Vitron: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing

Hao Fei[1,2] **Shengqiong Wu**[1,2] **Hanwang Zhang**[1,3]
Tat-Seng Chua[2], **Shuicheng Yan**[1]

[1] Skywork AI, Singapore
[2] National University of Singapore
[3] Nanyang Technological University

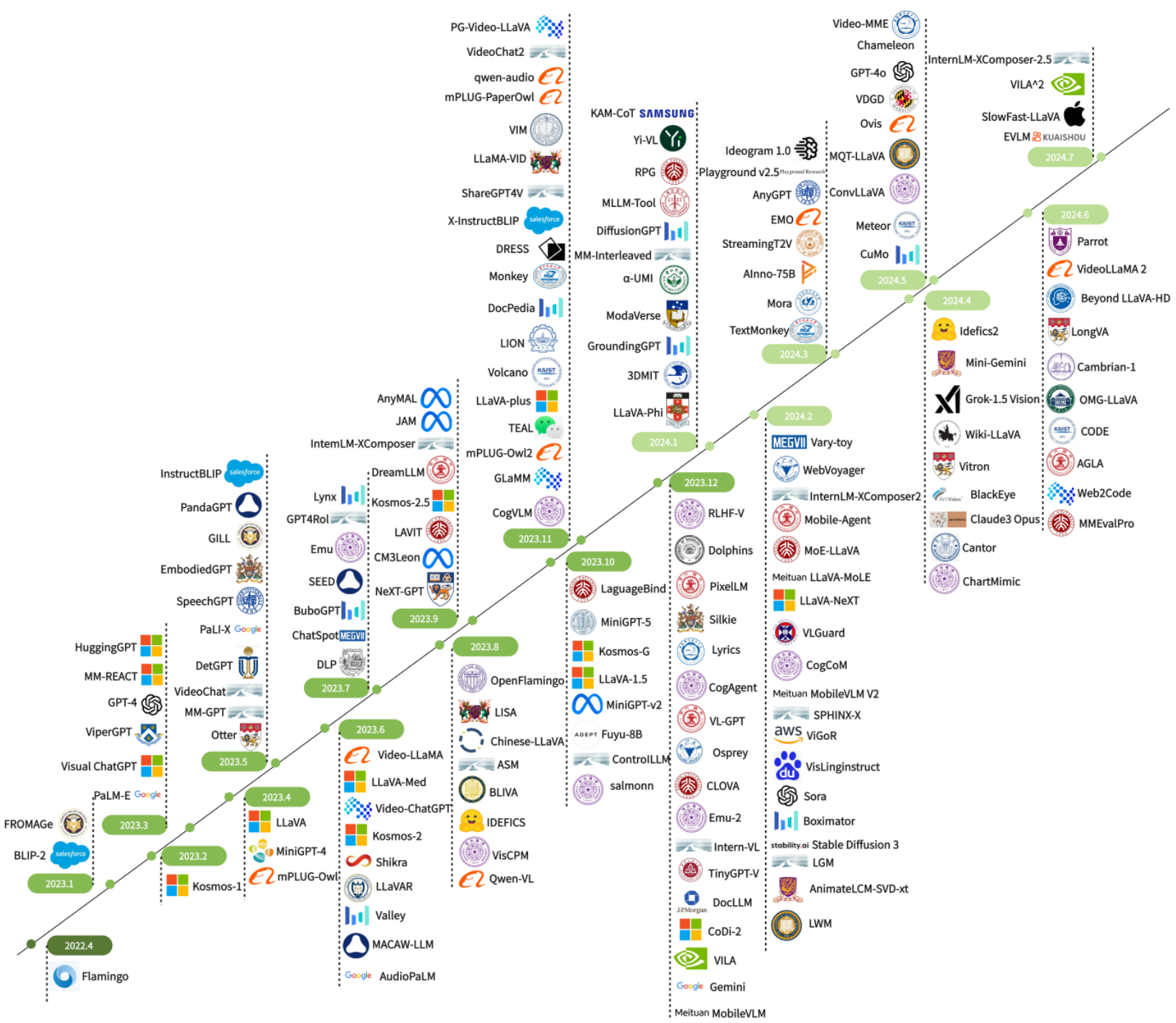
Project: https://vitron-llm.github.io/

Paper: https://is.gd/aGu0VV

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

1. Motivation

- Hot research attention ever in **MLLM**



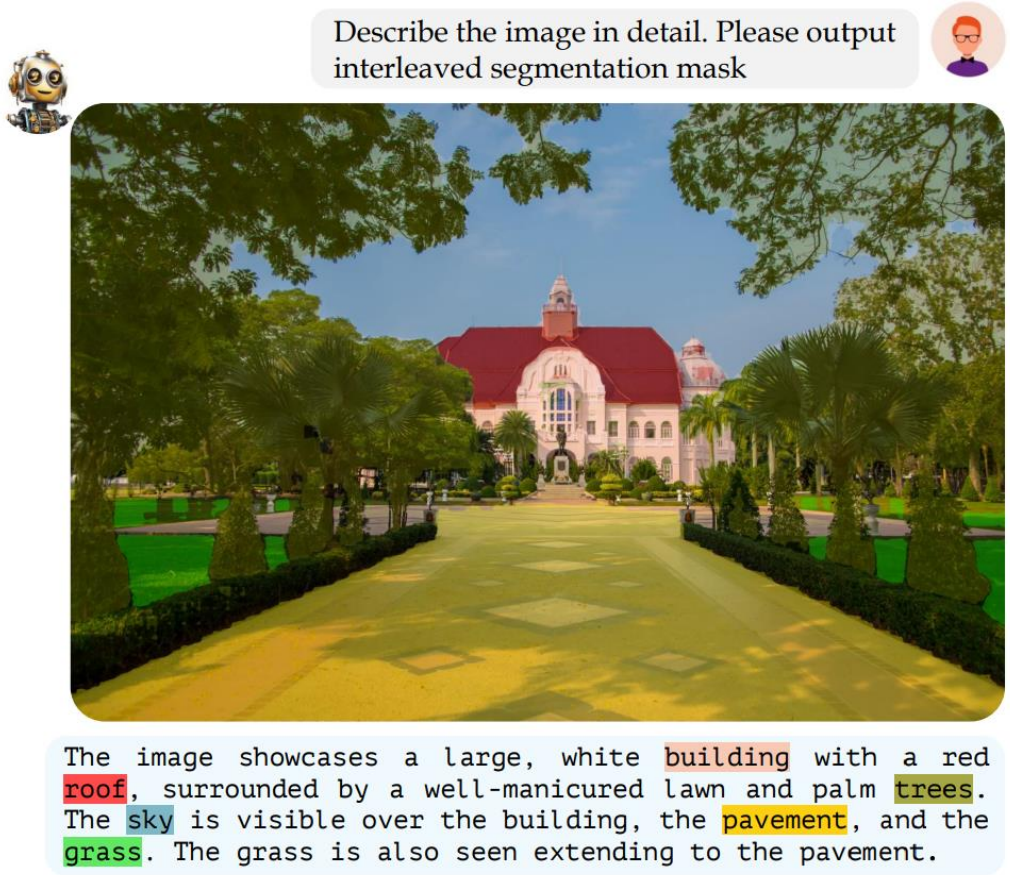
1. Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

➢ Existing vision LLM: not professional enough in visual task unification

Coarse-grained instance-level understanding

- *Most vision MLLMs only support coarse-grained, instance-level visual understanding.*
- *This can lead to* ***imprecise visual interpretations****.*
- *Also due to the lack of visual grounding, these MLLMs will potentially* ***produce hallucinations****.*

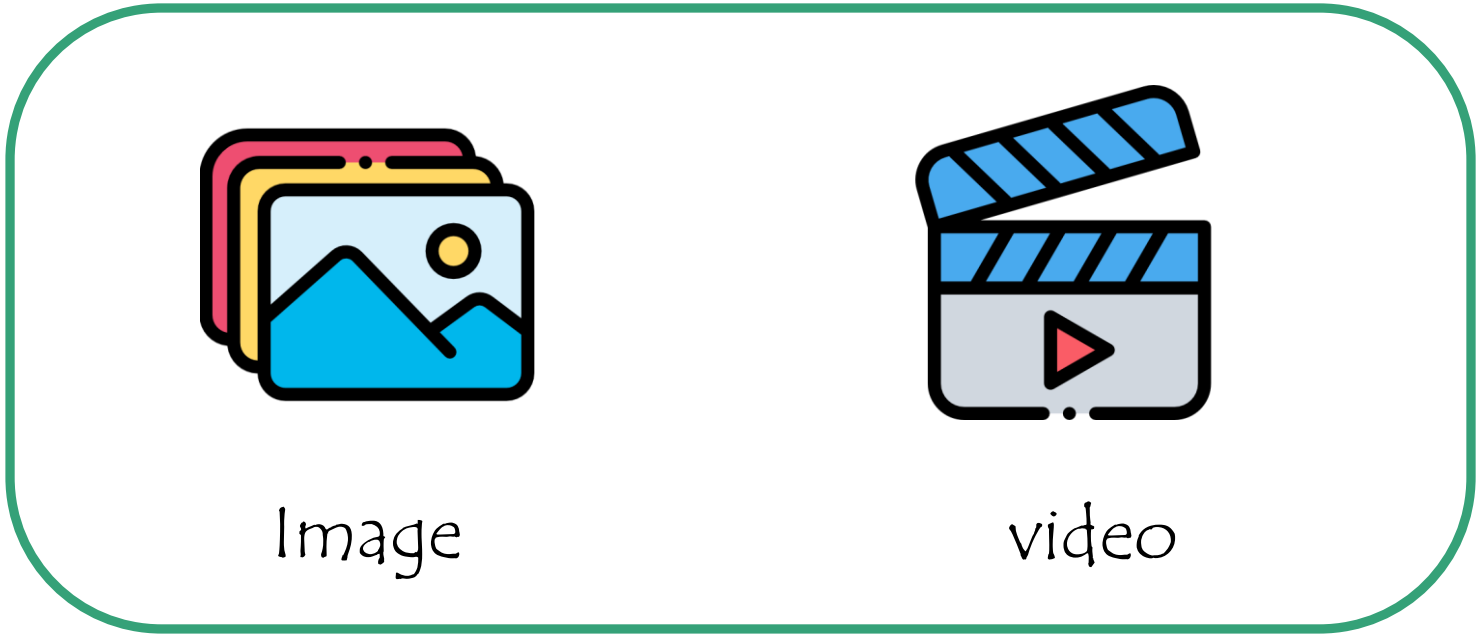


1. Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

- Existing vision LLM: not professional enough in visual task unification

Lack of unified support for both images and videos



Unified support

1. Motivation

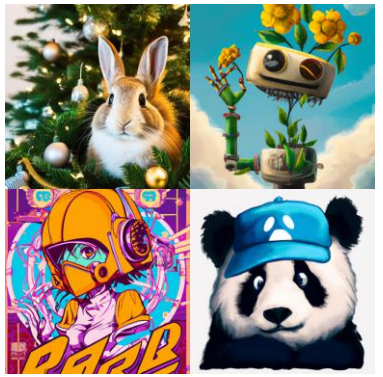
NEURAL INFORMATION PROCESSING SYSTEMS

➢ Existing vision LLM: not professional enough in visual task unification

Insufficient coverage across various vision tasks

- *Vision Segmentation & Grounding*
- *Vision Semantic Understanding & Reasoning*
- *Vision Synthesis & Generation*
- *Vision Editing & Inpainting*

Q: How many chromosomes do these creatures have?
A: 23



1. Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

- Existing vision LLM: not professional enough in visual task unification

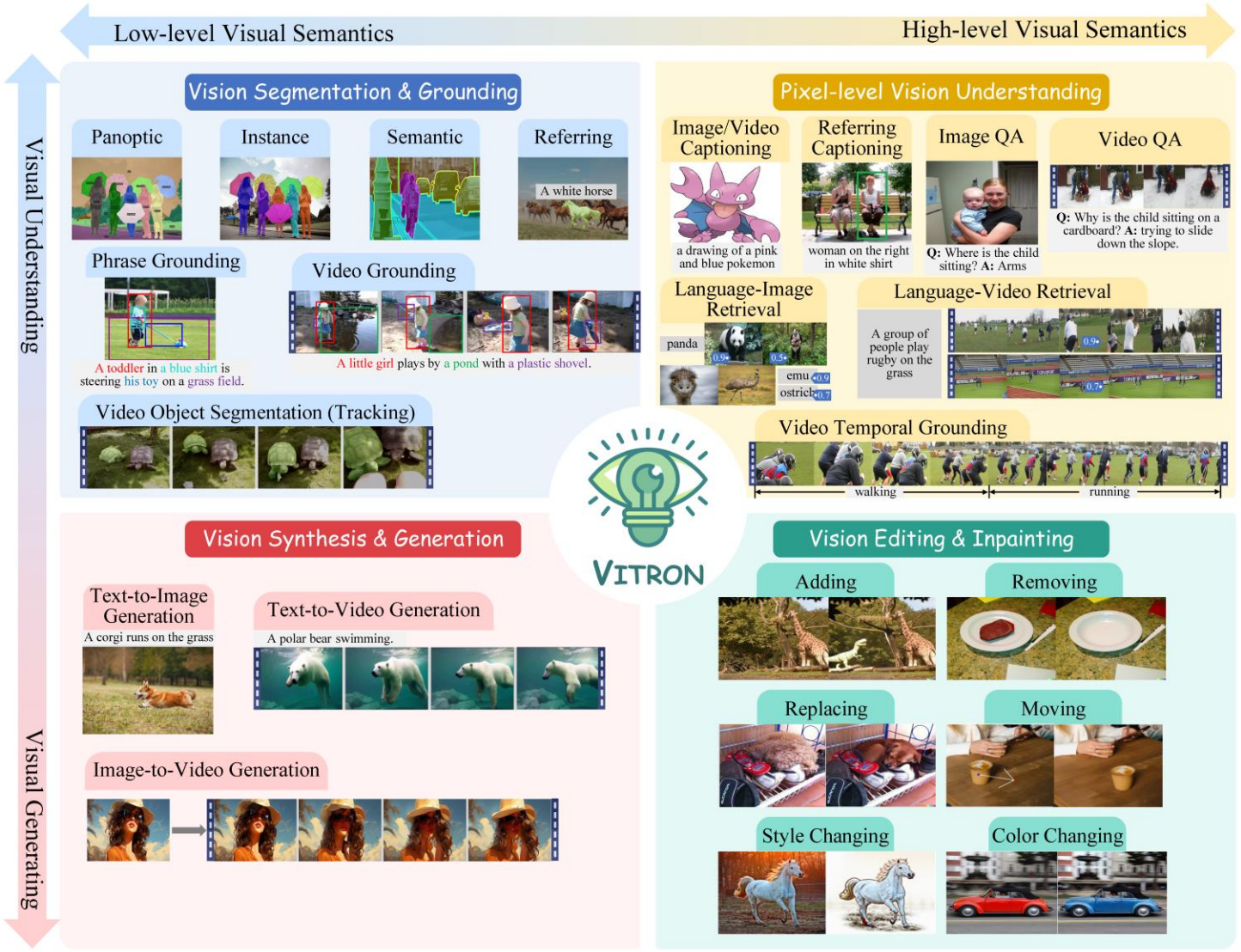
Model	Vision Supporting		Pixel/Regional Understanding	Segmenting/ Grounding	Generating	Editing
	Image	Video				
Flamingo [1]	✓	✗	✗	✗	✗	✗
BLIP-2 [45]	✓	✗	✗	✗	✗	✗
MiniGPT-4 [126]	✓	✗	✗	✗	✗	✗
LLaVA [57]	✓	✗	✗	✗	✗	✗
GILL [39]	✓	✗	✗	✗	✓	✗
Emu [90]	✓	✗	✗	✗	✓	✗
MiniGPT-5 [125]	✓	✗	✗	✗	✓	✗
DreamLLM [23]	✓	✗	✗	✗	✓	✗
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)	✓	✓	✓	✓	✓	✓

2. Proposed Model: Vitron

NEURAL INFORMATION PROCESSING SYSTEMS

- A Unified Pixel-level Vision MLLM

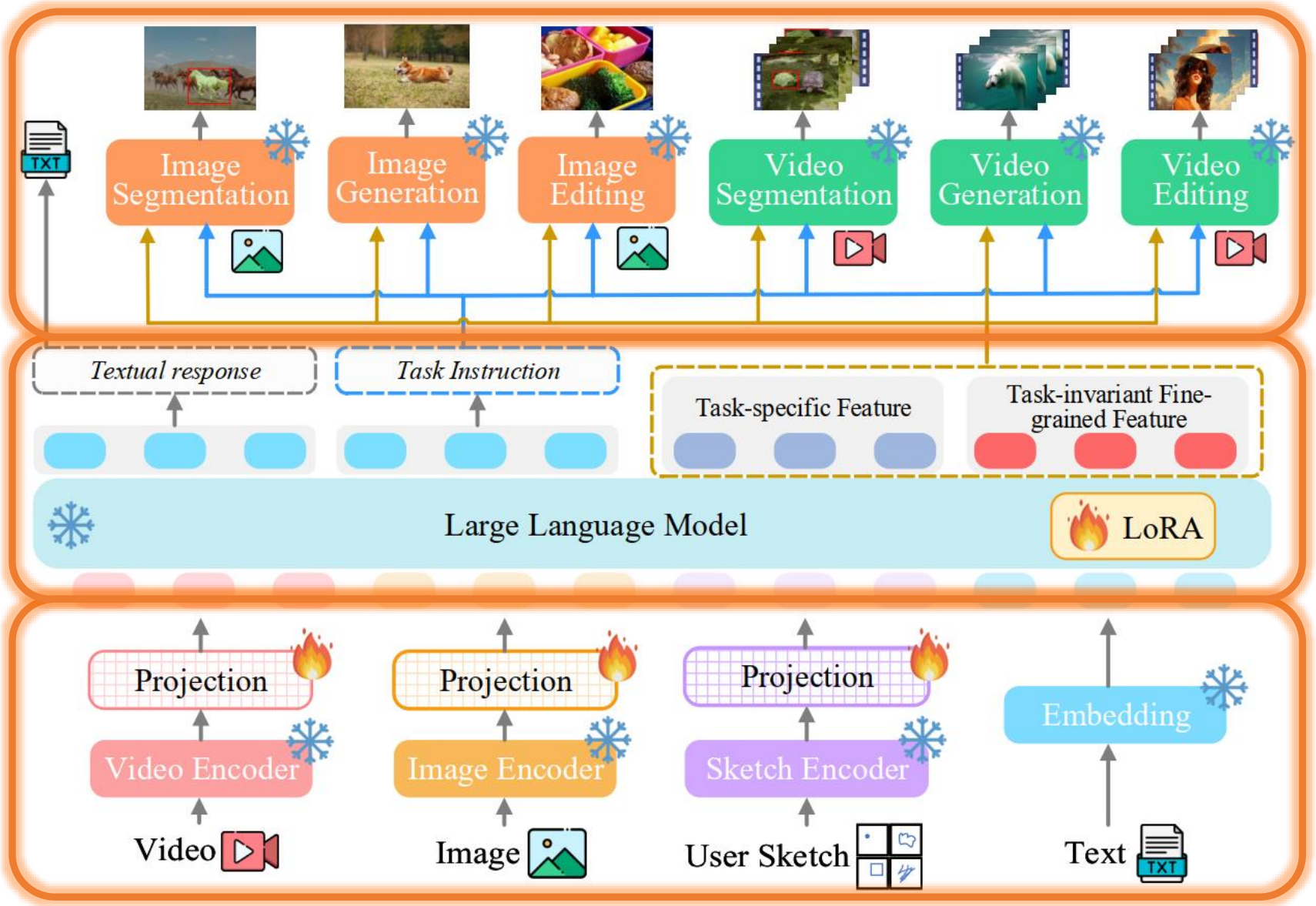
*A universal pixel-level vision LLM designed for comprehensive understanding, generating, segmenting, and editing of both static **images** and dynamic **videos**.*



2. Proposed Model: Vitron

NEURAL INFORMATION PROCESSING SYSTEMS

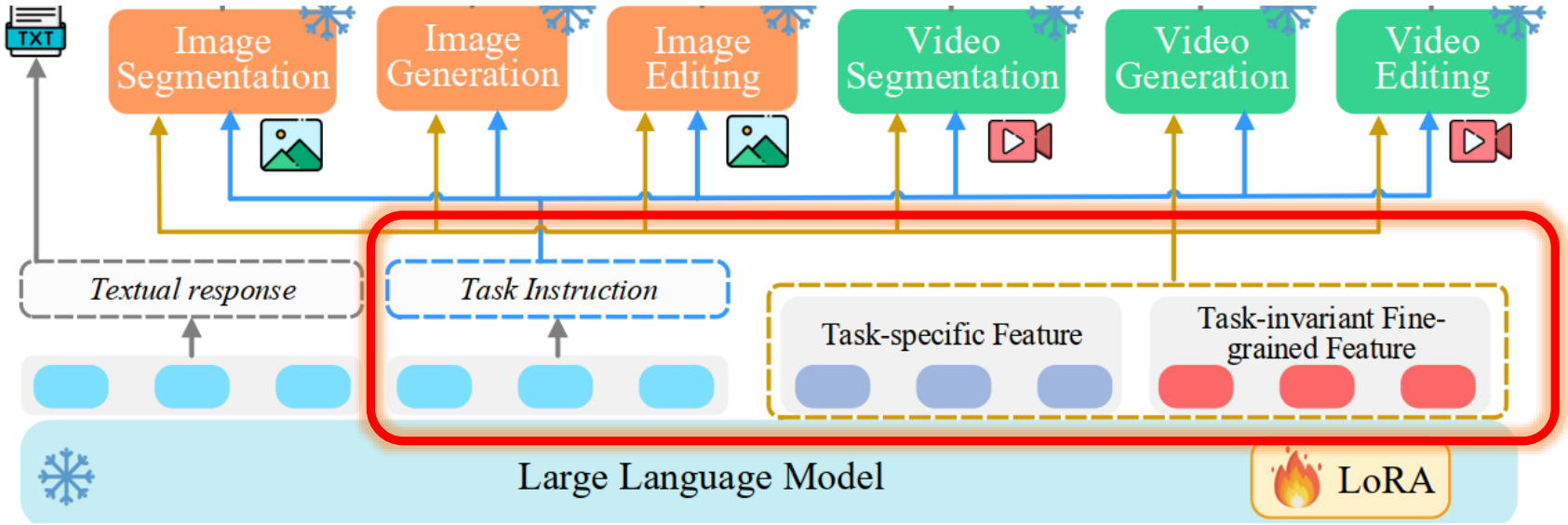
- Architecture



2. Proposed Model: Vitron

NEURAL INFORMATION PROCESSING SYSTEMS

➢ Hybrid LLM-to-decoder instruction-passing mechanism



- *Via Discrete texts:*

Accurately invoking different backbone modules

- *Via Continuous signal embeddings :*

Supplementing with richer modality-preserved visual features that cannot be directly described through discrete text

2. Proposed Model: Vitron

NEURAL INFORMATION PROCESSING SYSTEMS

- Pixel-aware Synergistic Vision-Language Understanding Tuning
  - *Basic Multimodal Comprehension and Generation Skill Training*
    - *Overall Vision-Language Alignment Learning*
    - *Text Invocation Instruction Tuning*
    - *Embedding-oriented Decoder Alignment Tuning*
  - *Fine-grained Spatiotemporal Vision Grounding Instruction Tuning*
    - *Image Spatial Grounding*
    - *Video Spatial-Temporal Grounding*
    - *Grounding-aware Vision QA*
  - *Cross-task Synergy Learning*

2. Proposed Model: Vitron

NEURAL INFORMATION PROCESSING SYSTEMS

➢ 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?*

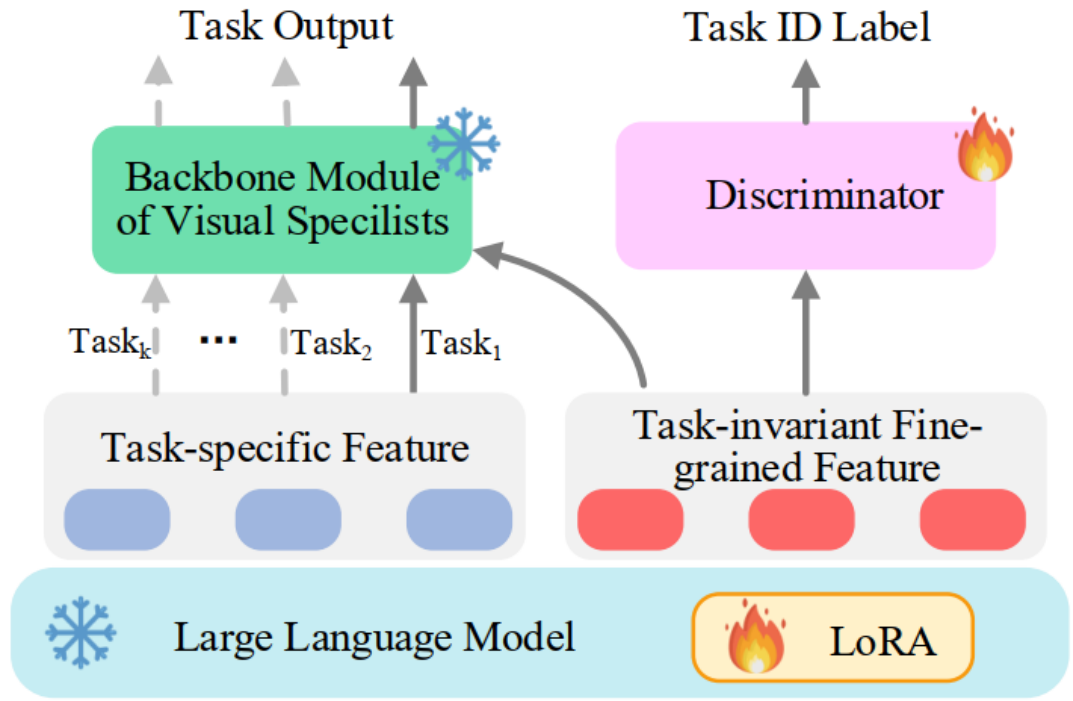


Figure 3: Illustration of the synergy module.

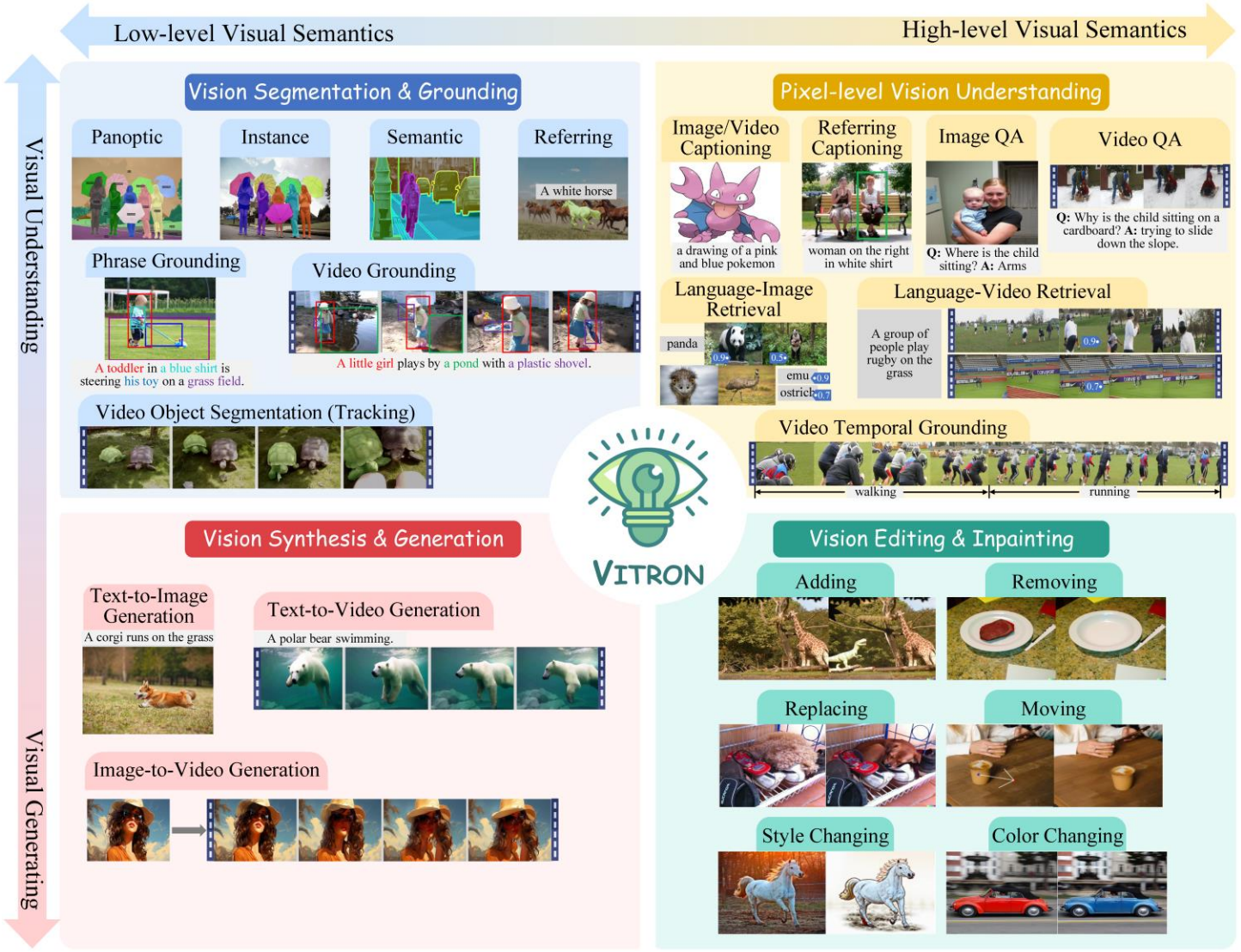
- *decoupling task-specific features from task-invariant features;*
- *then use a third-party **discriminator** to determine the current task based solely on the shared task-invariant feature representation.*

3. Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

- Main Evaluation

4 vision task groups, covering 12 tasks across 22 datasets



3. Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

- Main Evaluation

Method	FID (↓)
LAVT [120]	72.7
GRES [61]	73.8
LISA [46]	74.1
NExT-Chat [126]	74.7
VITRON	**75.5**
w/o syng.	-2.4

Method	FID (↓)	CLIPSIM (↑)
LAVT [120]	72.7	62.1
GRES [61]	73.8	66.0
LISA [46]	74.1	62.4
NExT-Chat [126]	74.7	65.1
VITRON	**75.5**	**66.7**
w/o syng.	-2.4	-1.7

Method	RefCOCO [40]			RefCOCO+ [123]			RefCOCOg [68]	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
LAVT [120]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [61]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
LISA [46]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
NExT-Chat [126]	74.7	78.9	69.5	65.1	71.9	56.7	67.0	67.0
VITRON	**75.5**	**79.5**	**72.2**	**66.7**	**72.5**	**58.0**	**67.9**	**68.9**
w/o syng.	-2.4	-2.0	-1.9	-1.7	-2.1	-1.5	-1.8	-1.6

Table 2: Results (cIoU) of referring image segmentation. 'w/o syng.': without synergy learning.

3. Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

- Ana

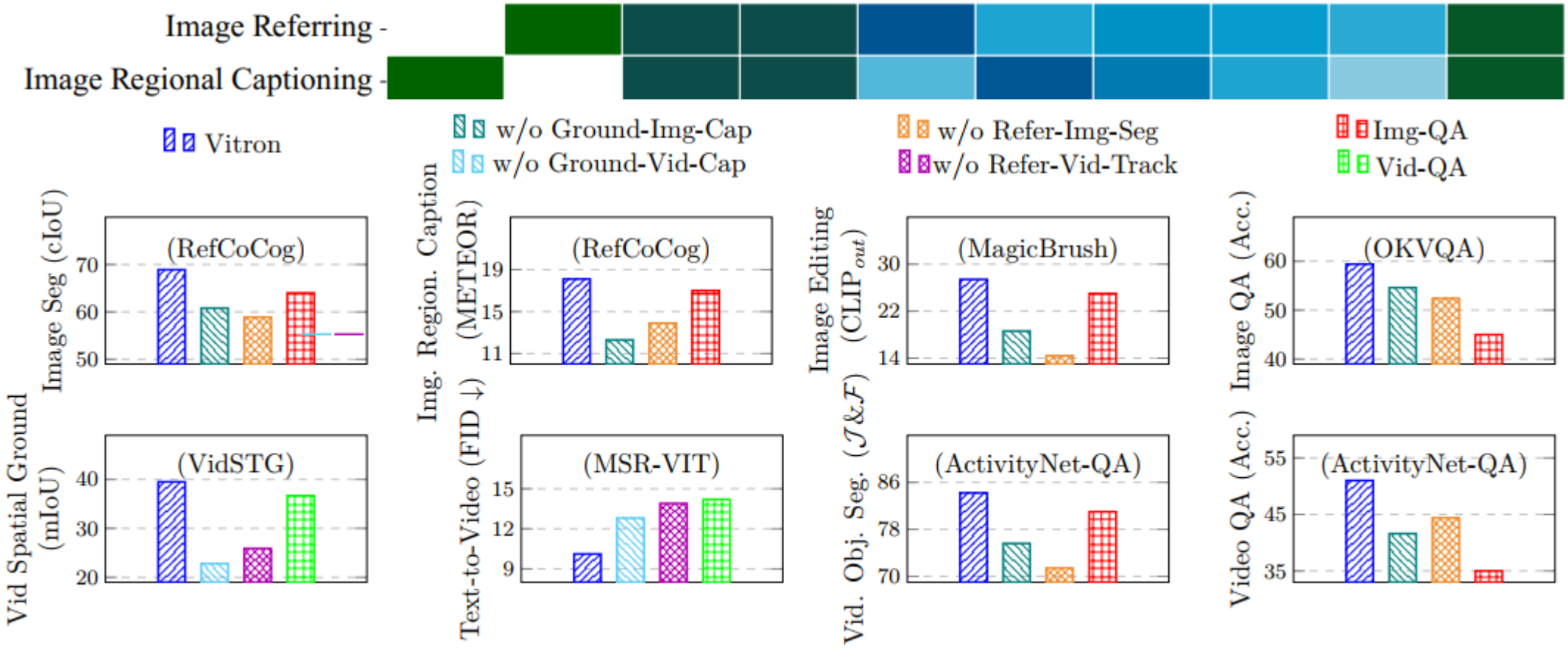


Figure 5: The impact of various fine-grained visual grounding learning strategies.

Figure 4: The influences of using different strategies for message passing.

the more synergistic they are in between.

Image Segmentation

Video Segmentation

Video Understanding

Video Editing

Thanks
Q&A

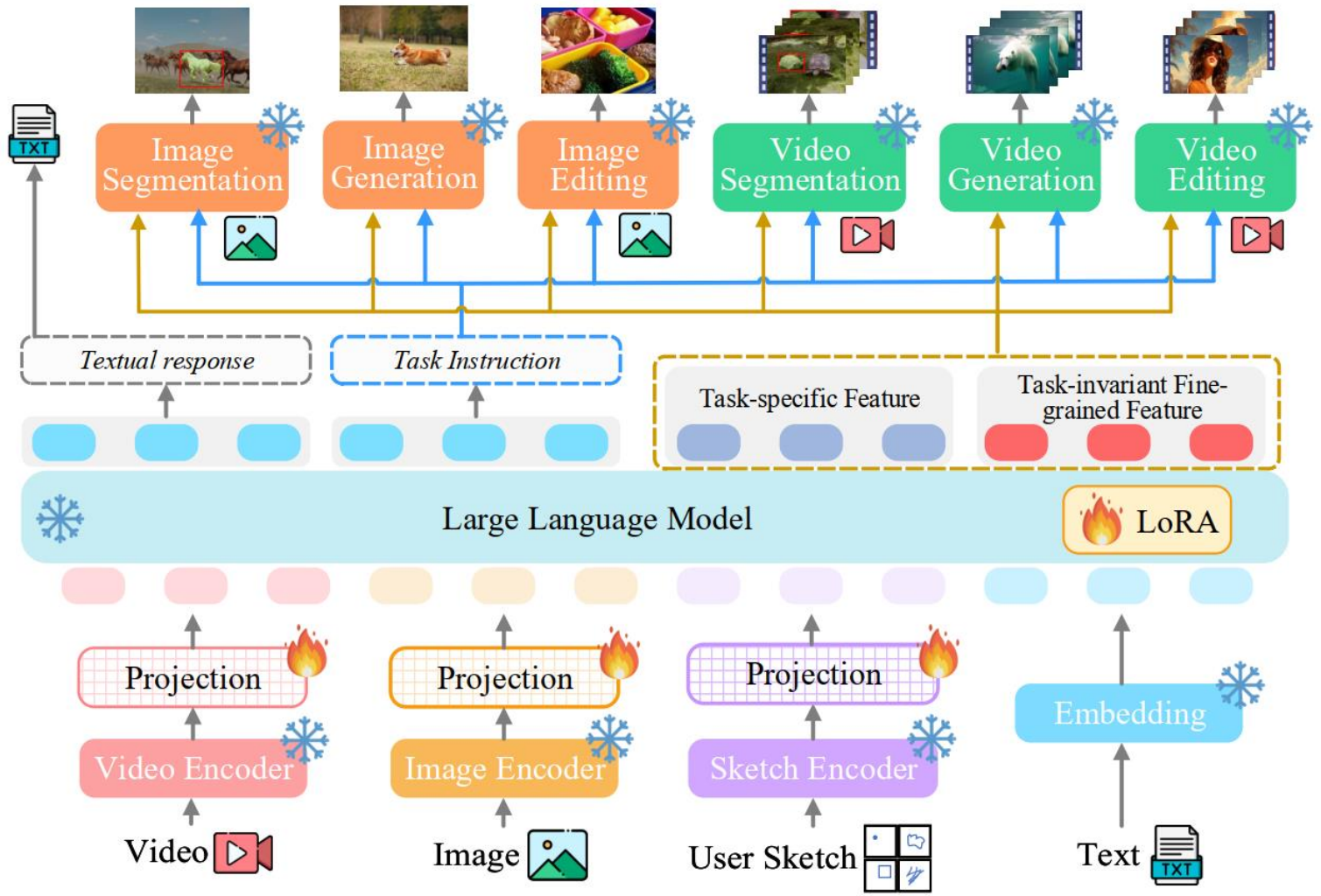


Image Segmentation
Image Generation
Image Editing
Video Segmentation
Video Generation
Video Editing
TXT
Textual response
Task Instruction
Task-specific Feature
Task-invariant Fine-grained Feature
Large Language Model
LoRA
Projection
Projection
Projection
Embedding
Video Encoder
Image Encoder
Sketch Encoder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