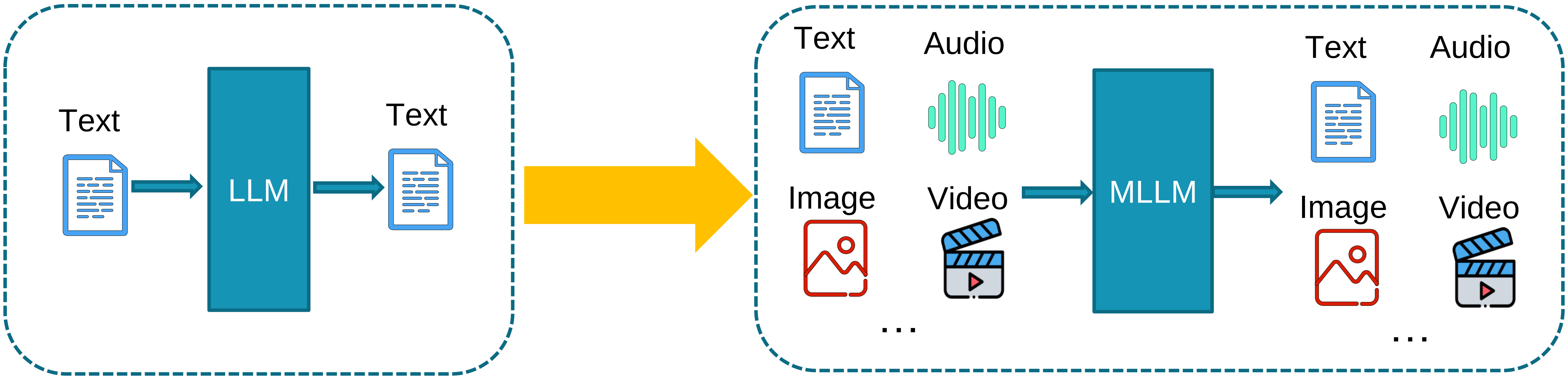
Preliminary on MLLM

☞ This world we live in is replete with multimodal information & signals, not just language.



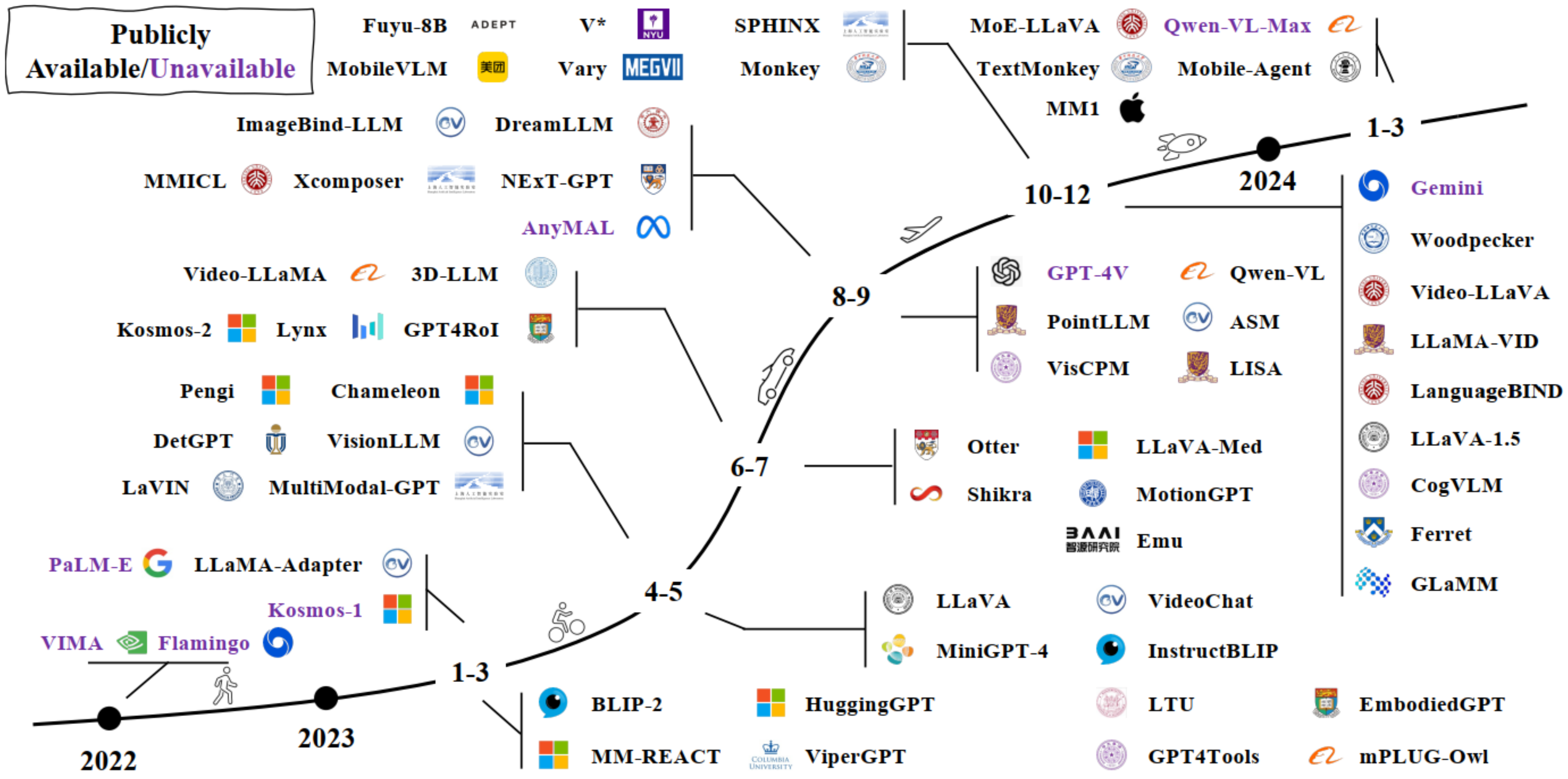
Preliminary on MLLM

- Extending **Language** LLM to **Multimodal** LLM (MLLM)



Preliminary on MLLM

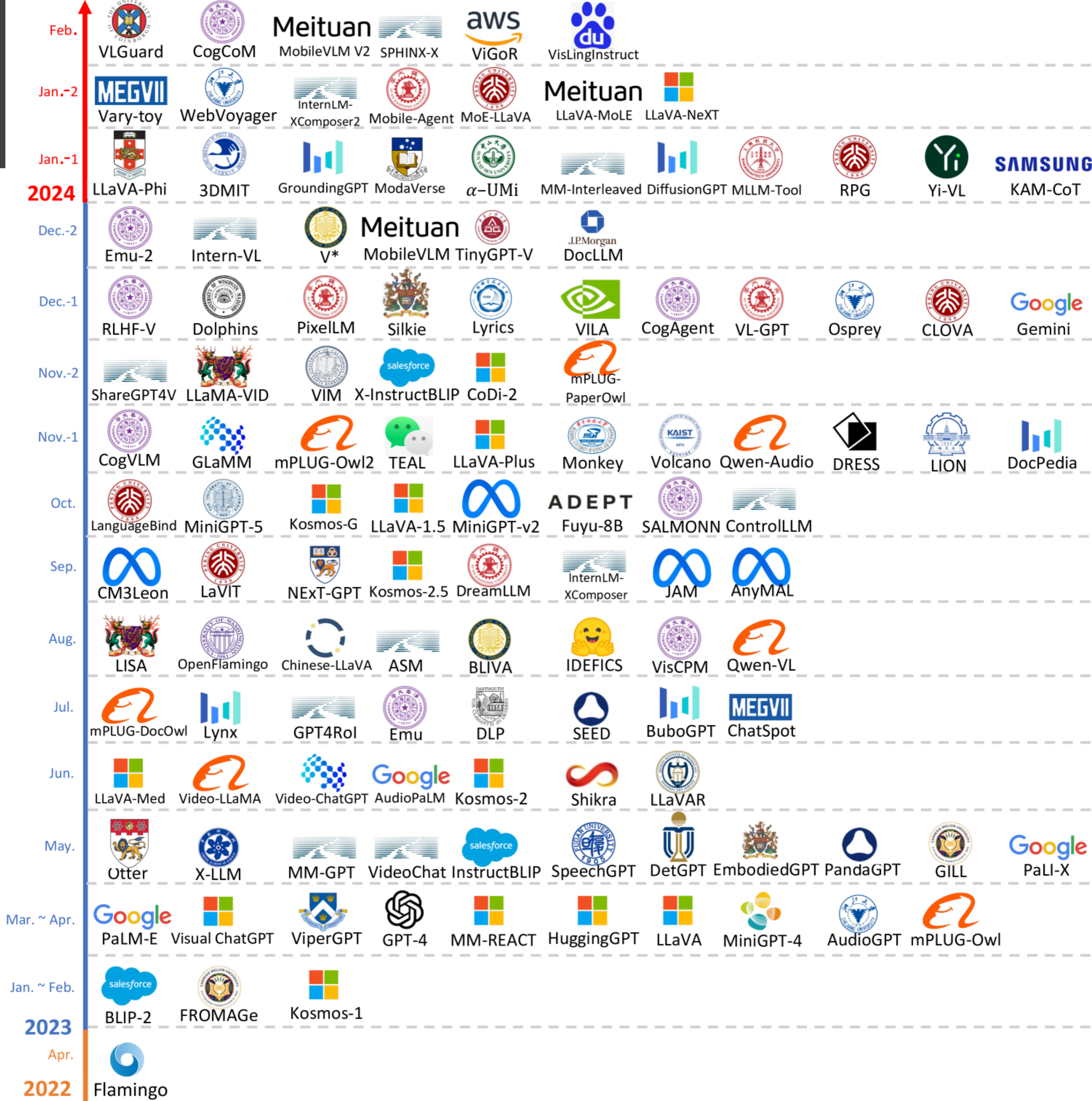
Research Trends on MLLM



[1] A Survey on Multimodal Large Language Models. https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models, 2023.

Preliminary on MLLM

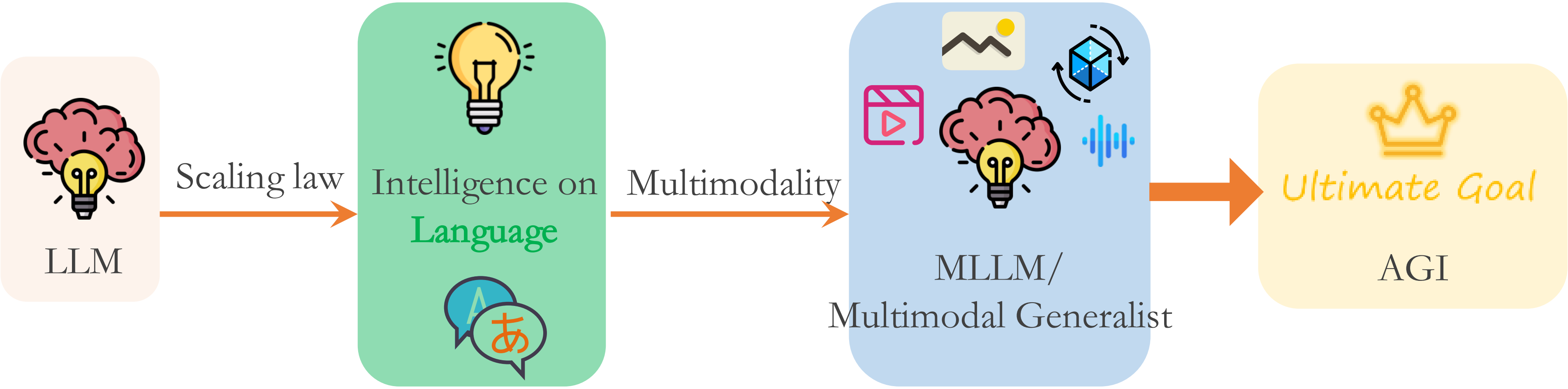
Research Trends on MLLM



[1] MM-LLMs: Recent Advances in MultiModal Large Language Models, 2023.

Preliminary on MLLM

- Existing MLLMs (almost) all stand on the **Language** Intelligence



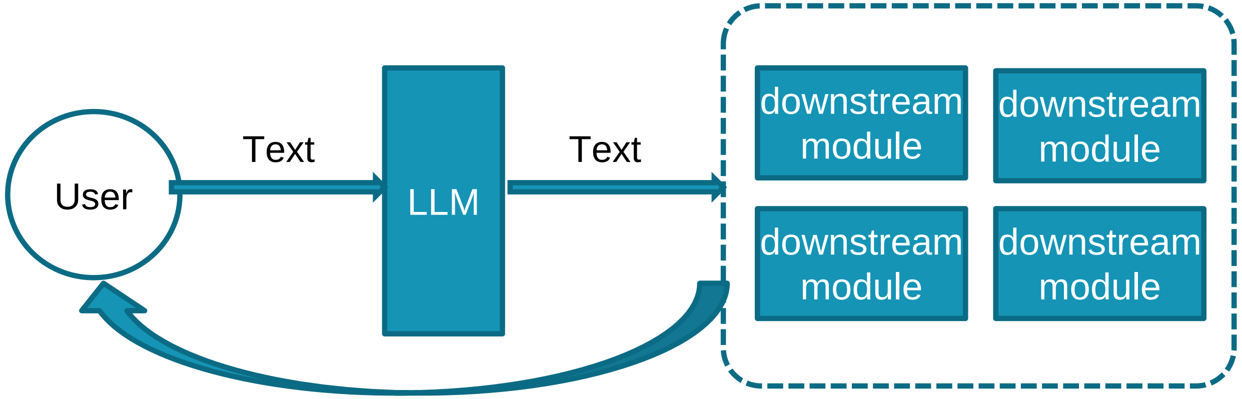
Preliminary on MLLM

Architecture-I: LLM as Discrete Task Scheduler/Controller (Agent)

☞ *The role of the LLM is to receive textual signals and instruct textual commands to call downstream modules.*

- Key feature:

  All message passing within the system, such as "multimodal encoder to the LLM" or "LLM to downstream modules", is facilitated through pure textual commands as the medium.



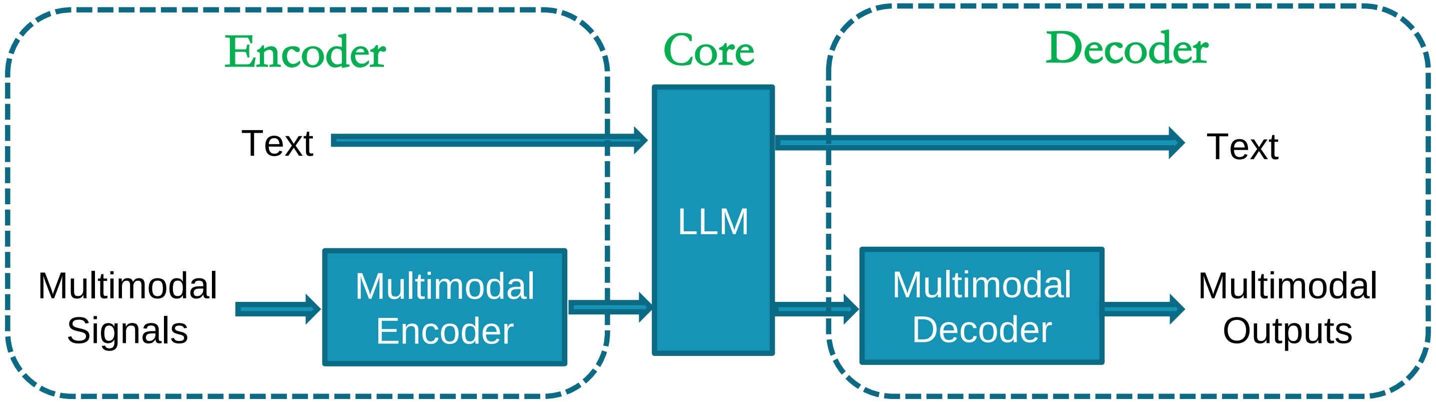
Preliminary on MLLM

Architecture-II: LLM as Joint Part of System

☞ *The role of the LLM is to perceive multimodal information, and **react by itself**, in an structure of* **Encoder-LLM-Decoder**.

- Key feature:

  LLM is the key joint part of the system, receiving multimodal information directly from outside, and delegating instruction to decoders/generators in a more smooth manner.



Taxonomy of existing MLLMs

<table>
<tr><th></th><th colspan="4">Modality (w/ Language)</th></tr>
<tr><th></th><th>Image</th><th>Video</th><th>Audio</th><th>3D</th></tr>
<tr><td rowspan="4">Input-side Perceiving</td><td>Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, …</td><td>VideoChat, Video-ChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID, Momentor, …</td><td>AudioGPT, SpeechGPT, VIOLA, AudioPaLM, SALMONN, MU-LLaMA, …</td><td>3D-LLM, 3D-GPT, LL3DA, SpatialVLM, PointLLM, Point-Bind, …</td></tr>
<tr><td>[Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT, Osprey, PixelLM, …</td><td>[Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, …</td><td>–</td><td>–</td></tr>
<tr><td colspan="2">Video-LLaVA, Chat-UniVi, LLaMA-VID</td><td>–</td><td>–</td></tr>
<tr><td colspan="3">Panda-GPT, Video-LLaMA, AnyMAL, Macaw-LLM, Gemini, VideoPoet, ImageBind-LLM, LLMBind, LLaMA-Adapter, …</td><td>–</td></tr>
<tr><td rowspan="3">Perceiving + Generating</td><td>GILL, EMU, MiniGPT-5, DreamLLM, LLaVA-Plus, InternLM-XComposer2, SEED-LLaMA, LaVIT, Mini-Gemini, …</td><td>GPT4Video, Video-LaVIT, VideoPoet, …</td><td>AudioGPT, SpeechGPT, VIOLA, AudioPaLM, …</td><td>–</td></tr>
<tr><td colspan="2">[Pixel-wise] Vitron</td><td>–</td><td>–</td></tr>
<tr><td colspan="3">NExT-GPT, Unified-IO 2, AnyGPT, CoDi-2, Modaverse, ViT-Lens, …</td><td>–</td></tr>
</table>

Preliminary on MLLM

MLLM Tutorial Series

Homepage:

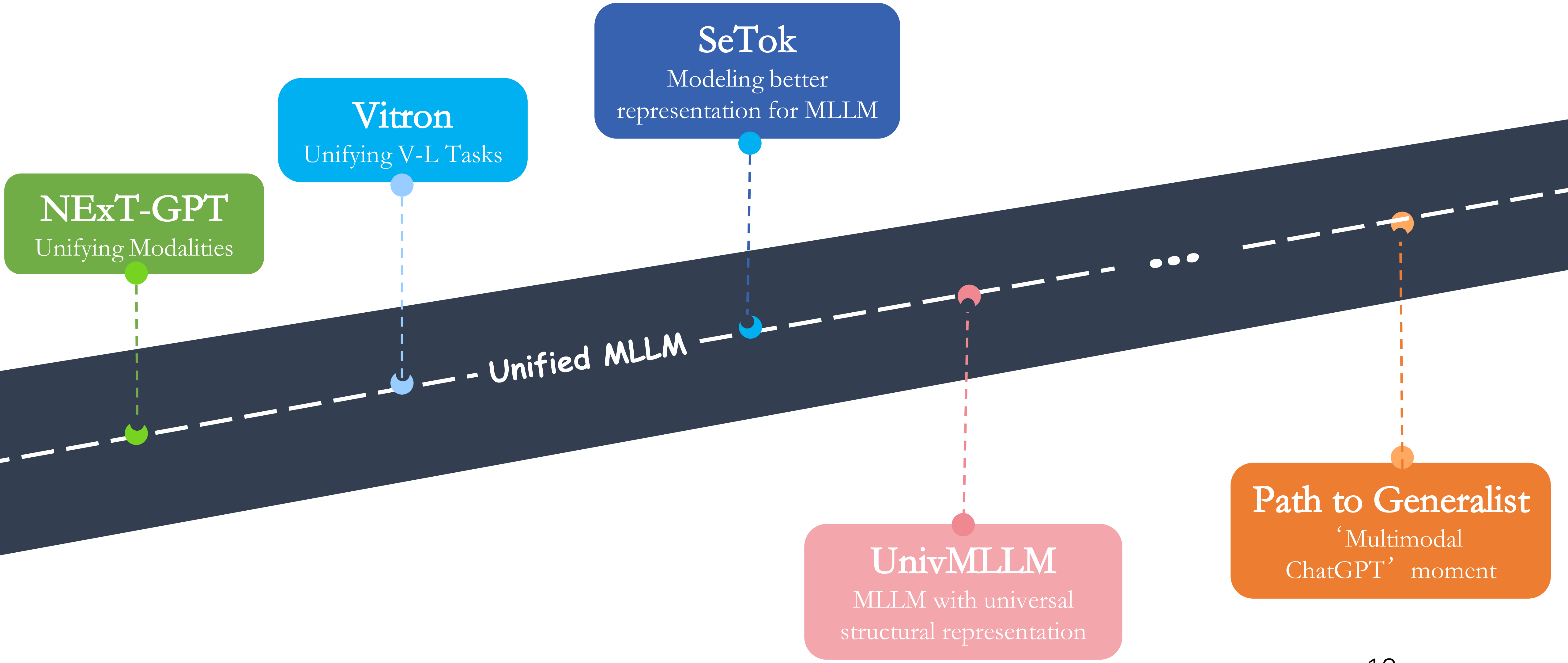
- COLING: https://mllm2024.github.io/COLING2024/
- CVPR: https://mllm2024.github.io/CVPR2024/
- ACM MM: https://mllm2024.github.io/ACM-MM2024/
- …

Oct 31, 2024

Video: https://www.youtube.com/watch?v=pHBT3zXxQX8

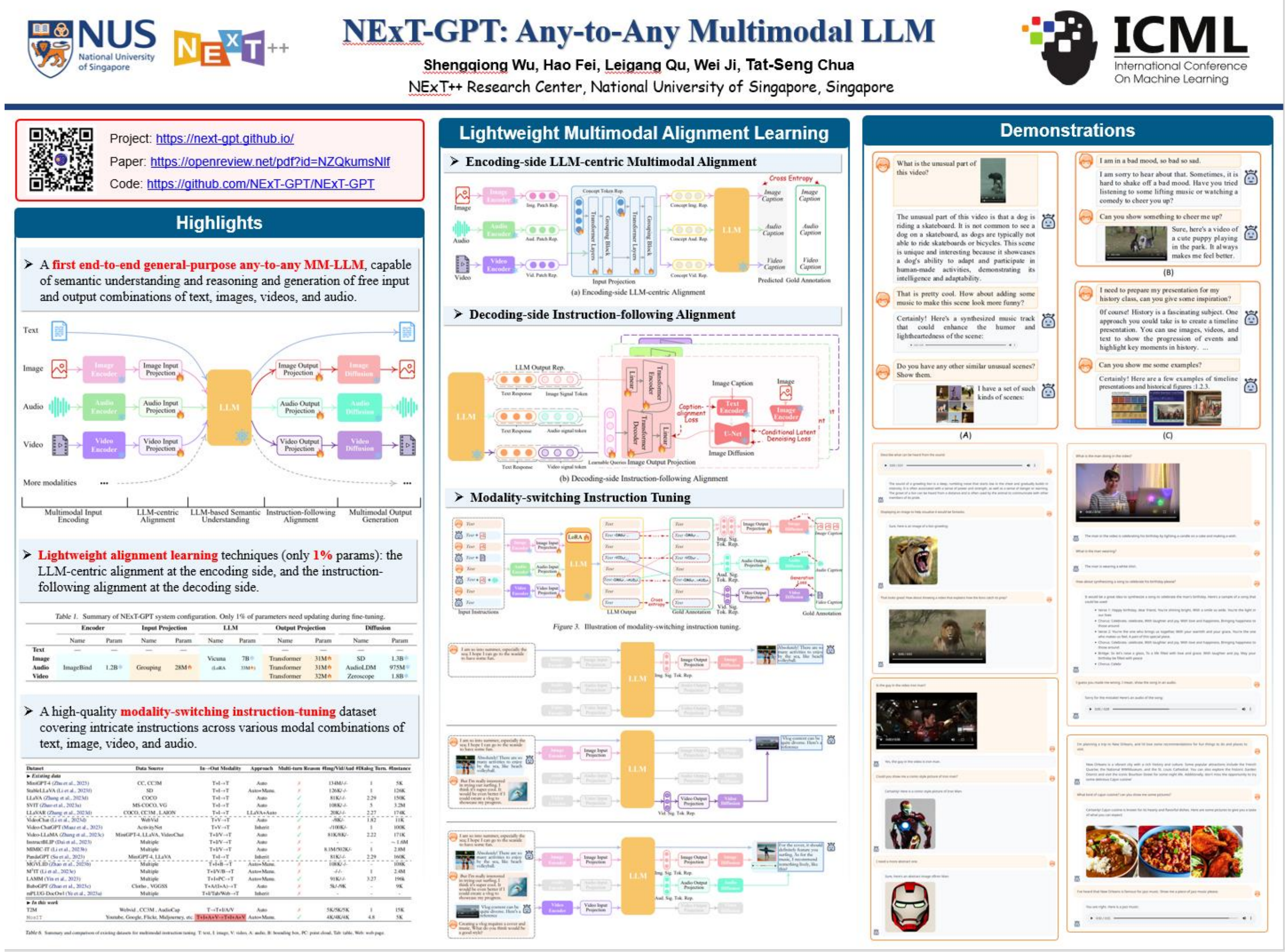
Content

Milestone on Unified MLLM



Unified MLLM

- NExT-GPT: Any-to-Any MLLM



Project: https://next-gpt.github.io

Paper: https://arxiv.org/pdf/2309.05519

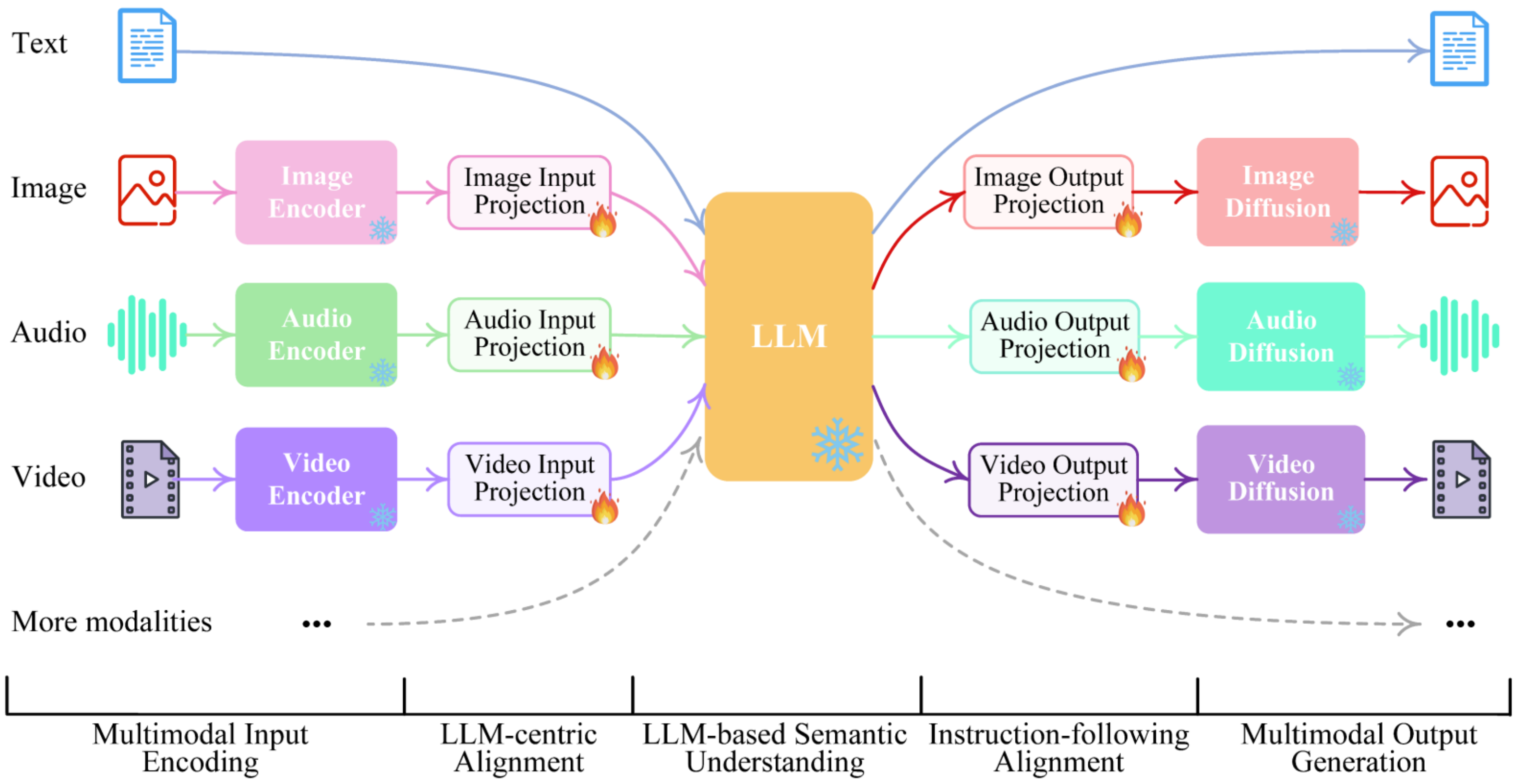
Code: https://github.com/NExT-GPT/NExT-GPT

- Shengqiong Wu, **Hao Fei**, Leigang Qu, Wei Ji, Tat-Seng Chua. "NExT-GPT: Any-to-Any Multimodal LLM". **ICML**. 2024.

Unified MLLM

NExT-GPT: Any-to-Any MLLM

☞ *The first end-to-end MLLM that **perceives** input and **generates** output in arbitrary combinations (any-to-any) of text, image, video, and audio and beyond.*



- Shengqiong Wu, **Hao Fei**, Leigang Qu, Wei Ji, Tat-Seng Chua. "NExT-GPT: Any-to-Any Multimodal LLM". **ICML**. 2024.

Uni

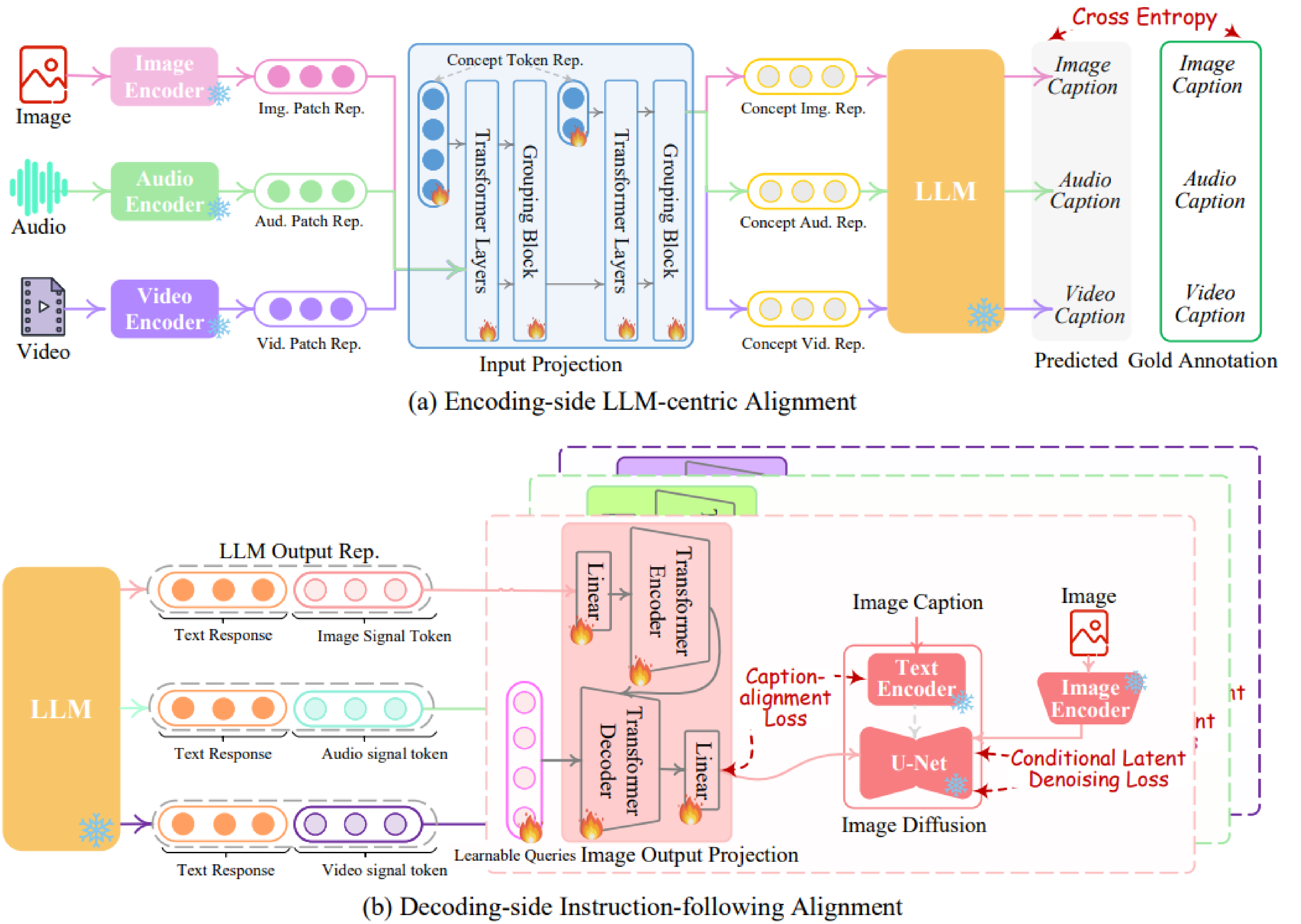


Figure 2: Illustration of the lightweight multimodal alignment learning of encoding and decoding, respectively.

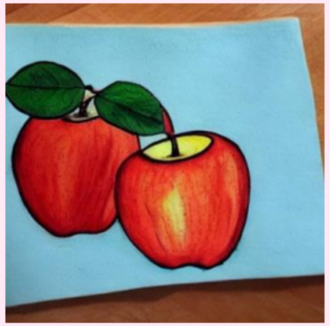
Unified MLLM/Generalist

NExT-GPT: Any-to-Any MLLM

- ***Instruction***: *There were originally 7 apples on the table, but I ate one. Then, how many apples are left now? Please generate a picture to describe the result.*

Response: Sure, I'd be happy to help you understand the process and results.
Based on your statement, "There were originally 7 apples on the table, but I ate one", I inferred that there are now 6 apples left on the table.
To visualize this, I used a text-to-image model to generate an image based on your description. The model I used is called 'runwayml/stable-diffusion-v1-5', which is currently the only available model for this task. The image generated can be found at the following path: '/images/ec4d.jpg'.
Please note that the image is a representation of the text and may not be 100% accurate. If there's anything else you need, feel free to ask!

HuggingGPT



{*Intermedia Caption:* 6 apples left on the table}

Unified MLLM/Generalist

NExT-GPT: Any-to-Any MLLM

- ***Instruction****: There were originally 7 apples on the table, but I ate one. Then, how many apples are left now? Please generate a picture to describe the result.*

Response: Certainly! In the end, you have the following outcomes: 6 apples on the table.

NExT-GPT

Uni

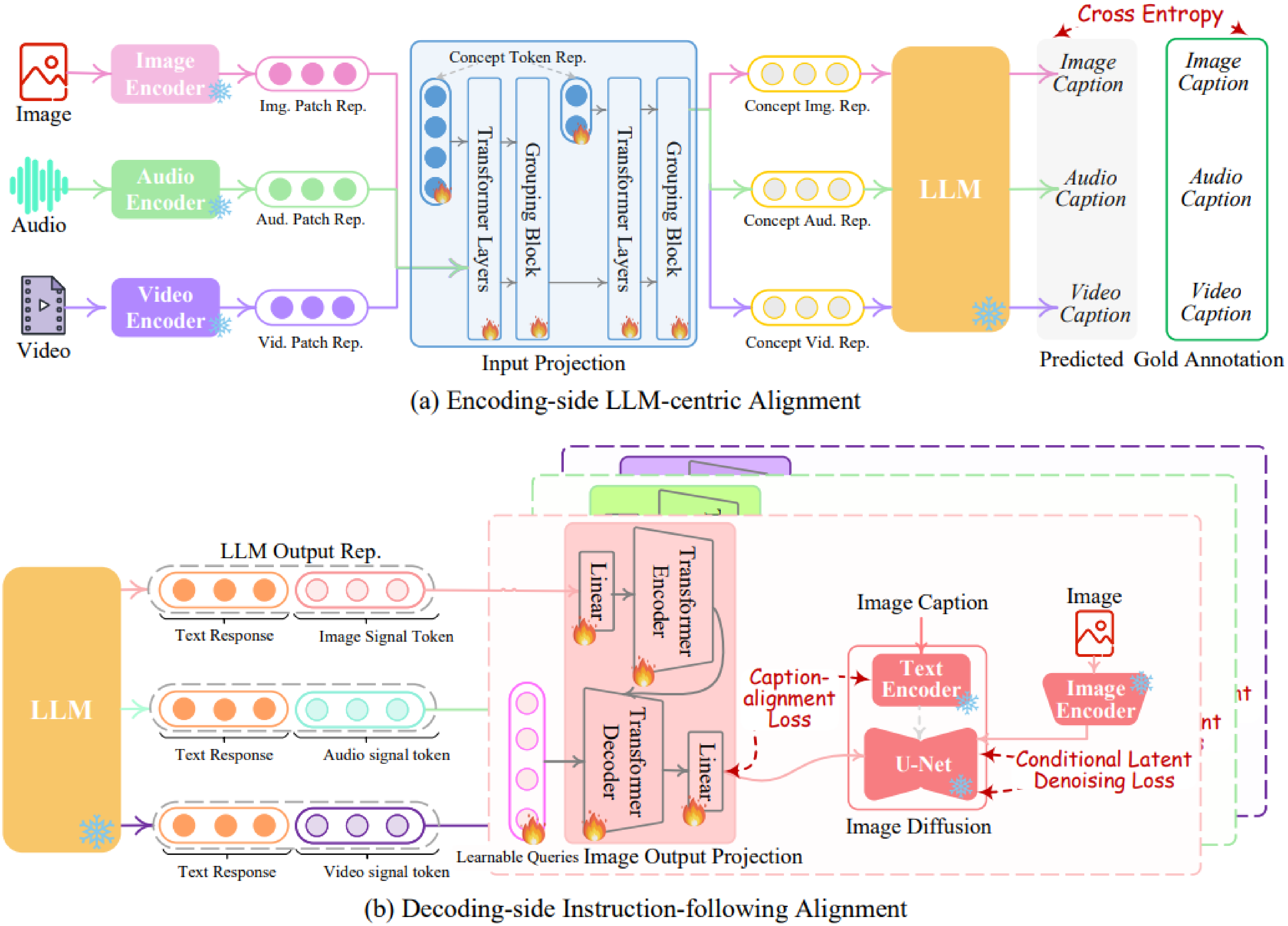


Figure 2: Illustration of the lightweight multimodal alignment learning of encoding and decoding, respectively.

Unified MLLM

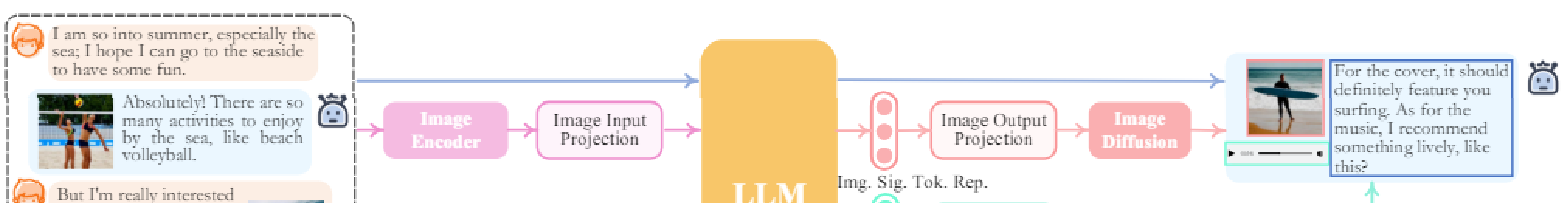
NExT-GPT: Any-to-Any MLLM

Lightweight fine-tuning alignment learning: only 1% parameter update is needed.

Table 1: Summary of NExT-GPT system configuration. Only 1% of parameters need updating during fine-tuning.

	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—			—	—	—	—
Image					Vicuna	7B❄	Transformer	31M🔥	SD	1.3B❄
Audio	ImageBind	1.2B❄	Grouping	28M🔥	(LoRA	33M🔥)	Transformer	31M🔥	AudioLDM	975M❄
Video							Transformer	32M🔥	Zeroscope	1.8B❄

Unified MLLM



➢ Mod

Instr

Mec

Dataset	Data Source	In→Out Modality	Approach	Multi-turn Reason	#Img/Vid/Aud	#Dialog Turn.	#Instance
▸ ***Existing data***							
MiniGPT-4 [70]	CC [7], CC3M [45]	T+I→T	Auto	✗	134M/-/-	1	5K
StableLLaVA [32]	SD [43]	T+I→T	Auto+Manu.	✗	126K/-/-	1	126K
LLaVA [65]	COCO [33]	T+I→T	Auto	✓	81K/-/-	2.29	150K
SVIT [67]	MS-COCO [33], VG [26]	T+I→T	Auto	✓	108K/-/-	5	3.2M
LLaVAR [65]	COCO [33], CC3M [45], LAION [44]	T+I→T	LLaVA+Auto	✓	20K/-/-	2.27	174K
VideoChat [29]	WebVid [4]	T+V→T	Auto	✓	-/8K/-	1.82	11K
Video-ChatGPT [36]	ActivityNet [17]	T+V→T	Inherit	✗	-/100K/-	1	100K
Video-LLaMA [64]	MiniGPT-4, LLaVA, VideoChat	T+I/V→T	Auto	✓	81K/8K/-	2.22	171K
InstructBLIP [11]	Multiple	T+I/V→T	Auto	✗	-	-	∼ 1.6M
MIMIC-IT [27]	Multiple	T+I/V→T	Auto	✗	8.1M/502K/-	1	2.8M
PandaGPT [49]	MiniGPT-4, LLaVA	T+I→T	Inherit	✓	81K/-/-	2.29	160K
MGVLID [68]	Multiple	T+I+B→T	Auto+Manu.	✗	108K/-/-	-	108K
M^{3}IT [30]	Multiple	T+I/V/B→T	Auto+Manu.	✗	-/-/-	1	2.4M
LAMM [61]	Multiple	T+I+PC→T	Auto+Manu.	✓	91K/-/-	3.27	196k
BuboGPT [69]	Clotho [13], VGGSS [8]	T+A/(I+A)→T	Auto	✗	5k/-/9K	-	9K
mPLUG-DocOwl [60]	Multiple	T+I/Tab/Web→T	Inherit	✗	-	-	-
▸ ***In this work***							
T2M	Webvid [4], CC3M [45], AudioCap [24]	T→T+I/A/V	Auto	✗	4.9K/4.9K/4.9K	1	14.7K
MosIT	Youtube, Google, Flickr, Midjourney, etc.	T+I+A+V→T+I+A+V	Auto+Manu.	✓	4K/4K/4K	4.8	5K

Table 2: Summary and comparison of existing datasets for multimodal instruction tuning. T: text, I: image, V: video, A: audio, B: bounding box, PC: point cloud, Tab: table, Web: web page.

Unified MLLM

- Realizing Human-like Multimodal Interaction Mode

NExT-GPT

Text + Audio

↓

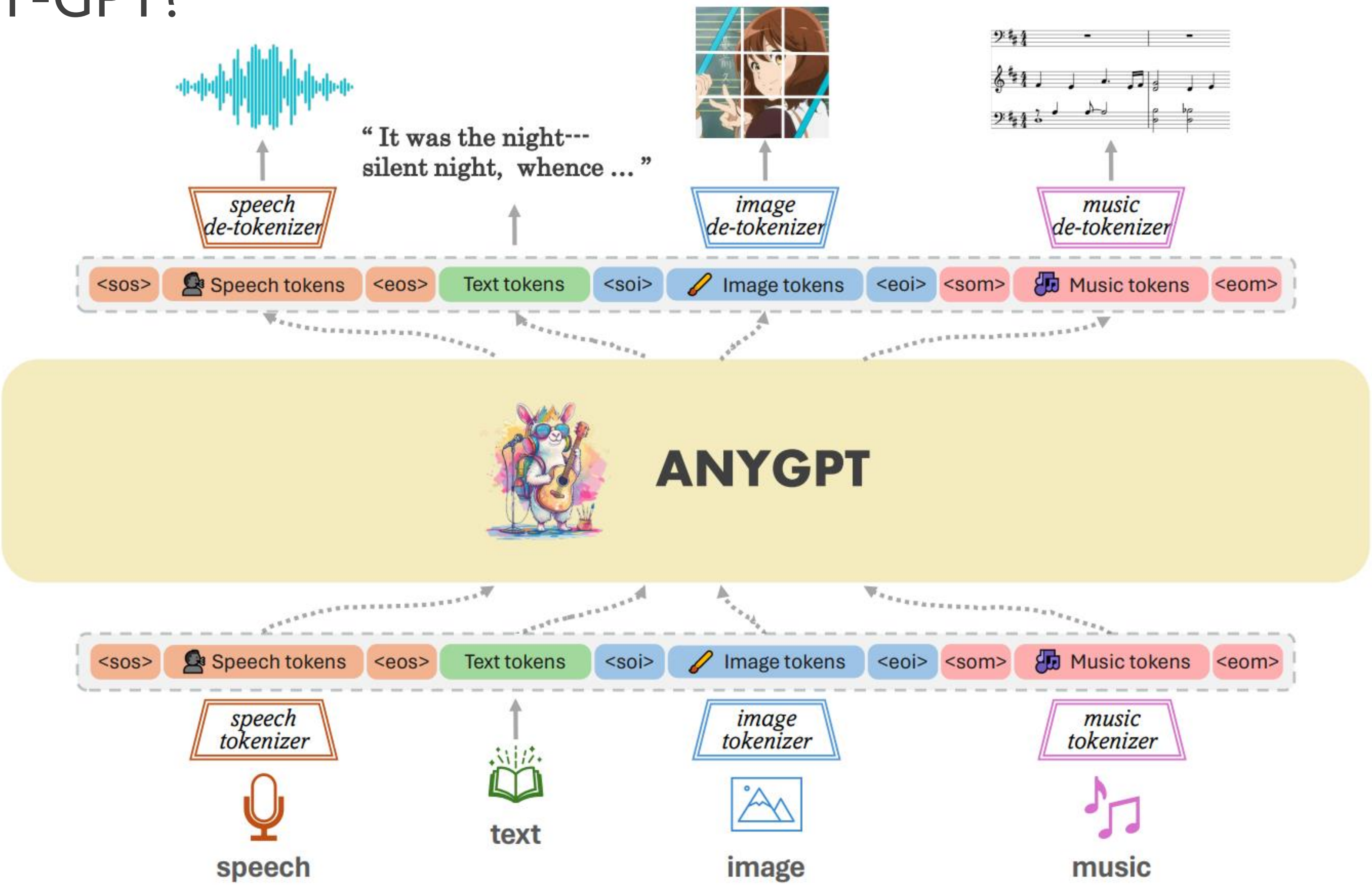
Text + Image + Video

Unified MLLM

What’s Next after NExT-GPT?

➢ AnyGPT

- Discrete Tokenization
- Autoregressive Generation



[1] AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2023

Unified MLLM

VITRON: A Unified Pixel-level Vision MLLM

Project: https://vitron-llm.github.io/

Paper: https://is.gd/aGu0VV

Code: https://github.com/SkyworkAI/Vitron

- **Hao Fei**, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, Shuicheng Yan. "VITRON: A Unified Pixellevel Vision LLM for Understanding, Generating, Segmenting, Editing". 2024

Unified MLLM

VITRON: A Unified Pixel-level Vision MLLM

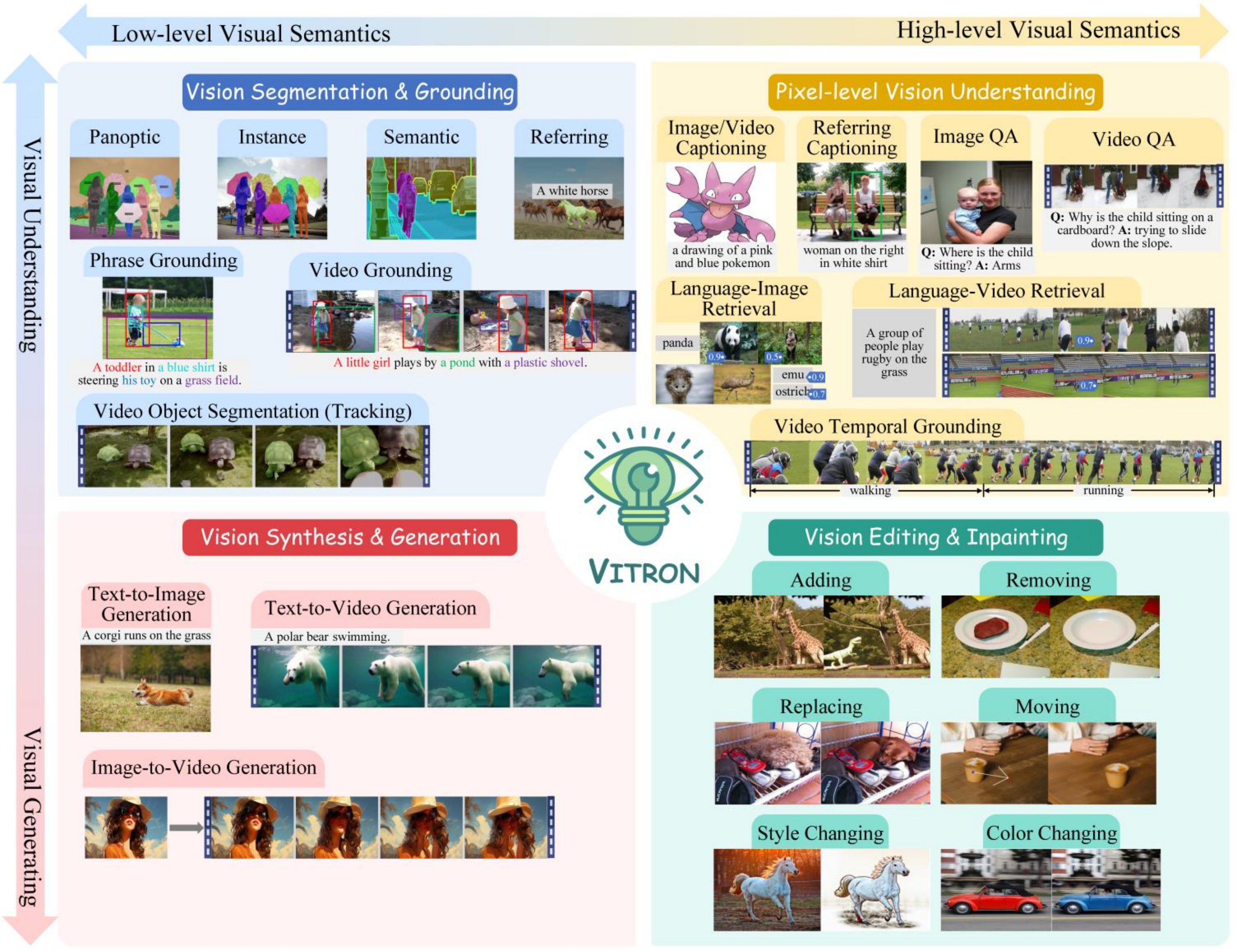
Existing vision MLLM:
not professional enough in visual task unification

Model	Vision Supporting		Pixel/Regional Understanding	Segmenting/ Grounding	Generating	Editing
	Image	Video				
Flamingo [1]	✓	✗	✗	✗	✗	✗
BLIP-2 [45]	✓	✗	✗	✗	✗	✗
MiniGPT-4 [126]	✓	✗	✗	✗	✗	✗
LLaVA [57]	✓	✗	✗	✗	✗	✗
GILL [39]	✓	✗	✗	✗	✓	✗
Emu [90]	✓	✗	✗	✗	✓	✗
MiniGPT-5 [125]	✓	✗	✗	✗	✓	✗
DreamLLM [23]	✓	✗	✗	✗	✓	✗
GPT4RoI [122]	✓	✗	✓	✓	✗	✗
NExT-Chat [118]	✓	✗	✓	✓	✗	✗
MiniGPT-v2 [13]	✓	✗	✓	✓	✗	✗
Shikra [14]	✓	✗	✓	✓	✗	✗
Kosmos-2 [72]	✓	✗	✓	✓	✗	✗
GLaMM [78]	✓	✗	✓	✓	✗	✗
Osprey [117]	✓	✗	✓	✓	✗	✗
PixelLM [79]	✓	✗	✓	✓	✗	✗
LLaVA-Plus [58]	✓	✗	✗	✓	✓	✓
VideoChat [46]	✗	✓	✗	✗	✗	✗
Video-LLaMA [120]	✗	✓	✗	✗	✗	✗
Video-LLaVA [52]	✓	✓	✗	✗	✗	✗
Video-ChatGPT [61]	✗	✓	✗	✗	✗	✗
GPT4Video [99]	✗	✓	✗	✗	✓	✗
PG-Video-LLaVA [67]	✗	✓	✓	✓	✗	✗
NExT-GPT [104]	✓	✓	✗	✗	✓	✗
VITRON (Ours)	✓	✓	✓	✓	✓	✓

- **Hao Fei**, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, Shuicheng Yan. "VITRON: A Unified Pixellevel Vision LLM for Understanding, Generating, Segmenting, Editing". Submitted. 2024

Unified MLLM

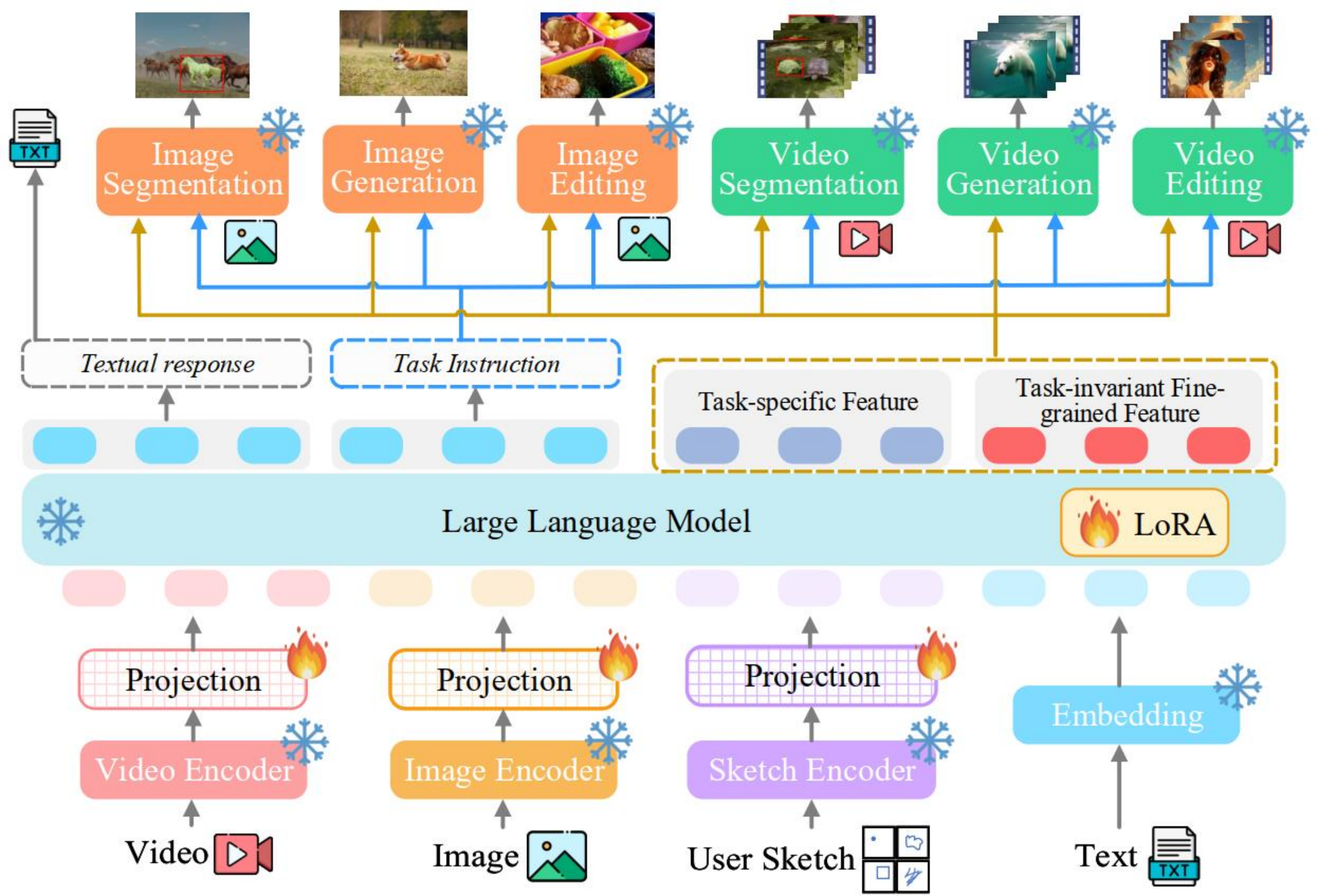
VITRON: A Unified Pixel-leve



*A universal pixel-level vision LLM designed for comprehensive understanding, generating, segmenting, and editing of both static **images** and dynamic **videos**.*

- **Hao Fei**, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, Shuicheng Yan. "VITRON: A Unified Pixellevel Vision LLM for Understanding, Generating, Segmenting, Editing". Submitted. 2024

Uni



TXT
Image Segmentation
Image Generation
Image Editing
Video Segmentation
Video Generation
Video Editing
Textual response
Task Instruction
Task-specific Feature
Task-invariant Fine-grained Feature
Large Language Model
LoRA
Projection
Projection
Projection
Embedding
Video Encoder
Image Encoder
Sketch Encoder
Video
Image
User Sketch
Text
TXT

Unified MLLM

VITRON: A Unified Pixel-level Vision MLLM

- Cross-task Synergy Learning
  - *Without any collaboration, integrating all existing specialists together might be meaningless.*
  - *How to ensure the different modules (tasks) work together synergistically?*

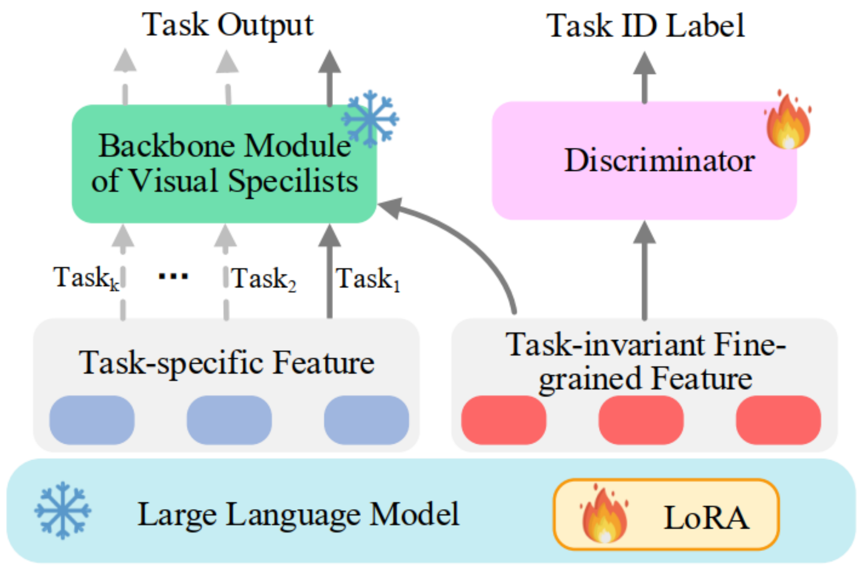


Figure 3: Illustration of the synergy module.

☞
- *decoupling task-specific features from task-invariant features;*
- *then use a third-party **discriminator** to determine the current task based solely on the shared task-invariant feature representation.*

Image Segmentation

Video Segmentation

Video Understanding

Video Editing

Unified MLLM

SeTok: Semantic Equivalence of Tokenization in MLLM

Project: https://chocowu.github.io/SeTok-web/

Paper: https://arxiv.org/abs/2406.05127

Code: https://github.com/ChocoWu/SeTok

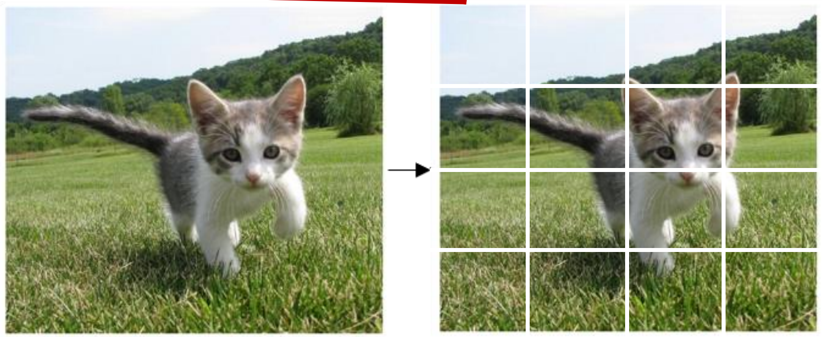
- Shengqiong Wu, **Hao Fei**, Hanwang Zhang, Tat-Seng Chua, Shuicheng Yan. Towards Semantic Equivalence of Tokenization in Multimodal LLM. 2024

Unified MLLM

SeTok: Semantic Equivalence of Tokenization in MLLM

- Existing Visual Tokenization: ***Patchify***

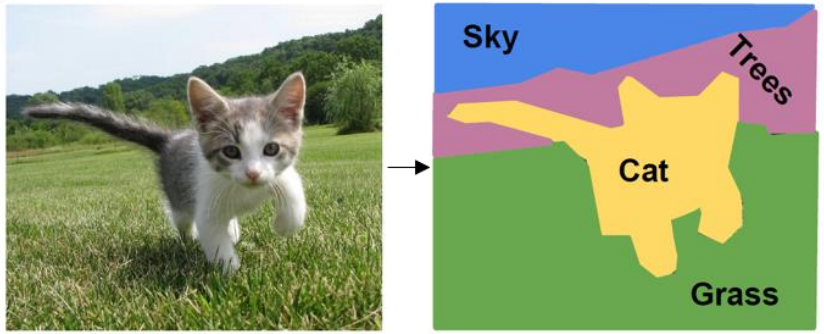
Integrity of visual semantic units is damaged.



Vision and language is not semantically equivalent

A cat is taking a walk on the grass

- Idea Visual Tokenization: ***Semantically Equivalent***



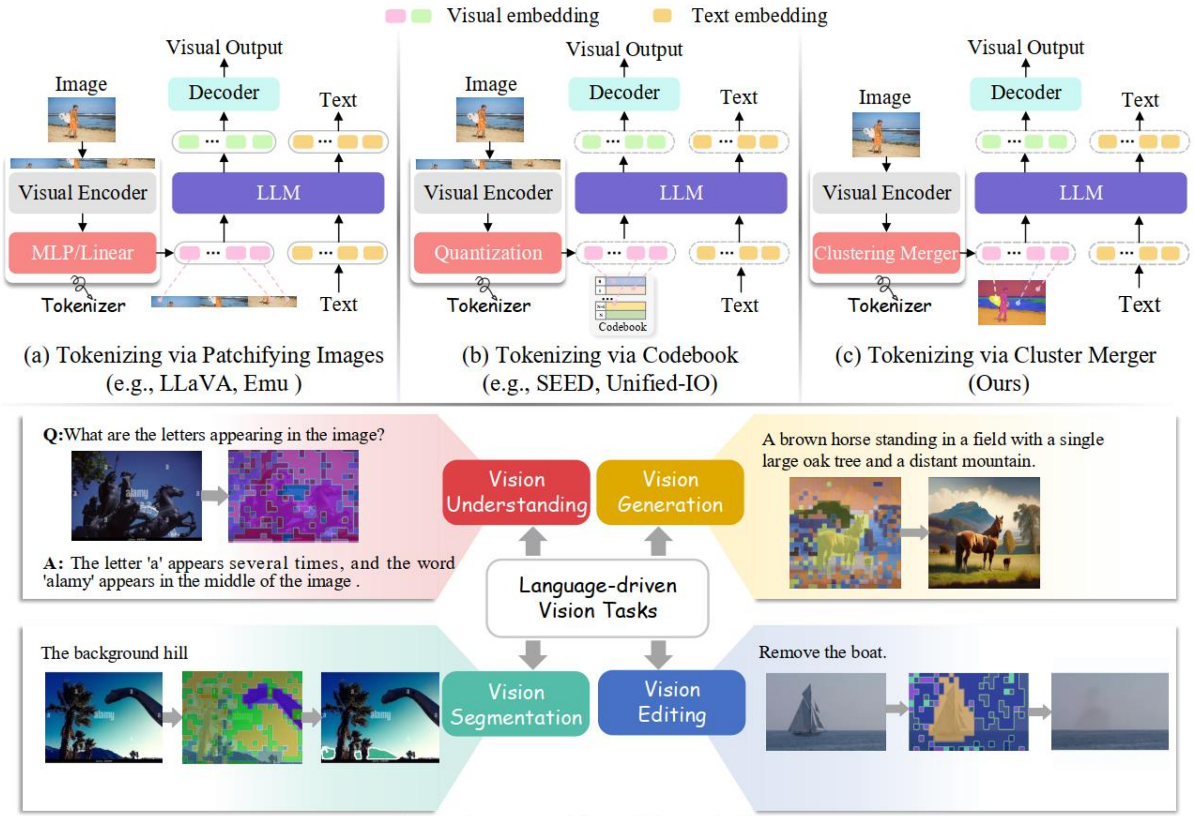
A cat is taking a walk on the grass

- Shengqiong Wu, **Hao Fei**, Hanwang Zhang, Tat-Seng Chua, Shuicheng Yan. Towards Semantic Equivalence of Tokenization in Multimodal LLM. 2024

Unified MLLM

SeTok: Semantic Equivalence of Tokenization in MLLM

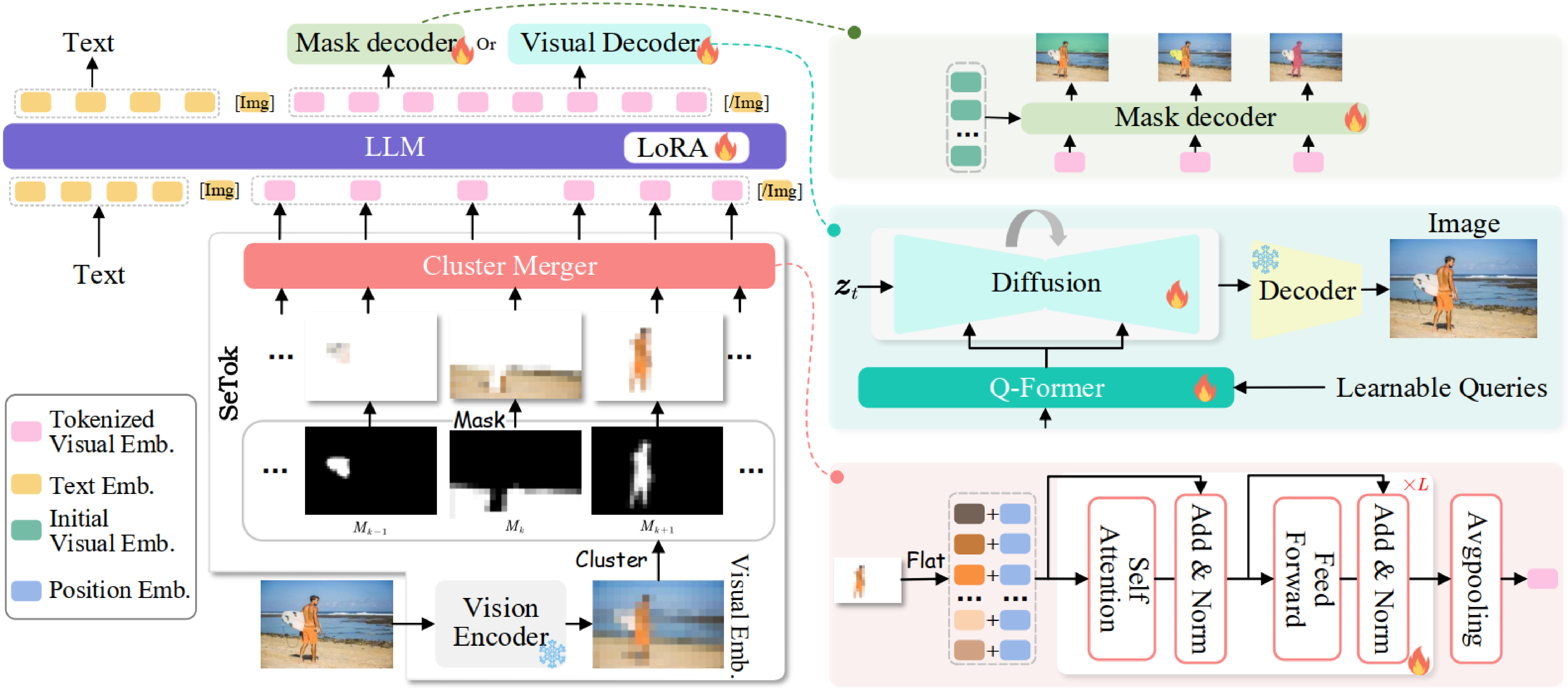
a Semantic-Equivalent Vision Tokenizer to achieve finer-grained semantic alignment between __visual__ and __text__ tokens, facilitating to improve various vision-language tasks.



(d) Four language-driven vision task clusters

- Shengqiong Wu, **Hao Fei**, Hanwang Zhang, Tat-Seng Chua, Shuicheng Yan. Towards Semantic Equivalence of Tokenization in Multimodal LLM. 2024

Unified MLLM



U

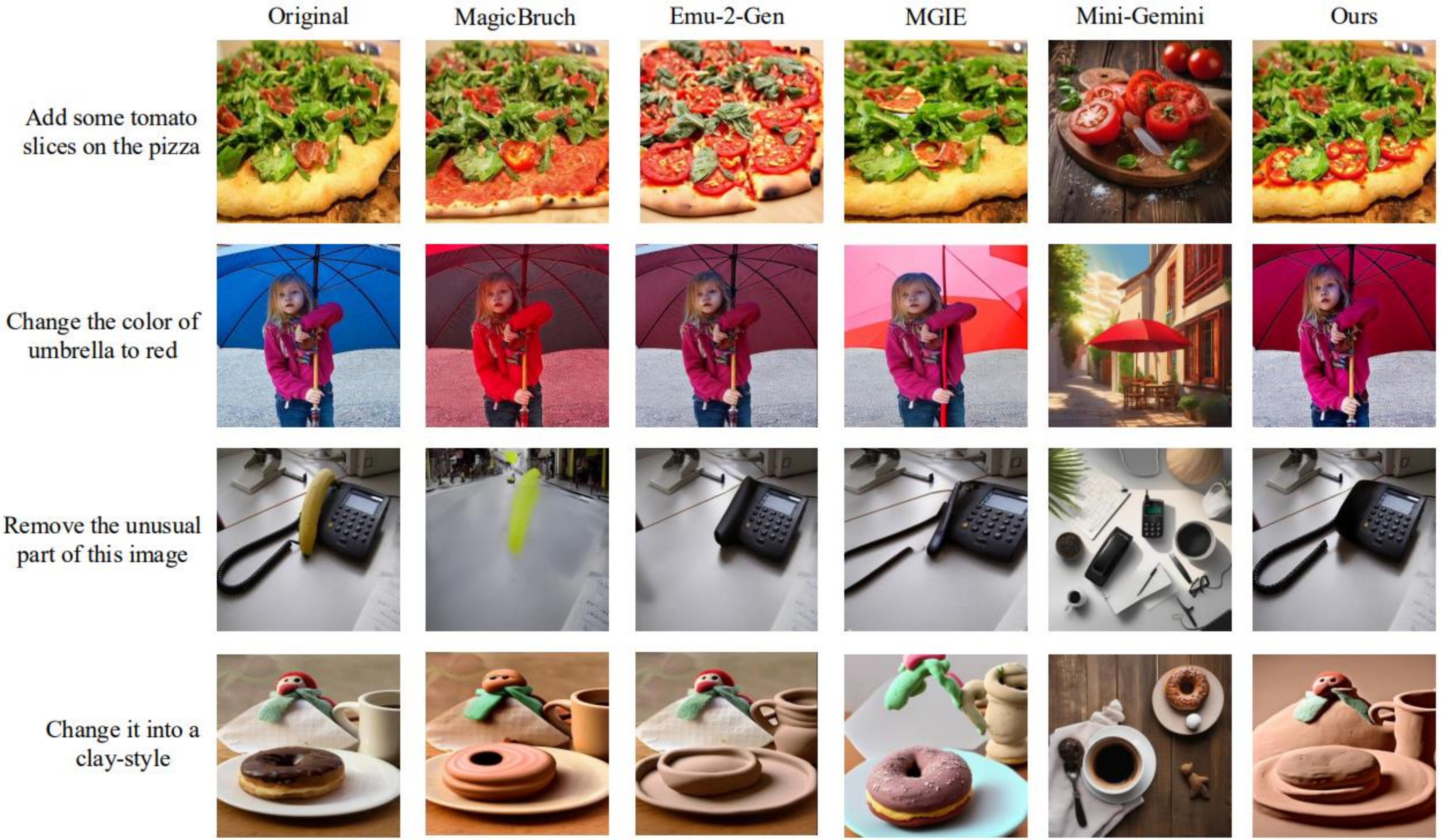


Figure 5: Qualitative comparison between MLLMs for the image editing. SETOKIM excels in adhering to instructions and preserving low-level image details.

Unified MLLM

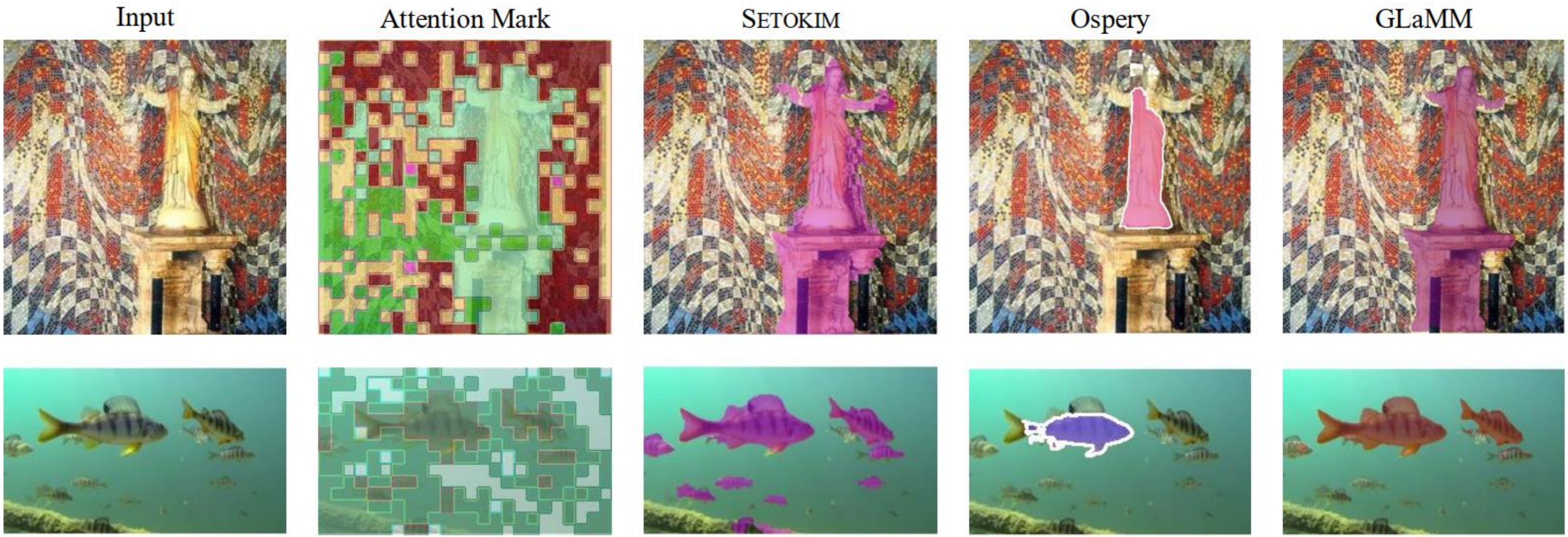


Figure 5: The visualizations for segmentation results compared with GLaMM[72] and Ospery [93].

Unified MLLM

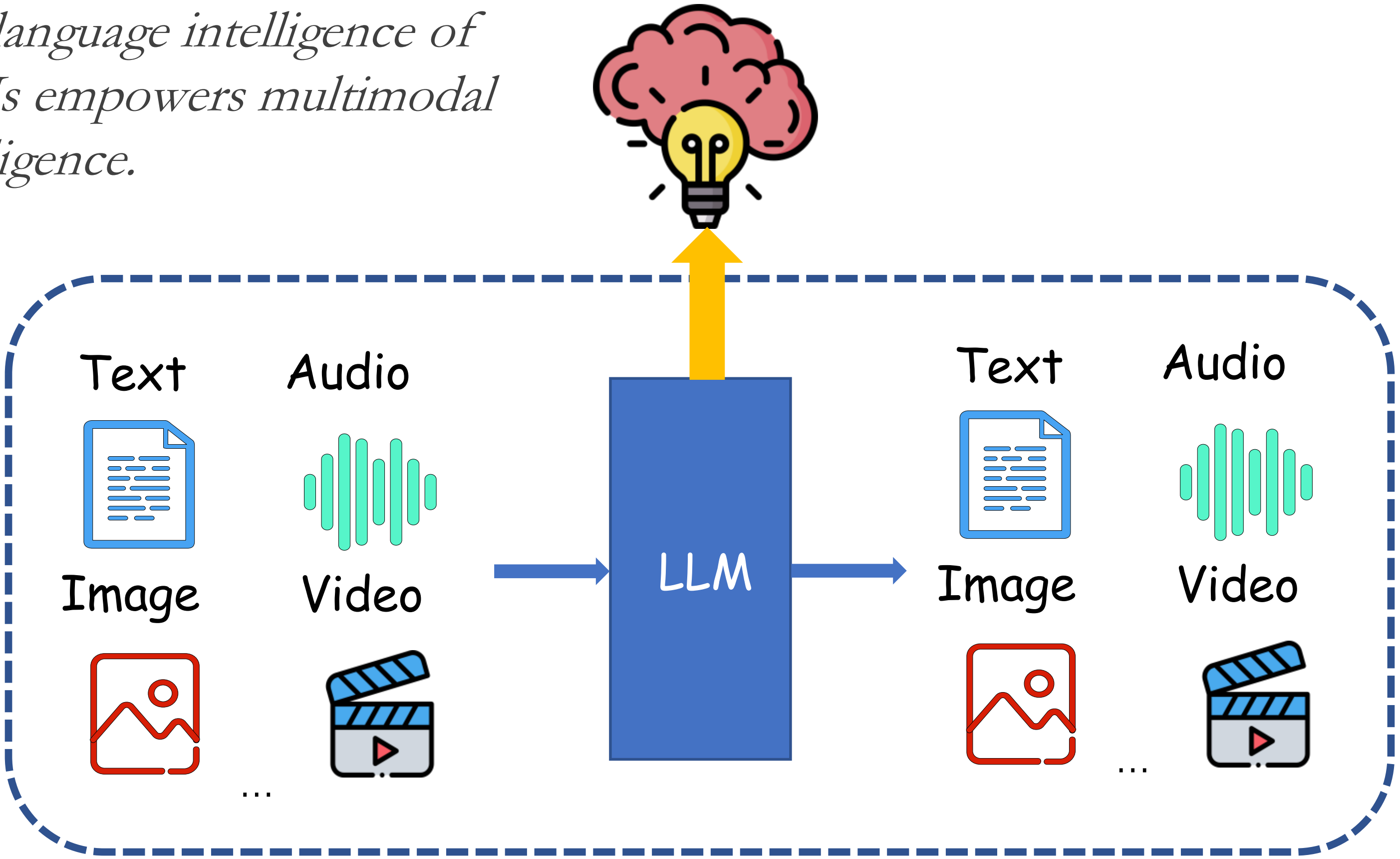
Figure 6: The visualizations for visual tokens.

Content

Towards Building Native MLLM

- Multimodal intelligence of MLLM relies on language’ s intelligence

The language intelligence of LLMs empowers multimodal intelligence.

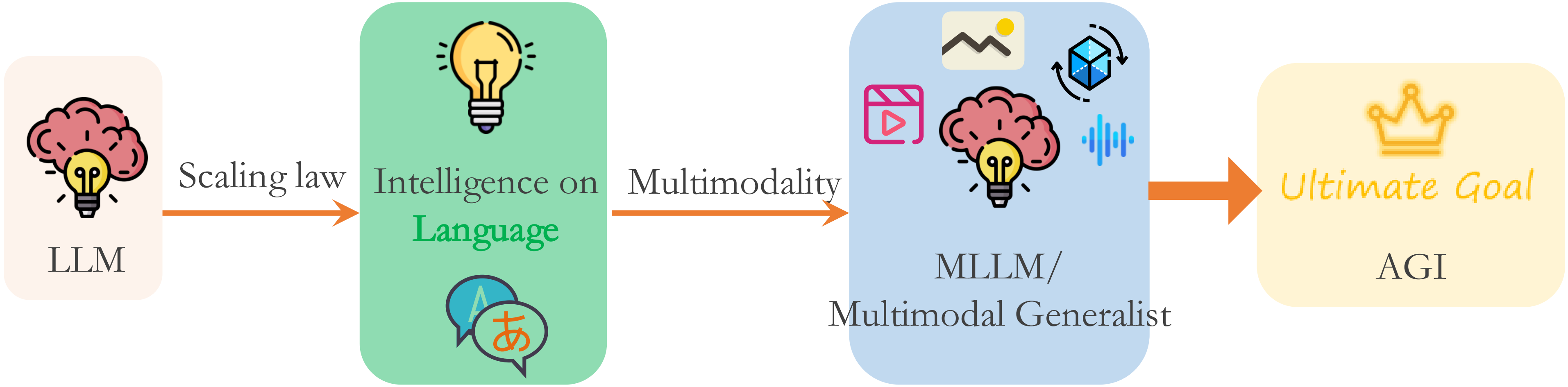


MLLM/Generalist

Towards Building Native MLLM

- Multimodal intelligence of MLLM relies on language's intelligence

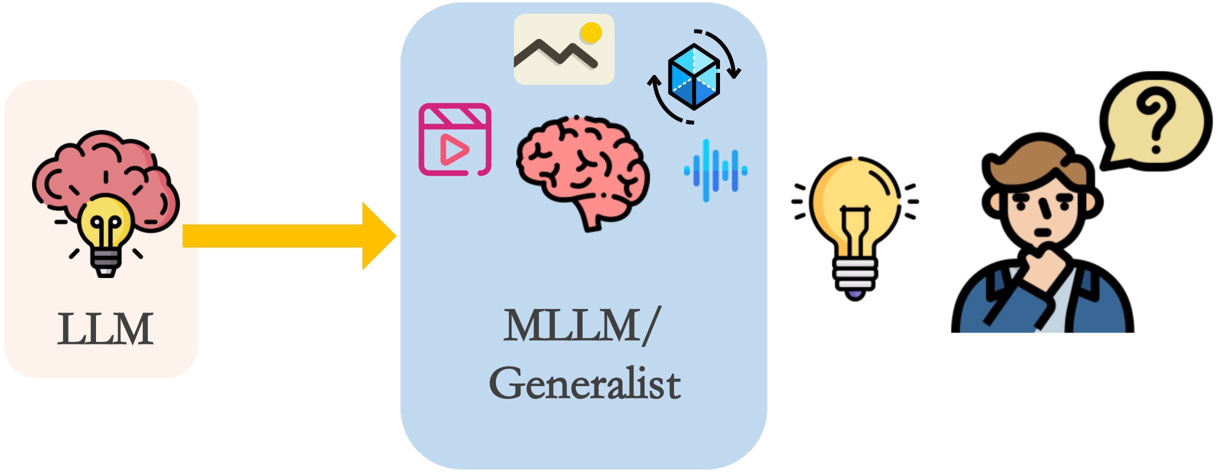
The language intelligence of LLMs empowers multimodal intelligence.



Towards Building Native MLLM

Multimodal intelligence of MLLM relies on language’s intelligence

- Could the scaling law and emergence success of LLMs be replicated in multimodality to achieve the intelligence of native MLLMs?

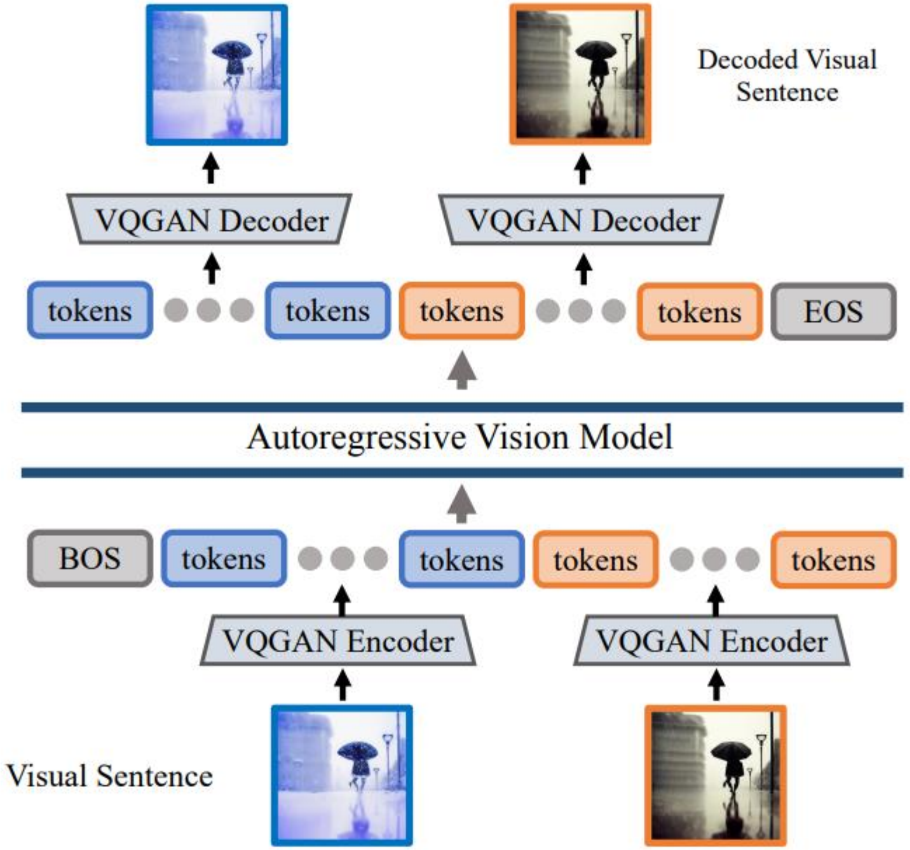


Towards Building Native MLLM

Exploration#1

➢ Large Vision Model (LVM)

- mimicking LLM pretraining
- next visual token prediction



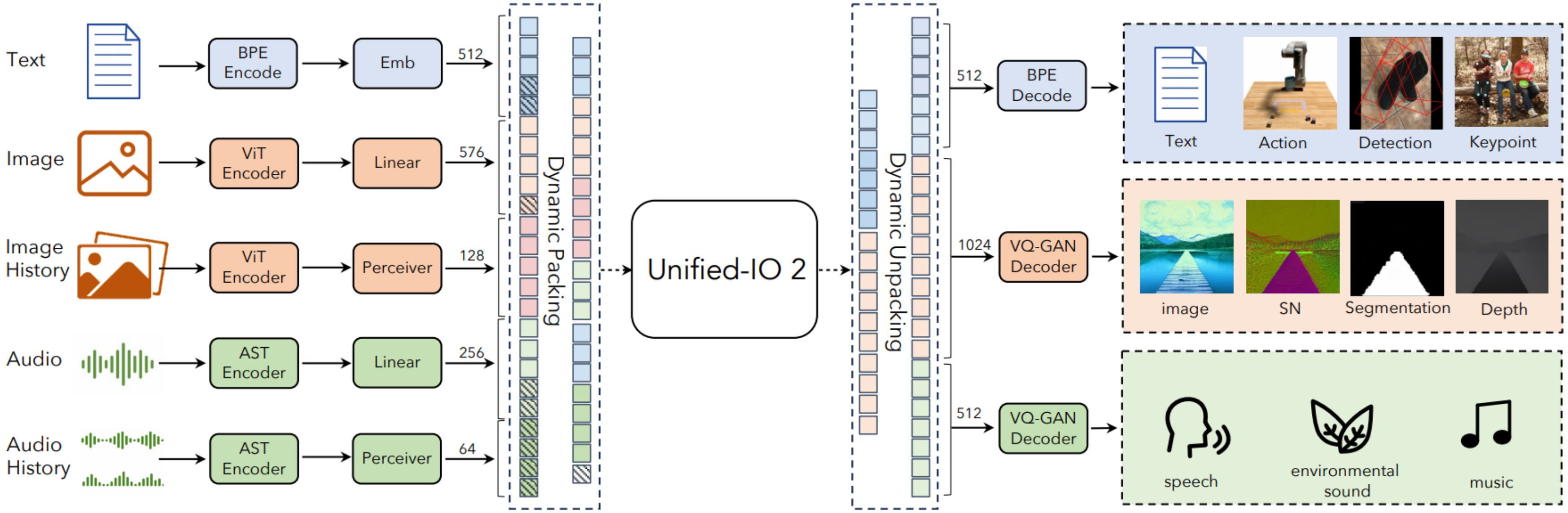
- Yutong Bai, Xinyang Geng, Karttikeya Mangalam etc. Sequential Modeling Enables Scalable Learning for Large Vision Models. CVPR. 2024

Towards Building Native MLLM

Exploration#2

➢ Unified IO-2

- mimicking LLM pretraining
- next visual token prediction



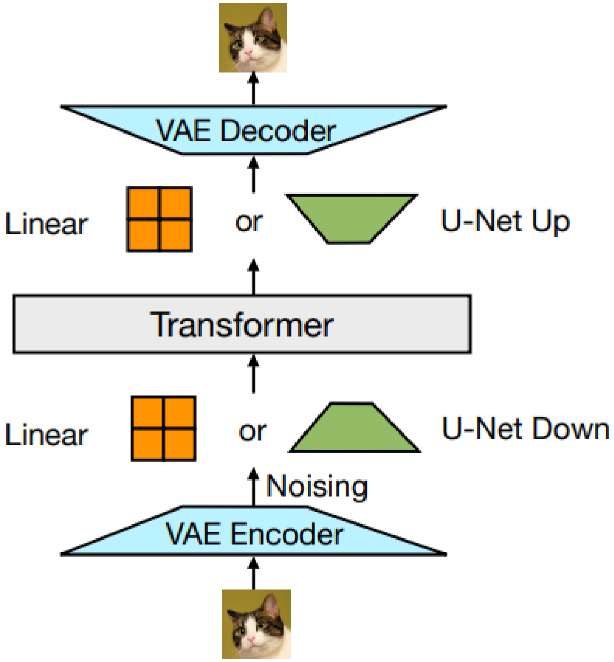
- Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., ... & Kembhavi, A. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action. CVPR. 2024

Towards Building Native MLLM

Open Question #1

What is the optimal model architecture under unified MLLM?

- Pipeline Agent
- Joint Encoder+LLM+Diffusion
- Joint LLM^{AR} Tokenization (VQ-VAE)
- Joint LLM^{AR}+Diffusion



- Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, Kaiming He. Autoregressive Image Generation without Vector Quantization. 2024.
- Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. 2024.
- Zhou, Chunting, et al. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. 2024.

Towards Building Native MLLM

Open Question #2

What scale of dataset is required for pre-training from scratch?

Modality	LLM/MLLM	Amount
Language	Chat-GPT4	13 Trillion text tokens
Vision	LVM	420 Billion visual tokens
Multimodalities	Unified-IO 2	1 Trillion text tokens, 1 Billion image-text pairs, 180 Million video clips, 130 Million interleaved image & text, 3 Million 3D assets, 1 Million agent trajectories

Towards Building Native MLLM

Open Question #3

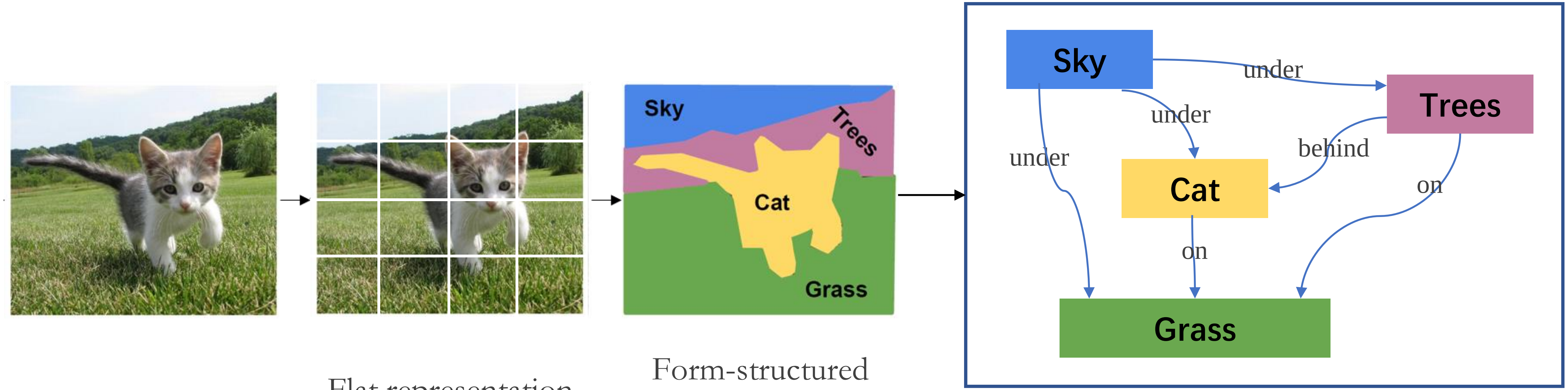
There is a gap of the downstream task performance between **native MLLMs** and **SoTA "LLM+encoder/decoder" architecture MLLMs**.

How can this gap be bridged?

Towards Building Native MLLM

Open Question #4

☞ What is the optimal representation method for multimodal data?



Flat representation

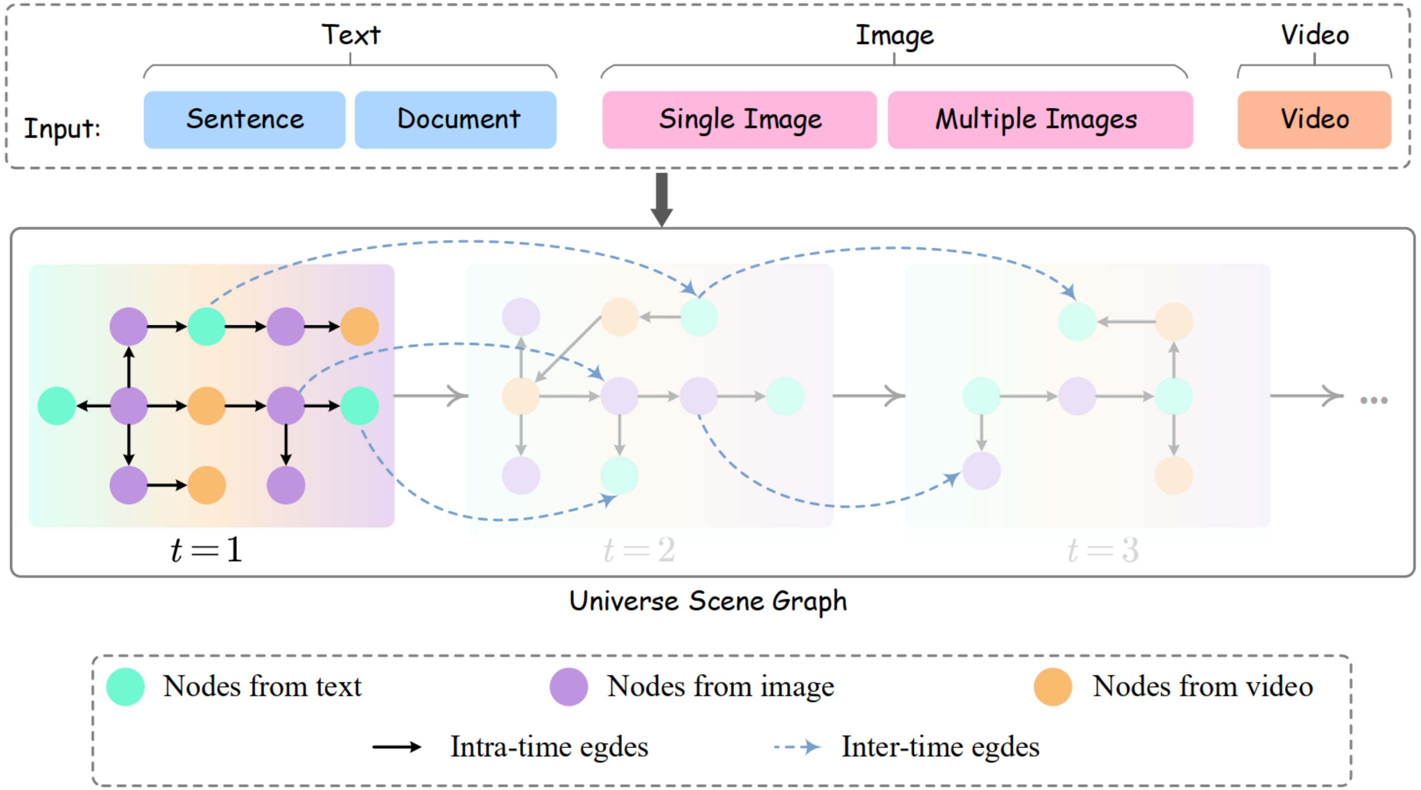
Form-structured representation

Semantically-structured representation

Towards Building Native MLLM

Pre-training MLLM over Universal Scene Graph (USG) Representation

☞ *USG: A topological structure of a scene description from text, image, video, or any combination of modalities.*



Towards Building Native MLLM

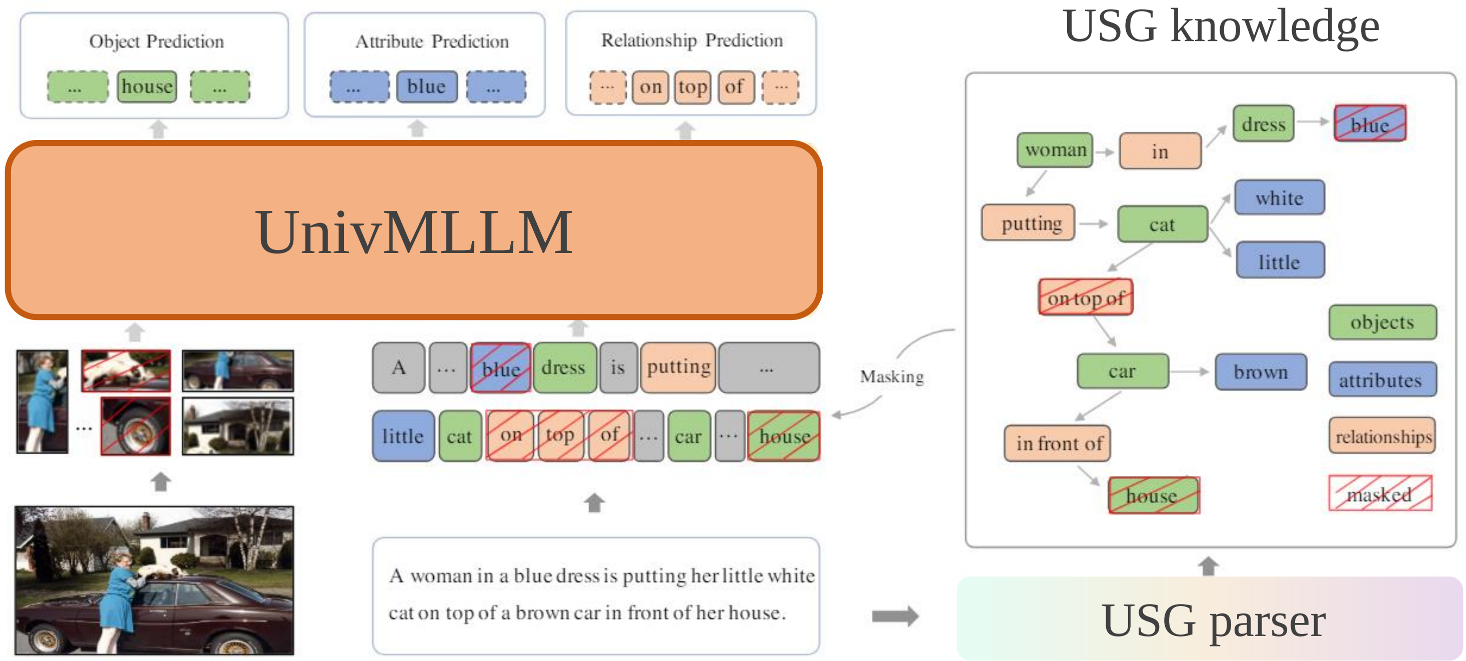
Pre-training MLLM over Universal Scene Graph (USG) Representation

☞ *masking and predicting different types of elements in the USG:*

1) masked object node prediction

2) masked attribute node prediction

3) masked relation prediction

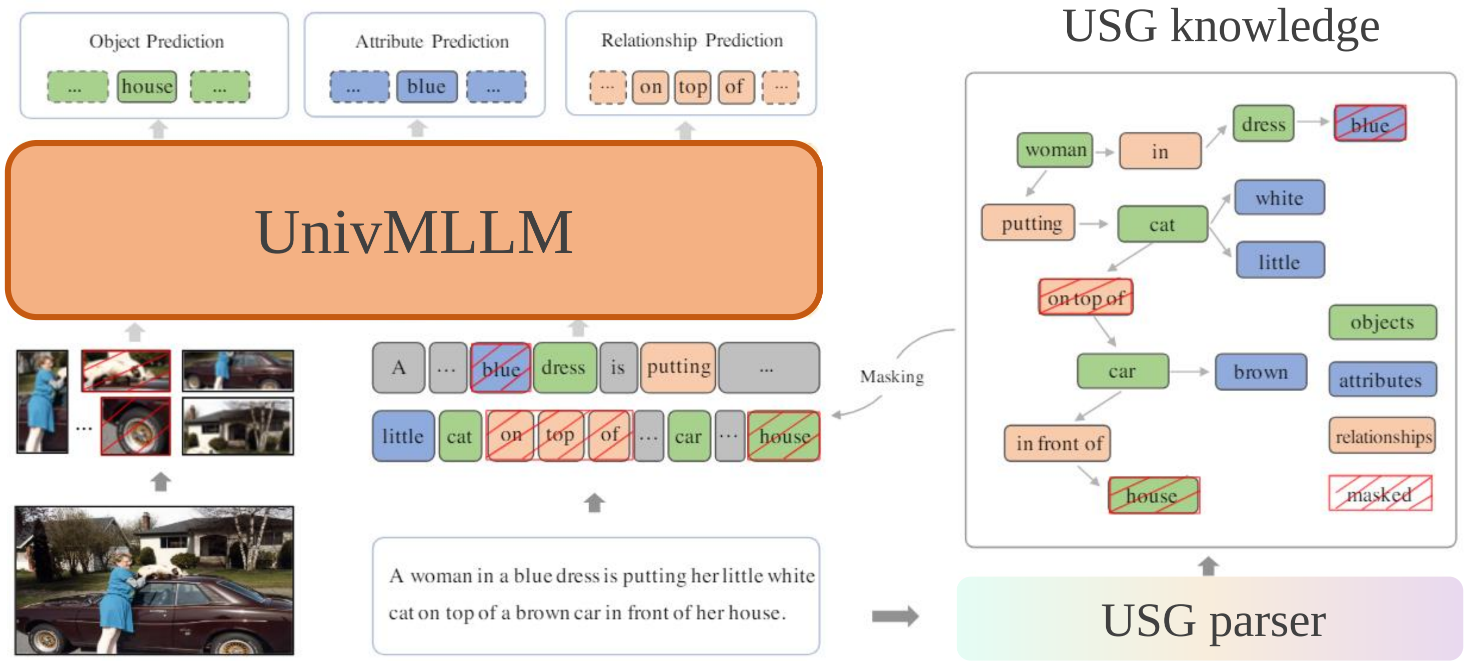


Towards Building Native MLLM

Pre-training MLLM over Universal Scene Graph (USG) Representation

Advances:

- *Seamlessly universal cross-modal representation*
- *Fine-grained semantical alignment between various modalities*
- *Universal modeling of various modalities and tasks*

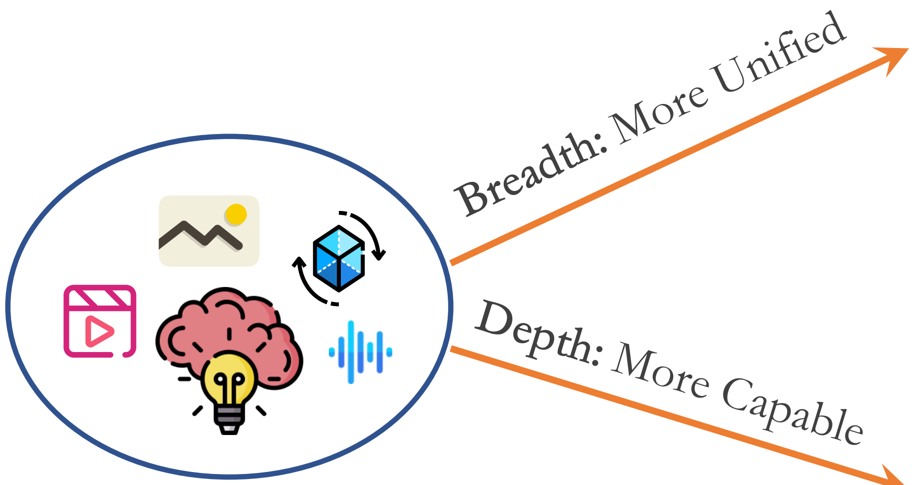


Content

Path to Multimodal Generalist

Multimodal Generalist Capability

➢ MLLMs should further enhance capabilities both in **breadth** and **depth**.



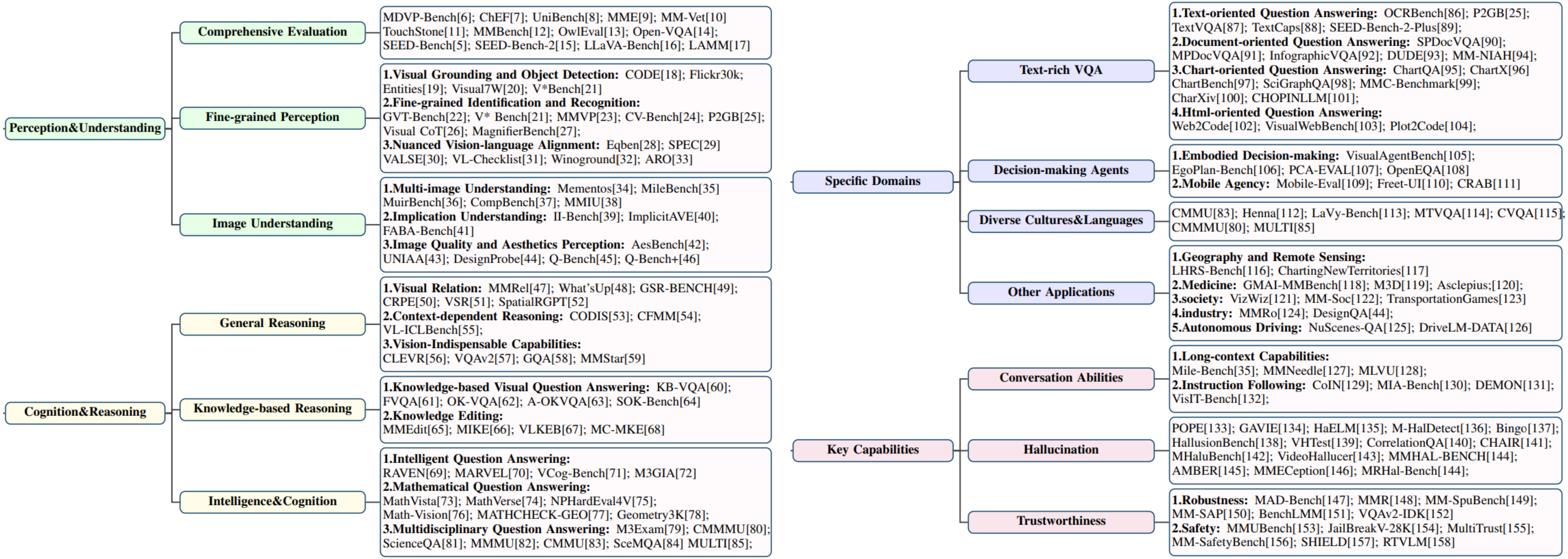
MLLM/
Multimodal Generalist

- Supporting More Tasks
- Supporting More Modalities

- Longer Context/Memory
- Reasoning
- Cross-modal Generalizability
- …

Path to Multimodal Generalist

MLLM Evaluation



- Li, J., & Lu, W. (2024). A Survey on Benchmarks of Multimodal Large Language Models. 2024.
- Huang, J., & Zhang, J. (2024). A Survey on Evaluation of Multimodal Large Language Models. 2024

Path to Multimodal Generalist

MLLM Evaluation

➤ *Higher performance simply indicate a stronger MLLM capability?*

- **Multimodal Comprehension vs. Multimodal Comprehension+Generation**

  An MLLM that only has multimodal comprehension capabilities represents the most basic and primitive level; we believe that the more powerful an MLLM is, the more it should support advanced functionalities, capable of both multimodal comprehension and generating content across various modalities.

- **More and Broader Modalities and Task Paradigms**

  The stronger the MLLM and the closer it is to AGI, the more task types it can support, the more modalities it can handle, and the stronger its task performance.

- **A Strong Synergy Effect is the Core Aspect of an MLLM**

  Synergy is the most critical aspect when assessing whether a multimodal generalist is stronger! An MLLM should be able to achieve a synergistic effect where 1+1 > 2, such as understanding a single modality & task that can be transferred to understanding other tasks & modalities, similar to the ChatGPT series, which can achieve robust generalization abilities with minimal training examples.

Path to Multimodal Generalist

Project: https://path2generalist.github.io/

Paper: Coming soon

Benchmark: https://github.com/path2generalist/GeneralBench

- **Hao Fei**, Yuan Zhou, ···, Hanwang Zhang, Shuicheng Yan. Path to Multimodal Generalist: Level, Benchmark and Model. TBD. 2024

Path to Multimodal Generalist

Path to Generalist

Overview Level Benchmark Leaderboard Contact

Path to Multimodal Generalist: *Levels*, *Benchmarks* and *Models*

Hao Fei[*,1,2], Yuan Zhou[*,3], Juncheng Li[1], Xiangtai Li[2], Yucheng Han[3], Wentao Hu[3], Liyu Jia[3], Shengqiong Wu[1], Peng Zhou[6], Lin Liu[7], Haobo Yuan[3], Tao Zhang[4], Bobo Li[4], Zixiang Meng[4], Chengjie Zhou[4], Minghe Gao[5], Kaihang Pan[5], Yaobo Ye[5], Mingze Zhou[5], Zhiqi Ge[5], Hanwang Zhang[†,2,3], Shuicheng Yan[2]

(* Equal contribution † Correspondence)

[1]National University of Sigapore, [2]Skywork AI, Singapore,
[3]Nanyang Technological University, [4]Wuhan University,
[5]Zhejiang University, [6]Shanghai Jiao Tong University,
[7]University of Science and Technology of China

Paper Code&Data

🔔News

🔥**[2024-06-16]: We release our first version of benchmark datasets, and the leaderboard!**

Path to Multimodal Generalist

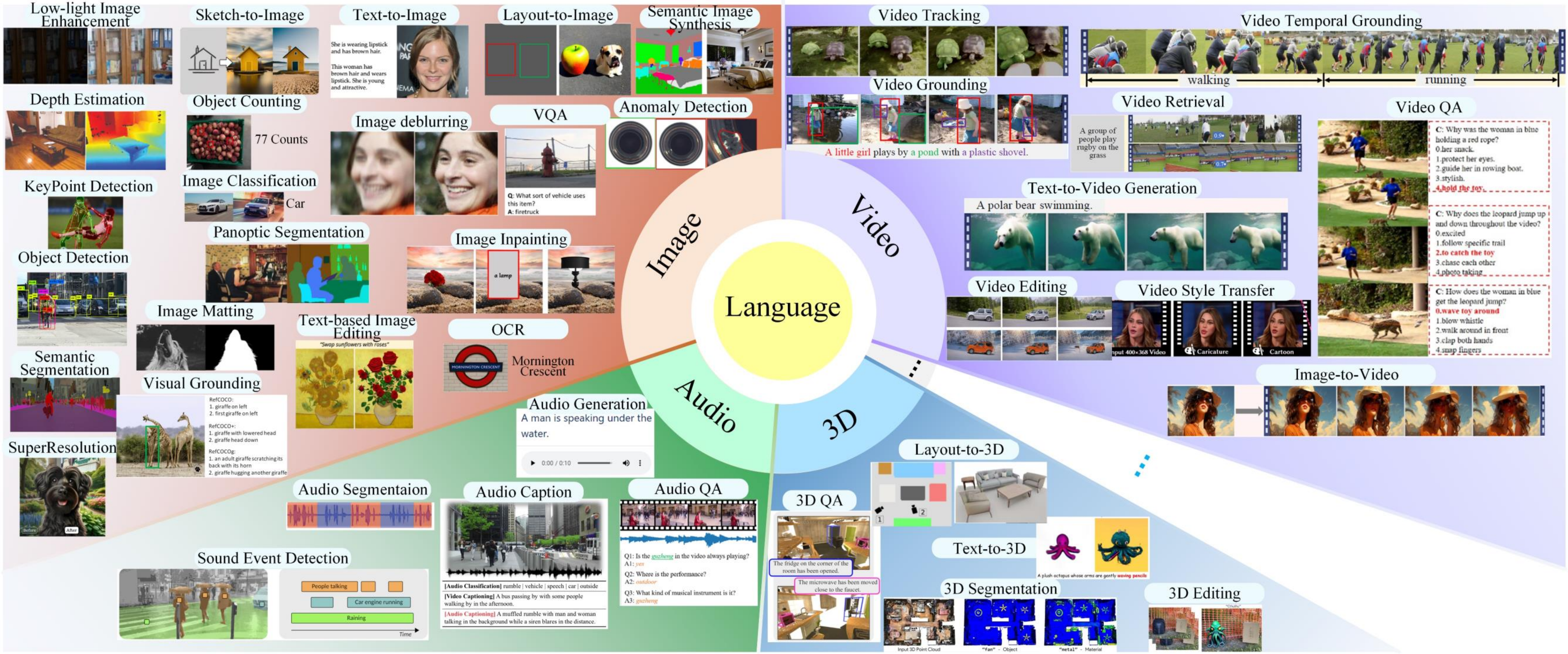
Level of Multimodal Generalist

Level	Definition	Scoring	Example
Level-1: Specialist	Various current models, each fine-tuned on a specific task or dataset of specific modalities, are task-specific players (i.e., state-of-the-art (SoTA) specialists). This includes various AI processing tasks, such as recognition, classification, text generation, image generation, video segmentation, grounding, inpainting, and more.	For each task in the benchmark (i-th task), record the current SoTA specialist's score: σ_i^{sota}	SAM, Dino, DALLe, ChatGPT
↓ **Upgrading Conditions:** *LLM as intelligence medium (Comprehension or/and Generation)*			
Level-2: Generalist of Unified Comprehension and Generation	Models are task-unified players, e.g., MLLMs, capable of supporting different modalities and tasks. Such MLLMs can integrate various models through existing encoding and decoding technologies to achieve aggregation and unification of various modalities and tasks (such as comprehension and generation tasks).	The average score across all datasets is used as the model's score at this level. A model that can support a task, or scores non-zero on a corresponding dataset, is considered capable of supporting that task. The more tasks a model supports and the higher its scores, the higher its overall score: $S_2 = \frac{1}{M+N}\sum_{i=1}^{M+N}\sigma_i$	GPT4v, llava, LVM
↓ **Upgrading Conditions:** *Realizing synergy: multi-task joint learning*			
Level-3: Generalist with Synergy in Comprehension and Generation	Models are task-unified players, and synergy is in Comprehension and/or Generation. MLLMs enhance several tasks' performance beyond corresponding SoTA scores through joint learning across multiple tasks due to the synergy effect.	Assign a mask weight of 0 or 1 to each task; assign mask=1 only if the corresponding score exceeds the SoTA specialist's score, otherwise assign mask=0. Then, calculate the average score across all tasks. The more tasks a model surpasses the SoTA specialist, the higher its score at this level: $S_3 = \frac{1}{M+N}\sum_{i=1}^{M+N}\begin{cases}\sigma_i, & \sigma_i \geq \sigma_i^{sota} \\ 0\end{cases}$	MM-GPT, SALOMNN, Midjourney
↓ **Upgrading Conditions:** *Reconstruction loss for generation should be disentangled from compression learning loss*			
Level-4: Generalist with Synergy across Comprehension and Generation	Models are task-unified players, and synergy is across Comprehension and Generation.	Calculate the average scores exceeding SoTA specialists separately in the Comprehension and Generation groups, obtaining S_c and S_g, and then compute their harmonic mean. The stronger a model is in Comprehension and Generation tasks, the higher its score at this level: $S_4 = \frac{2S_cS_g}{S_c+S_g}$, where $S_g = \frac{1}{M}\sum_{i=1}^{M}\begin{cases}\sigma_i, & \sigma_i \geq \sigma_i^{sota} \\ 0\end{cases}$, $S_c = \frac{1}{N}\sum_{j=1}^{N}\begin{cases}\sigma_j, & \sigma_j \geq \sigma_i^{sota} \\ 0\end{cases}$	Emu2, NExT-GPT, SEED
↓ **Upgrading Conditions:** *Acquiring the capable of abductive reasoning, being context consistent, everything synergy*			
Level-5: Generalist with Total Synergy across Comprehension, Generation, and NLP	Models are task-unified players, preserving the synergy effect across Comprehension, Generation, and NLP. In other words, the model not only achieves cross-modality synergy between Comprehension and Generation groups but also further realizes synergy with language. The NLP's intelligence can enhance multimodal intelligence and vice versa; understanding multimodal information can also aid in understanding language.	First, calculate the model's average score exceeding SoTA NLP specialists on NLP benchmark data, normalize it to a [0,1] weight, and multiply it by the score from level 4 to determine the level 5 score: $S_5 = S_4 \times w_L$, where $w_L = \frac{S_L}{S_{total}}$, $S_L = \frac{1}{T}\sum_{k=1}^{T}\begin{cases}\sigma_k, & \sigma_k \geq \sigma_k^{sota} \\ 0\end{cases}$	None, this is our goal!

- **Hao Fei**, Yuan Zhou, ⋯, Hanwang Zhang, Shuicheng Yan. Path to Multimodal Generalist: Level, Benchmark and Model. TBD. 2024

Path to Multimodal Generalist

General-Bench




LaVIT-V2
Vitron
miniGemini-7B
NExT-GPT
Emu2(37B)
Level-2
Level-4
Level-3
Level-5

Thank you!

Q&A