

Toward Grounding Anything in Images and Videos

Ming-Hsuan Yang

UC Merced

Google DeepMind

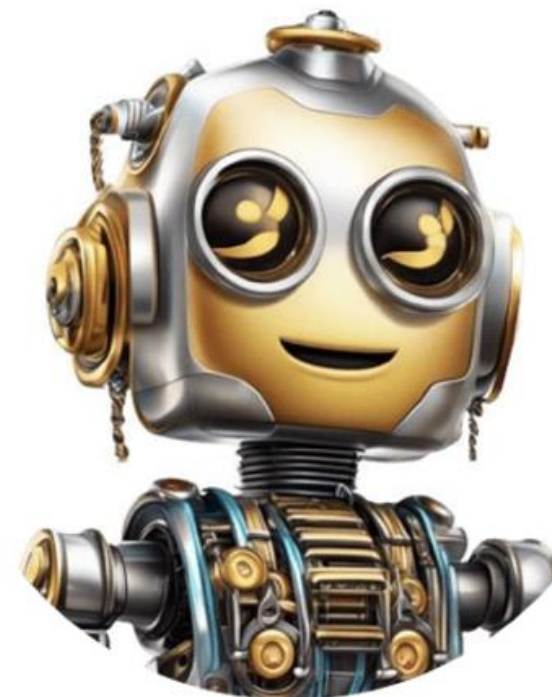
Recent Findings Related to Visual Grounding

- High-Quality Entity Segmentation, ICCV 2023
- AIMS: All-Inclusive Multi-Level Segmentation, NeurIPS 2023
- [GLaMM: Pixel Grounding Large Multimodal Model, CVPR 2024](#)
- VideoGrounding-DINIO: Towards Open-Vocabulary Spatio-Temporal Video Grounding, CVPR 2024
- [VideoPoet: A Large Language Model for Zero-Shot Video Generation, ICML 2024](#)
- [Distilling Spectral Graph for Object Context Aware Open-Vocabulary Semantic Segmentation, CVPR 2025](#)
- OmnixR: Evaluation Omni-Modality Language Models on Reasoning across Modalities, ICLR 2025
- RMP-SAM: Towards Real-Time Multi-Purpose Segment Anything, ICLR 2025
- [Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Video](#)
- Kitten: A Knowledge-Intensive Evaluation of Image Generation on Visual Entities

GLaMM: Pixel Grounding Large Multimodal Model

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, Fahad Shahbaz Khan

CVPR 2024

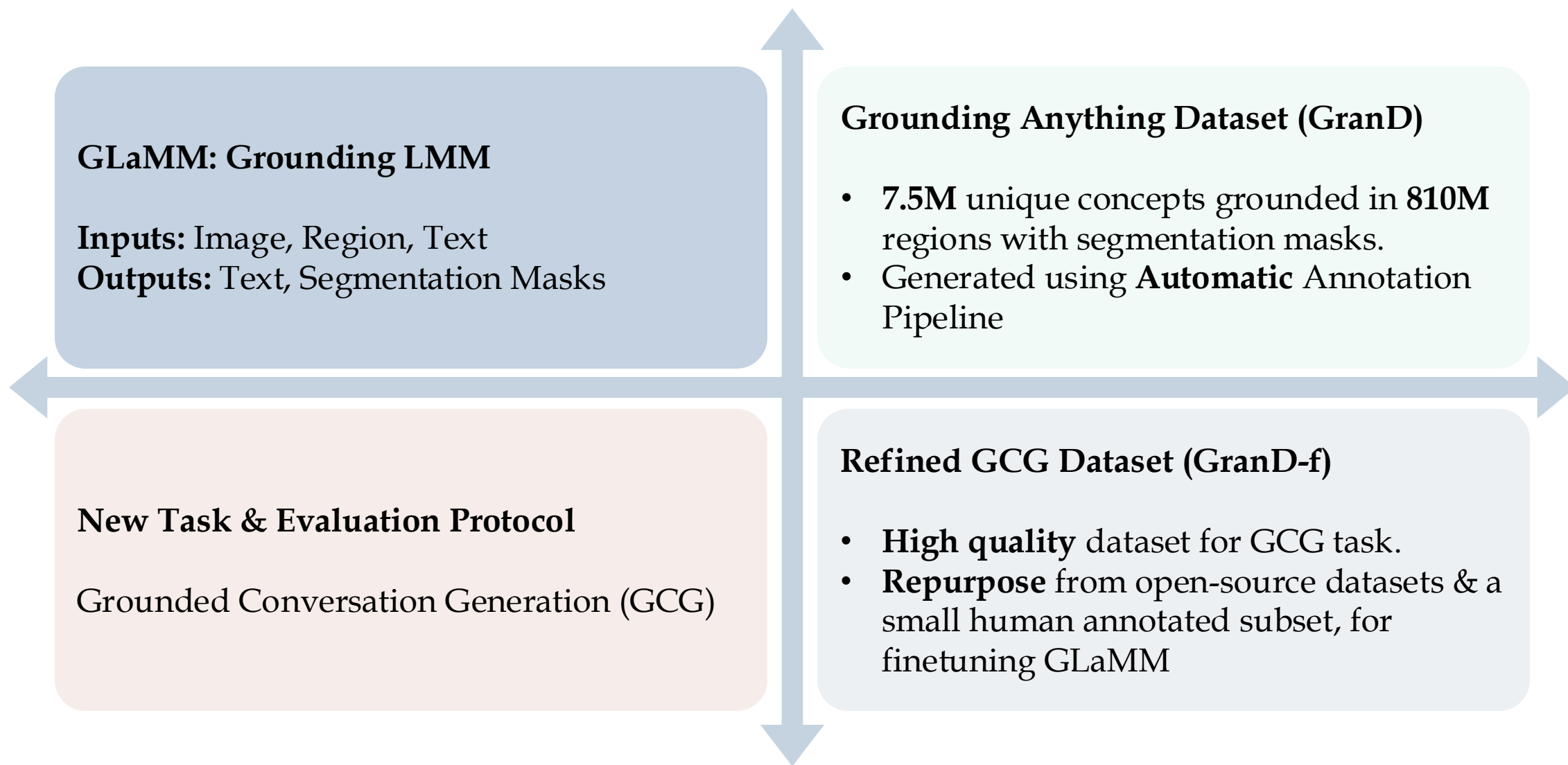


Large Multimodal Models for Visual Understanding

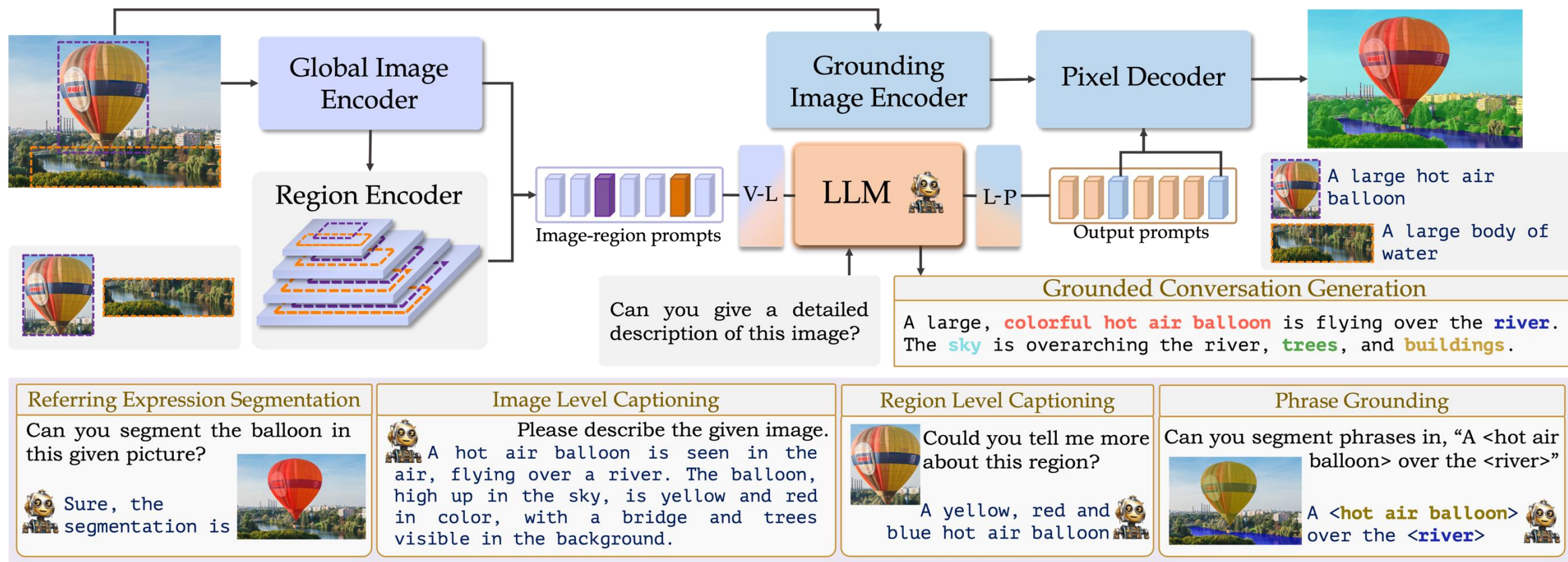
Method	Image	Input / Output		Region Enc. / Dec.	Pixel-Wise Grounding	Multi-turn Conversation	End-End Model	
		Region	Multi-Region					
MM-REACT (arXiv-23) [40]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✗	Do not handle regions
LLaVA (NeurIPS-23) [22]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓	
miniGPT4 (arXiv-23) [48]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓	
mPLUG-OWL (arXiv-23) [41]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓	
LLaMA-Adapter v2 (arXiv-23) [6]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓	
Otter (arXiv-23) [17]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✗	✓	Incorporate external vision modules
Instruct-BLIP (arXiv-23) [5]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓	
InternGPT (arXiv-23) [24]	✓	✓ / ✗	✗ / ✗	✗ / ✗	✗	✓	✗	
Bubo-GPT (arXiv-23) [47]	✓	✗ / ✓	✗ / ✓	✗ / ✗	✗	✓	✗	
Vision-LLM (arXiv-23) [33]	✓	✗ / ✓	✗ / ✓	✗ / ✗	✗	✗	✓	
Det-GPT (arXiv-23) [28]	✓	✓ / ✓	✓ / ✓	✗ / ✗	✗	✓	✓	Rely exclusively on LLMs for region understanding
Shikra (arXiv-23) [4]	✓	✓ / ✓	✗ / ✗	✗ / ✗	✗	✗	✓	
Kosmos-2 (arXiv-23) [27]	✓	✓ / ✓	✓ / ✓	✗ / ✗	✗	✗	✓	
GPT4RoI (arXiv-23) [46]	✓	✓ / ✗	✓ / ✗	✓ / ✗	✗	✓	✓	Combined with specialized vision modules with LLMs
ASM (arXiv-23) [34]	✓	✓ / ✗	✗ / ✗	✓ / ✗	✗	✗	✓	
LISA (arXiv-23) [16]	✓	✗ / ✓	✗ / ✗	✗ / ✓	✓	✗	✓	
GLaMM (ours)	✓	✓ / ✓	✓ / ✓	✓ / ✓	✓	✓	✓	

GLaMM: comprehensive region understanding, pixel-wise grounding, conversational capabilities, and an end-to-end training

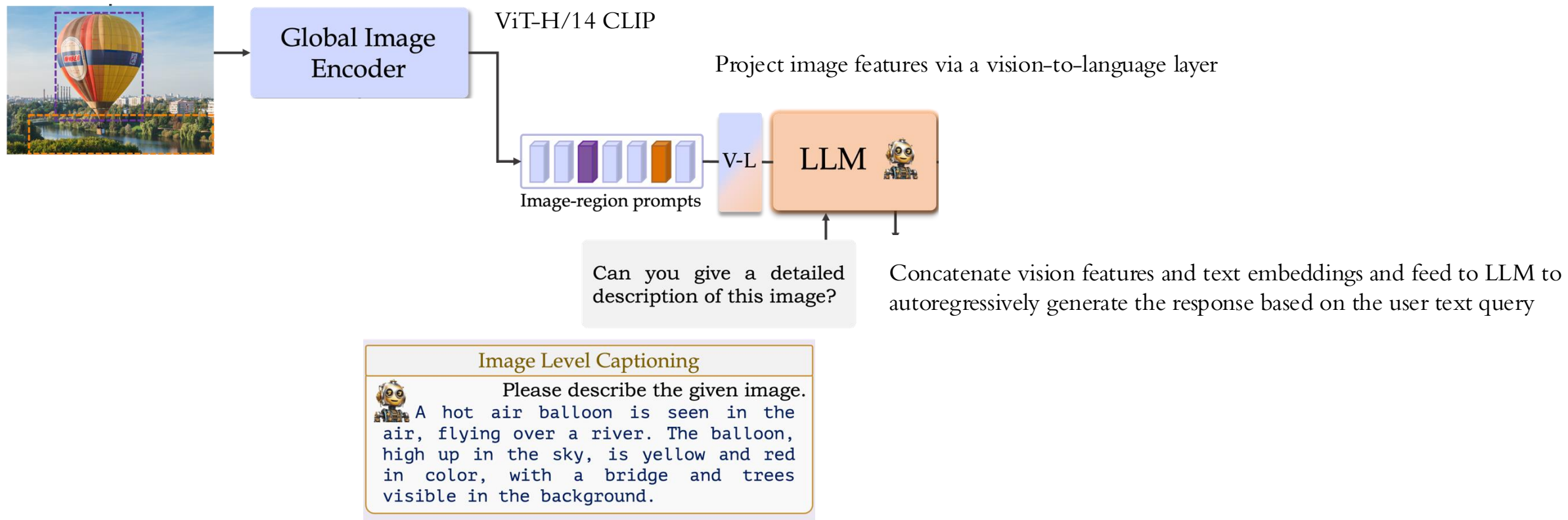
GLaMM: Contributions



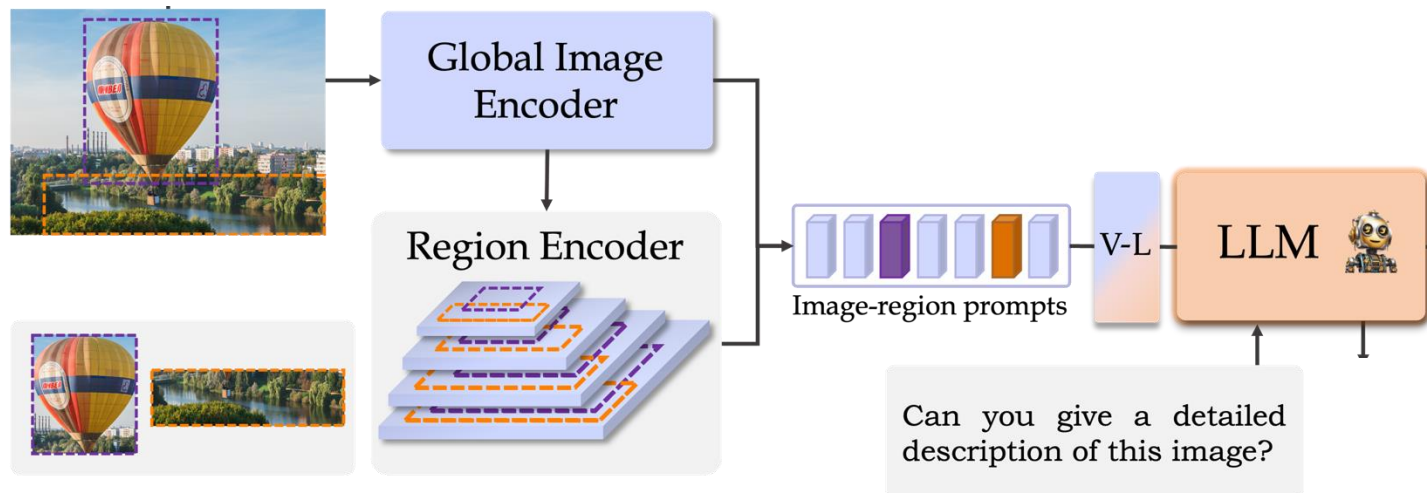
GLaMM: Architecture



GLaMM: Image Level Captioning / QA



GLaMM: Region Level Captioning



ROIAlign hierarchical features -> 14x14 feature map
Combine these features to have ROI representation
Use a special <bbox> to represent a box with ROI features

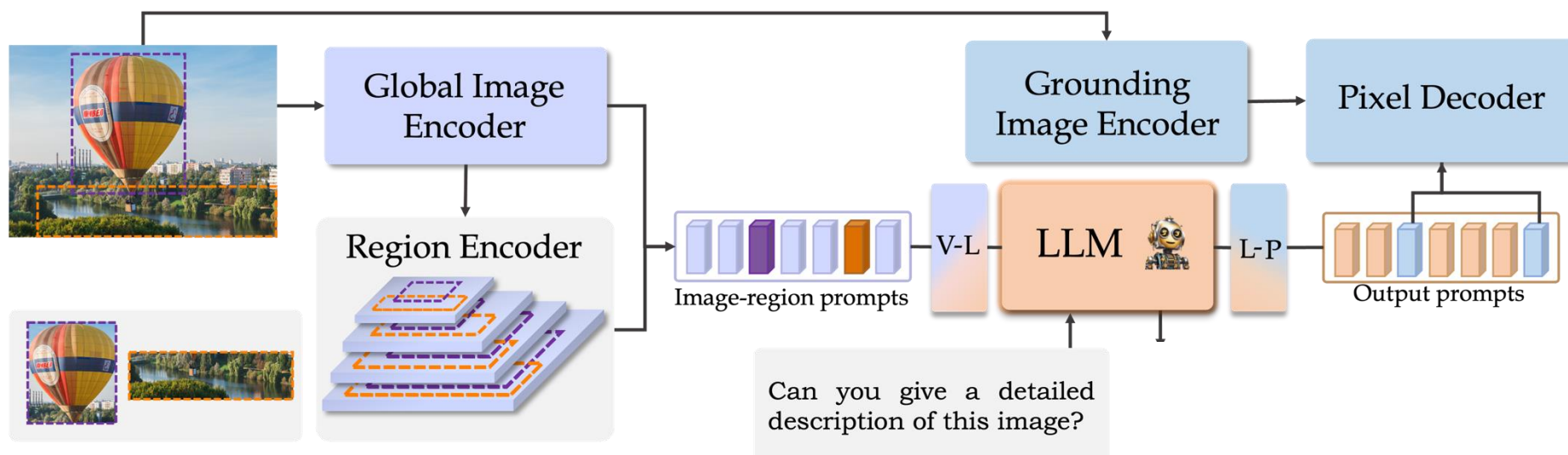
Region Level Captioning



Could you tell me more about this region?

A yellow, red and blue hot air balloon

GLaMM: Referring Expression Segmentation



Referring Expression Segmentation

Can you segment the balloon in this given picture?



Sure, the segmentation is



Grounding image encoder is initialized with SAM encoder

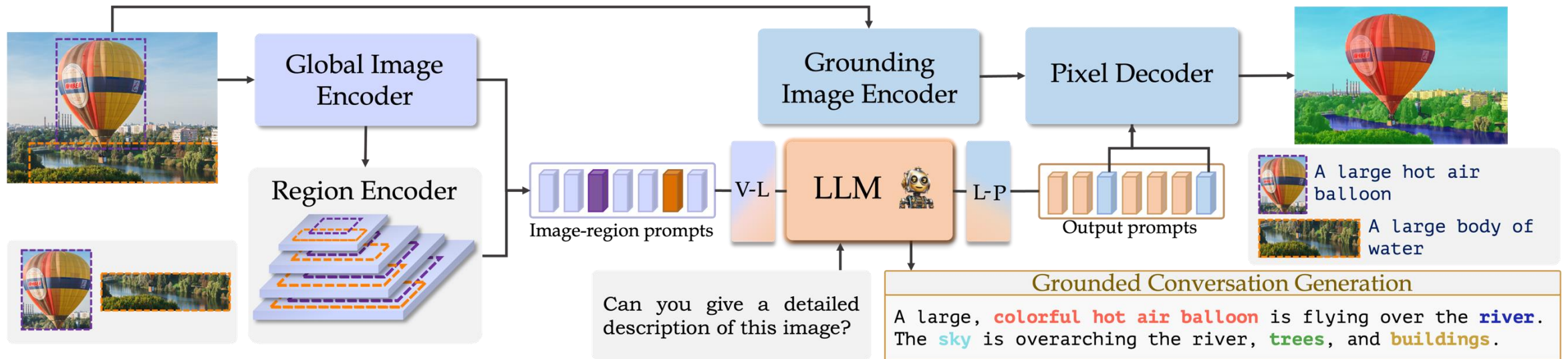
For pixel-level grounding, augment the model with special <SEG> token

Prompt "Please segment the 'man in red' in the given image," triggers to generate <SEG>

L-P: transform last layer embedding of <SEG> into decoder feature space

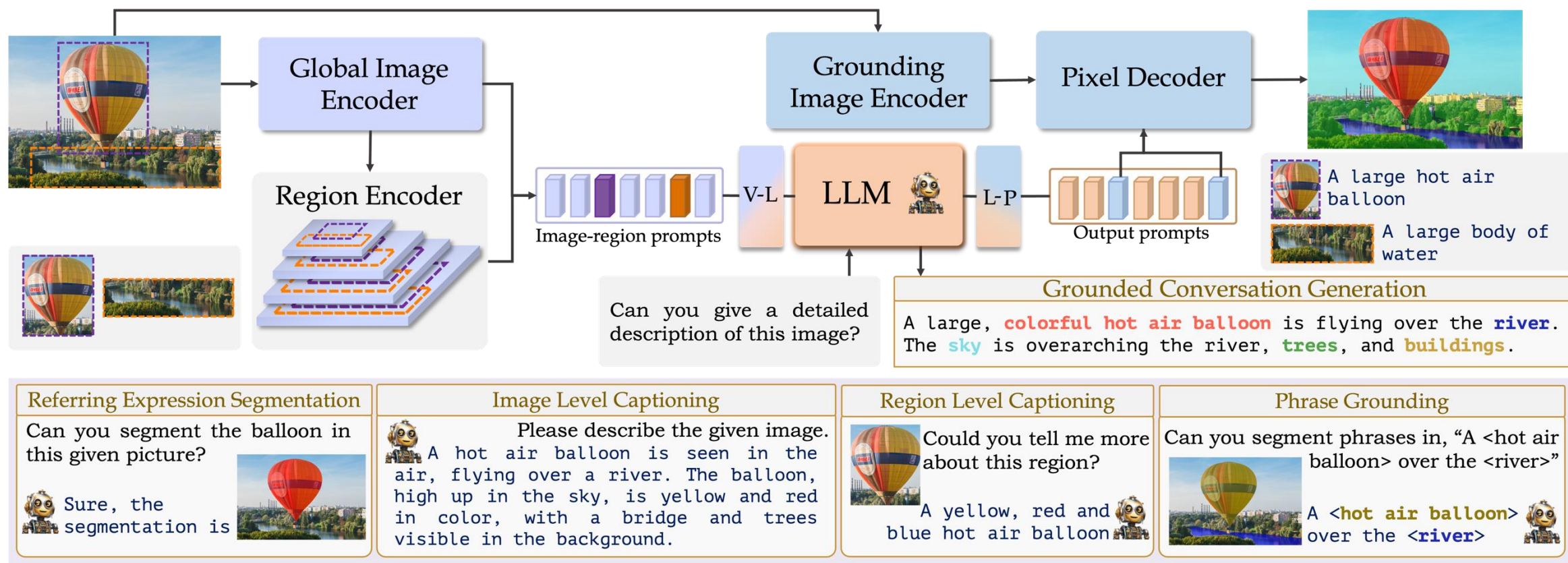
Pixel decoder takes respective tokens to generate masks

GLaMM: Grounded Conversation Generation



Objective: construct image-level captions with specific phrases tied to corresponding segmentation masks

GLaMM: Unifies Understanding and Grounding



Insight: One task help the others (i.e., joint training is useful)
See VideoPoet, ICML 2024

Automatic Annotation Pipeline of GranD

11M SAM images, 810M regions, 7.5M unique concepts, 84M referring expressions, 22M grounded short captions, 11M densely grounded captions

Objects and Attributes

- 1 dog, pub dog, a brown and white dog
- 2 dog collar, black color, chain collar
- 3 bell, cowbell
- 4 steps, stairs, the steps of a building
- 5 sack, a large white bag with black writing

Relationships and Landmarks

A dog sitting on the steps

A large brown dog wearing a chain collar

Cowbell attached to dog collar

Landmarks: Outdoor - Urban Landscape



Scene Graph



Dense Grounded Caption

A large brown dog is sitting on the steps of a building. It is wearing a black chain dog collar. The collar has a cowbell attached to it. There is a bag in the background with black writings on it.

Extra Context

Dogs, especially pugs and bulldogs, have been a part of human families for thousands of years, serving as loyal companions. They have been bred for specific traits, making them popular pets. Dogs have been trained for various tasks, including assisting people with disabilities and serving as search and rescue animals. Dog collars, often bearing identification tags, are essential for keeping pets safe and ensuring they can be returned home if lost. Cowbells, once used to signal the arrival of a cow, have been repurposed as dog collars, providing a distinct sound to help locate a dog if it wanders off. In outdoor urban landscape, dogs are often found sitting on steps, as they may choose to rest in spots that offer a good view of their surroundings.

Level-1

Object localization and attributes

- Image Tagging and Object Detection
- Open Vocabulary Detection
- Region Attribute Detection

Level-2

Relationships

- Short Captions and Phrase extraction
- Grounding expression
- Landmarks

Level-3

Scene Graph & Dense Captioning

- Hierarchical Scene Graph
- In-context Learning with LLM
- Verification Pipeline

Level-4

Extra Contextual Insights

- Landmark Details
- History and Background
- Precautionary Measures

GranD: Data Engine (Level # 1)

Level # 1

Object Localization and Attributes

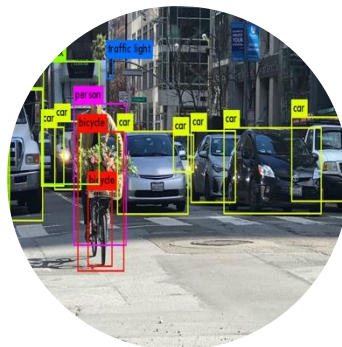
- Image tagging and object detection.
- Open-vocabulary detection.
- Region attribute detection.
- Depth estimation.

GranD: Data Engine (Level # 1)



family, ski slope, mountain,
skier, hill, ski, people,
snow, pose, stand, snowy

Image Tagging with
RAM and **Tag2Text**



Object Detection with **Co-DETR** & **EVA**



Open-Vocabulary Det
with **POMP** and **OWL-ViT**

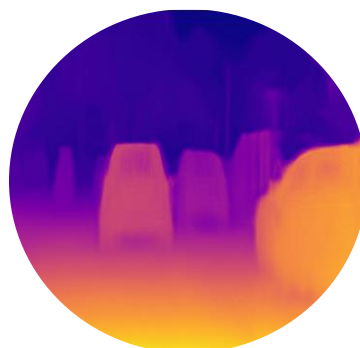


GT-1: a cat holding a
toothbrush in its mouth
GT-2: a cat chewing on a
packaged pink toothbrush
M²: a cat laying on top of a
pair of scissors
GRIT: a cat with a toothbrush
in its mouth on

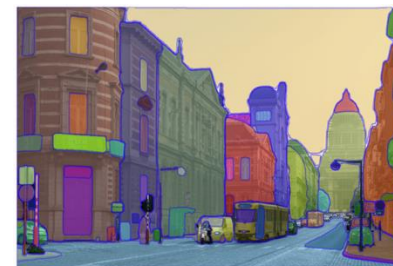


GT-1: the boy is playing video
games in his bedroom
GT-2: a young man is sitting in
a chair playing a video game
M²: a young man sitting in a
chair holding a wii remote
GRIT: a man sitting in a chair
playing a video game

Attribute Detection and
Grounding with **GRiT** &
GPT4RoI

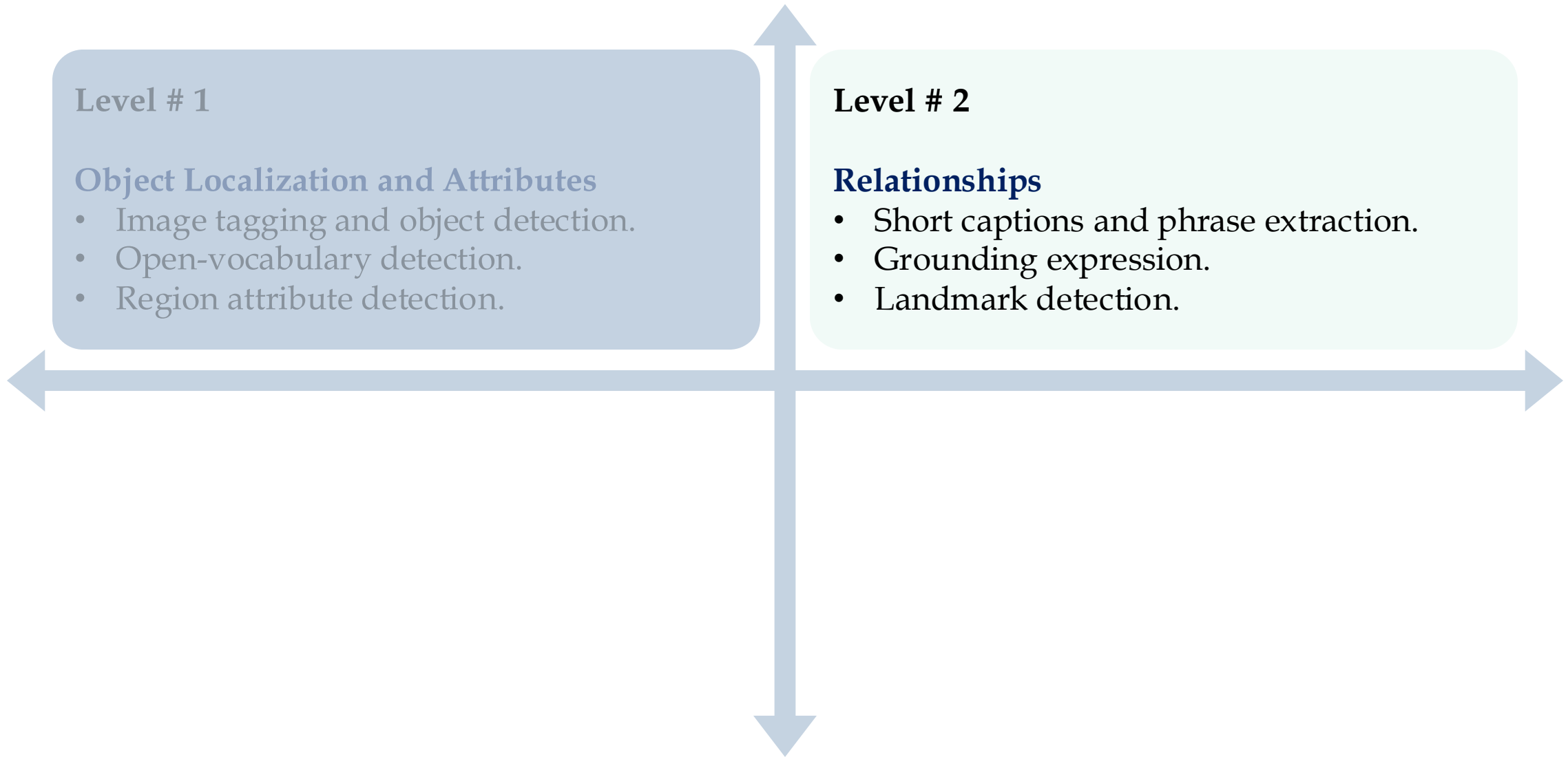


Depth Estimation with
MiDAS



Segmentation with **SAM**

GranD: Data Engine (Level # 2)



GranD: Data Engine (Level # 2)

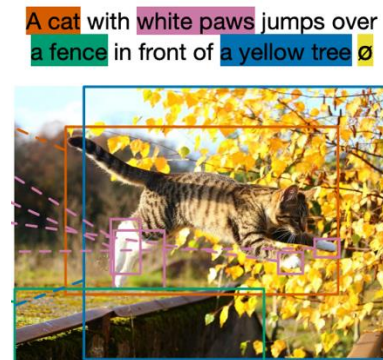


Image Captioning
with LLaVA



A digital
thermometer
resting on a
wooden table,
showing 38.5
degrees Celsius.

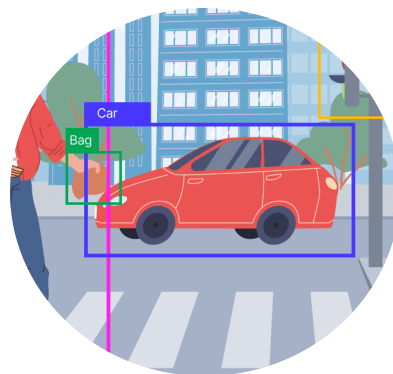
Image Captioning
with BLIP-2



Caption Grounding
with MDETR



Landmark Categorization
with LLaVA

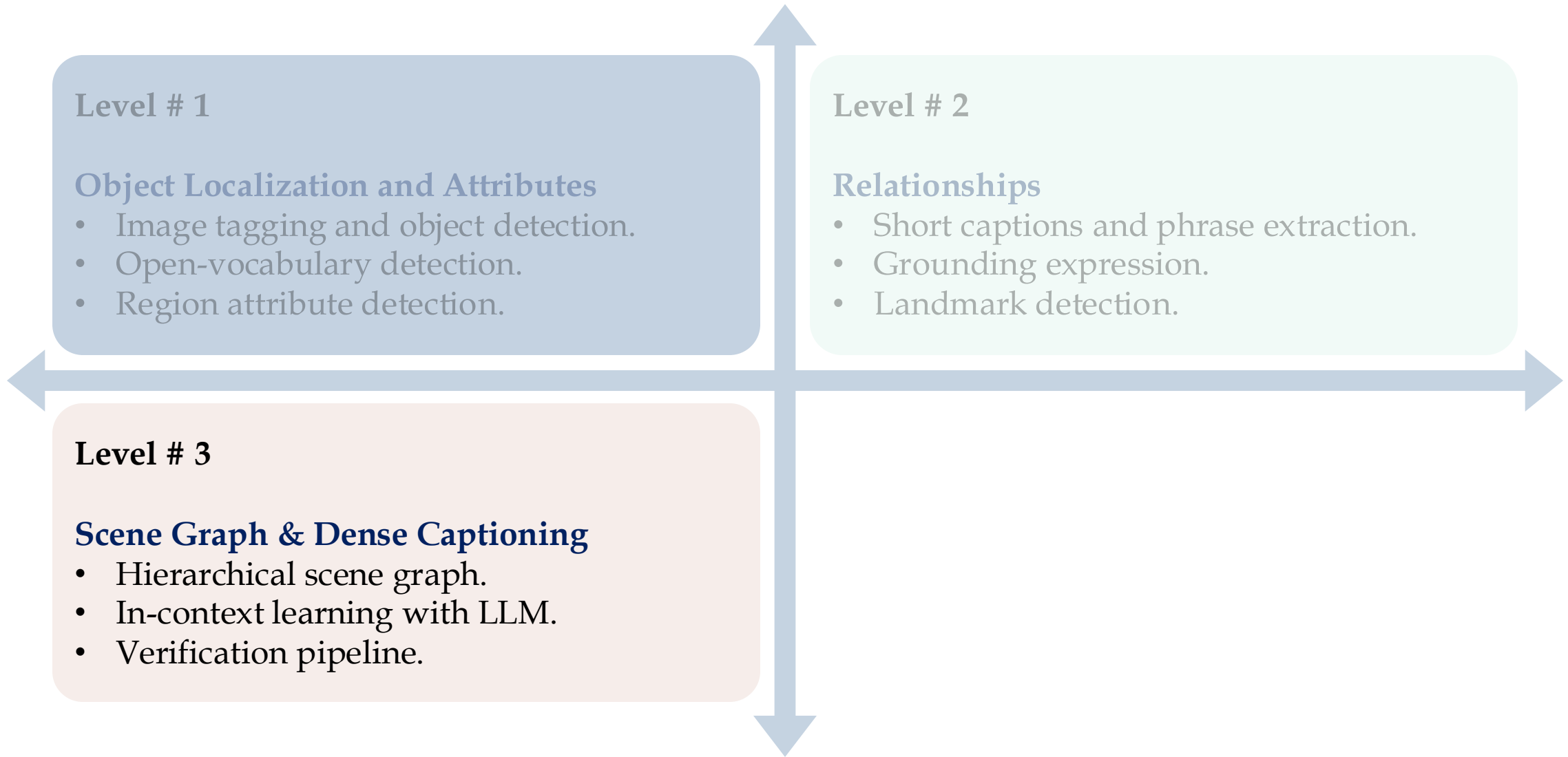


Object Attributes with
GPT4RoI



Label Assignment with
EVA-CLIP

GranD: Data Engine (Level # 3)



GranD: (Level # 3) Scene Graph

Level-1

Image Tagging

Object Detection

Open-Vocabulary Detection

Attribute Detection & Grounding

Open-vocabulary Detection

Depth Estimation

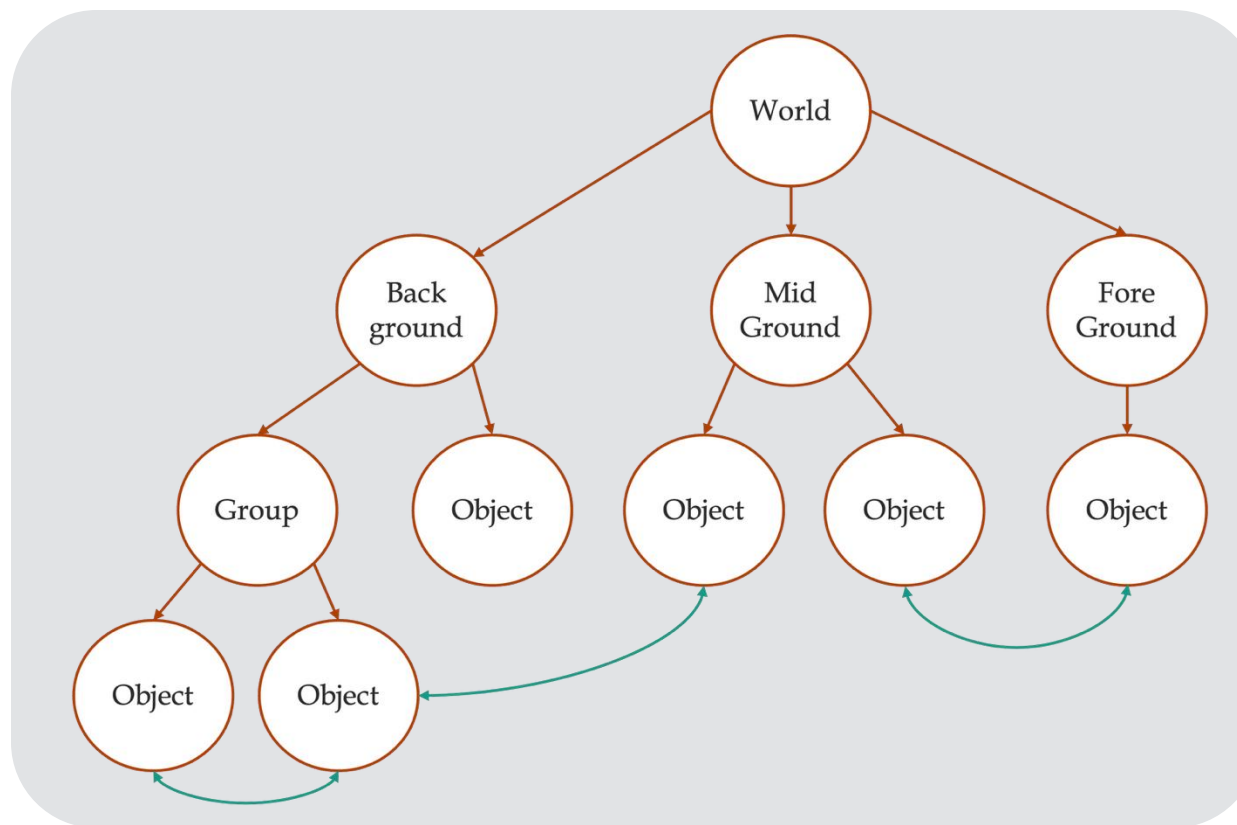
Level-2

Landmark

Scene-level Captioning

Caption Grounding

Object Attribute Assignment



Combine **object attributes and labels** with **relationships and phrases** to form a hierarchical scene graph, which serve as a query for LLM to generate dense image captions with in-context learning

GranD: (Level # 3) Additional Context

Level-1

Image Tagging

Object Detection

Open-Vocabulary Detection

Attribute Detection & Grounding

Open-vocabulary Detection

Depth Estimation

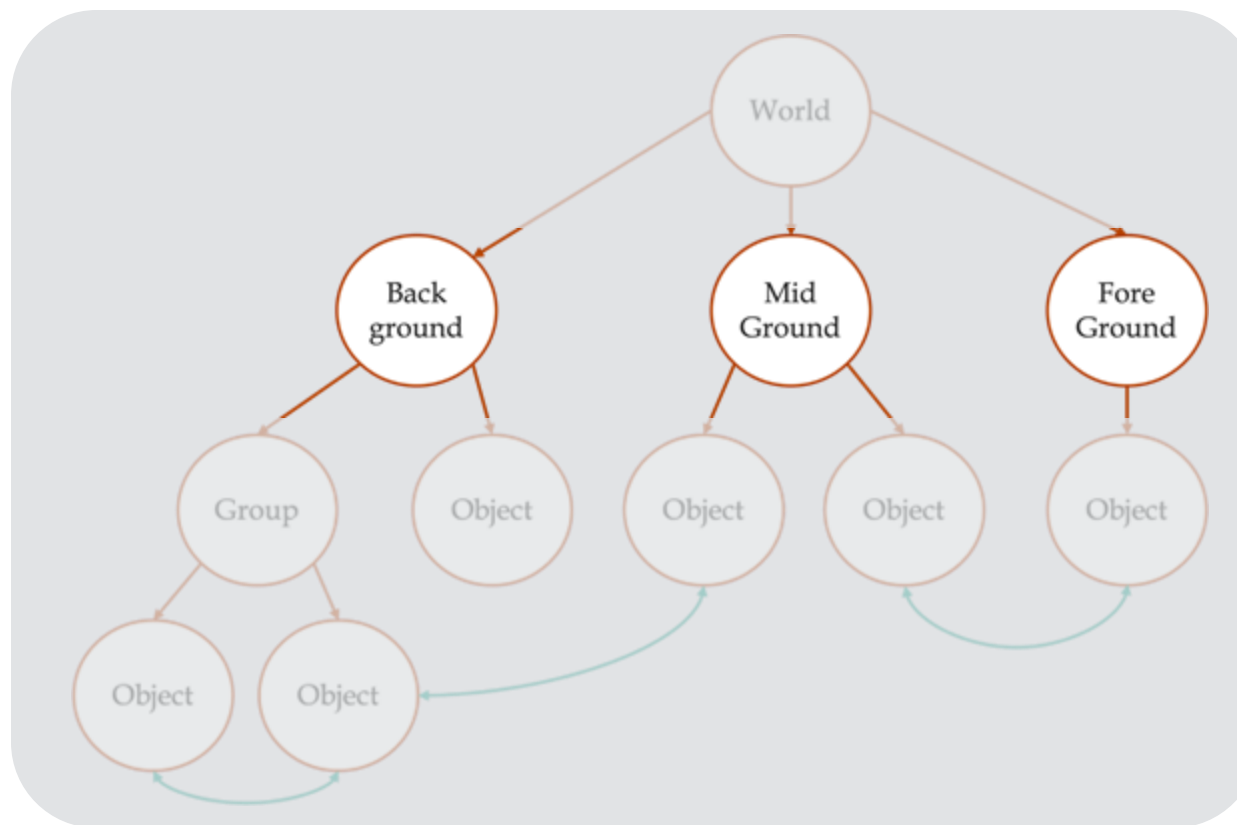
Level-2

Landmark

Scene-level Captioning

Caption Grounding

Object Attribute Assignment



Depth and bounding boxes are used to assign each object to specific layer

GranD: (Level # 3) Incorporate Scene-level Capations

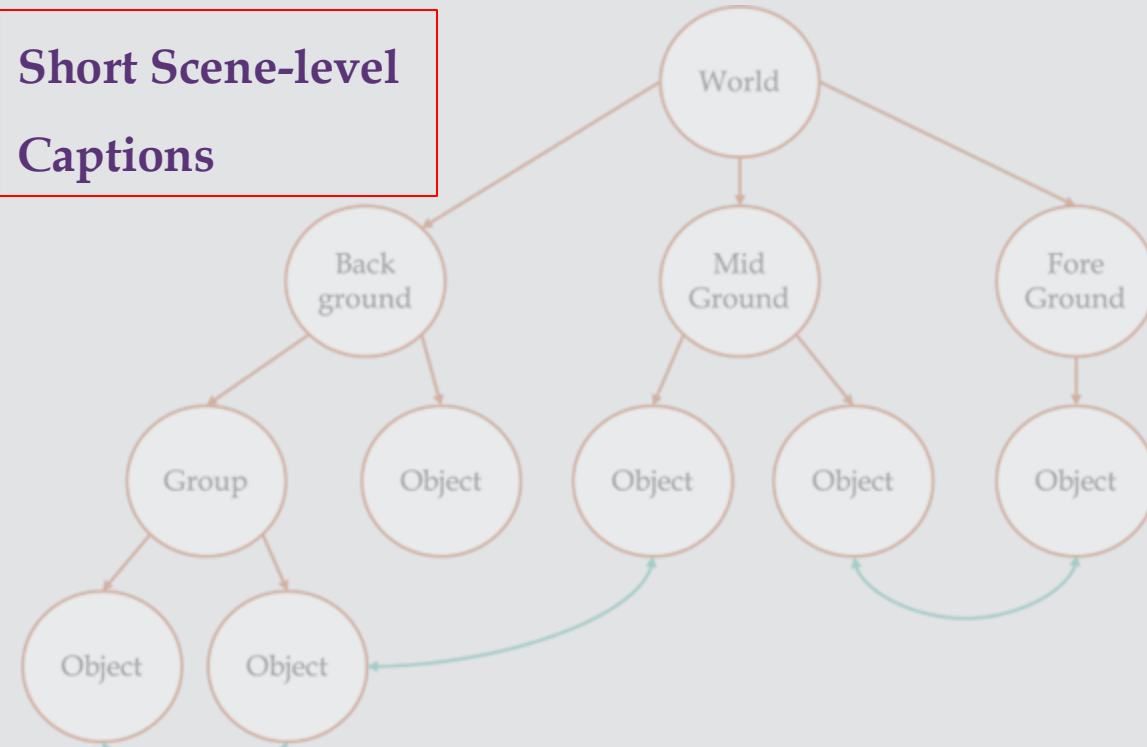
Level-1

Image Tagging
Object Detection
Open-Vocabulary Detection
Attribute Detection & Grounding
Open-vocabulary Detection
Depth Estimation

Level-2

Landmark
Scene-level Captioning
Caption Grounding
Object Attribute Assignment

Short Scene-level Captions



GranD: (Level # 3) Dense Caption Grounding

Dense Caption Grounding

- spaCy for phrase extraction.
- followed by direct comparison with BERT similarity for accurate matching.

Mask Assignment

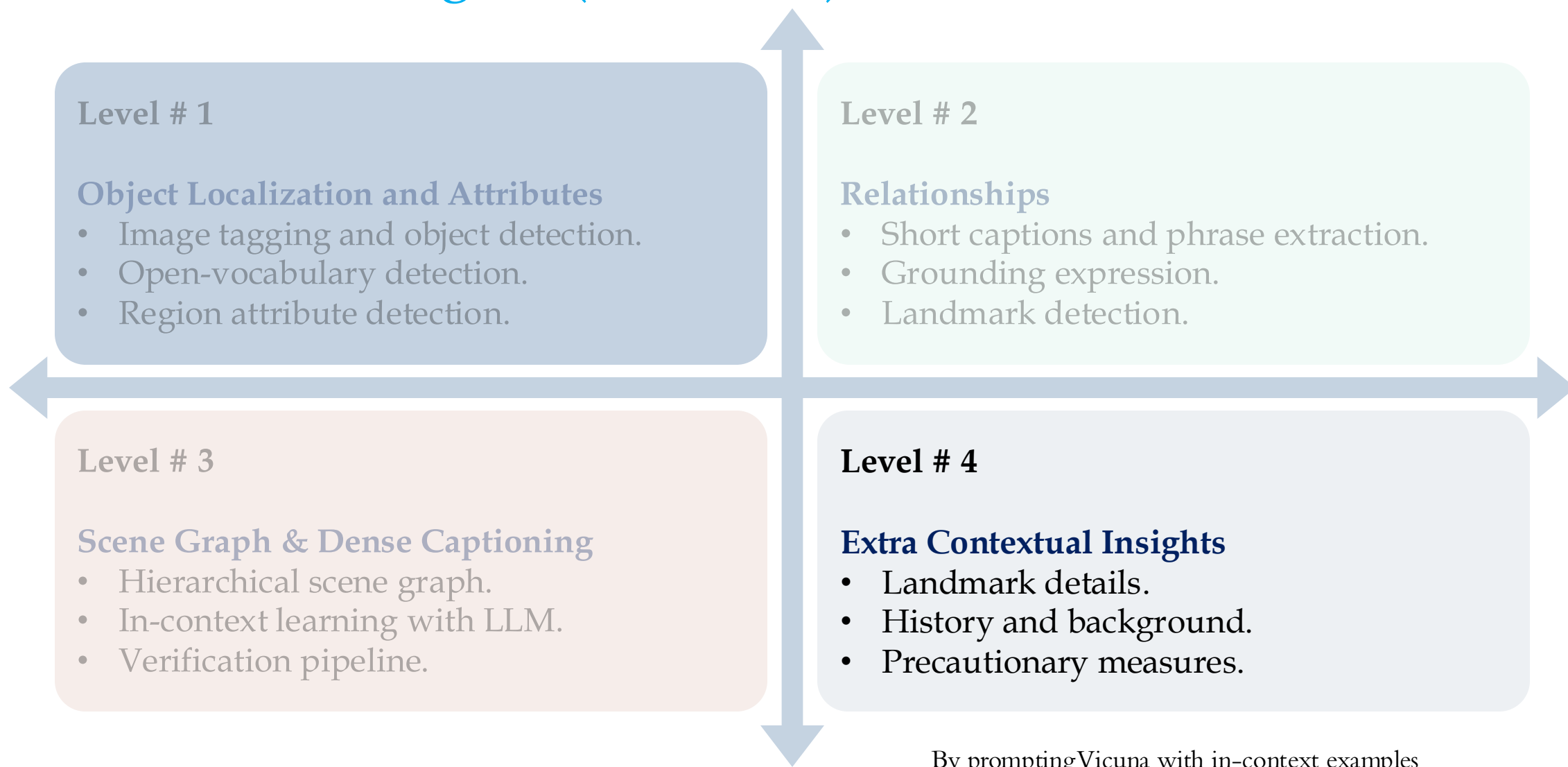
Three-step process

- Match with SAM annotations.
- Compare with EVA-02 detection.
- Use HQ-SAM for the rest.

Enhance fidelity of dense captions using chain-of-thought prompting

Produce a checklist of objects derived from the dense caption expected to be present in image

GranD: Data Engine (Level # 4)



GranD: Example

provide multiple semantic labels and attributes for detected objects, along with the grounded dense caption and additional context.

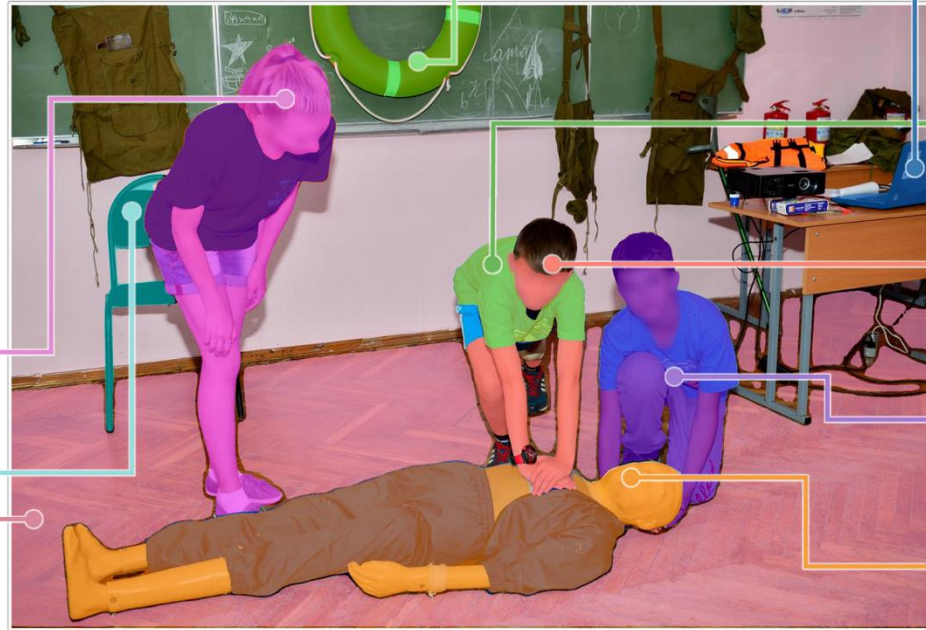


Person, Woman, Girl: A girl wearing orange t-shirt, the girl is bending.

Chair: A black chair with a girl in an orange shirt leaning on it.

Wood floor, Classroom: A first aid course in classroom.

Life buoy, Life belt, Life jacket: A red and white lifesaver.



Laptop, Laptop computer: Laptop on a brown table.

Jersey: A boy wearing blue shirt.

Boy, Person: A boy in blue shirt. The boy is bending.

Boy, Person: A boy wearing blue shirt. The boy is kneeling.

Mannequin, Person: A mannequin with blue clothes lying on the floor.

Dense Caption: A group of children are practicing a first aid course in a classroom. A mannequin wearing a blue shirt, surrounded by young boys, including two wearing blue shirts. There are various items on the table, such as a laptop, projector, and markers. A chair and a blackboard with a red and white lifesaver jacket is also seen in the room. The boys are bending over the mannequin, possibly practicing CPR, while a woman watches.

Additional Context: The image depicts a first aid training session, likely aimed at teaching children basic life-saving techniques like CPR. Given the classroom setting and the presence of a projector and laptop, it may be part of a structured educational program. First aid courses like this are crucial for empowering people to handle emergencies effectively, reducing the severity of injuries or even saving lives. However, improper technique can be potentially dangerous, emphasizing the importance of qualified supervision, in this case, provided by the watching woman. The lifesaver jacket on the blackboard hints at a broader scope of training, possibly including water safety. The brown table with a laptop on it serves as a functional workspace, allowing for remote work or study in a cozy environment.

Grounded Conversation Generation

- Grounded dense caption: specific phrases are linked to corresponding segmentation masks in the image.
- Bridges the gap between **textual** and **visual** understanding
- Fine-grained **visual grounding** alongside natural language captioning.
- Use a **Human-annotated** subset for benchmarking.



Describe the image in detail. Please output interleaved segmentation mask



The image showcases a large, white **building** with a red **roof**, surrounded by a well-manicured lawn and palm **trees**. The **sky** is visible over the building, the **pavement**, and the **grass**. The grass is also seen extending to the pavement.

GranD-f Dataset Construction

Prompt: You are given five base captions for an image briefly describing the image from different perspective. You are also given a number of relationships between objects in the image. Each relationship consists of [subject, relation/verb, object]. Note that each subject and object follows the format <entity name>-<object number in image>. For example, if there are five persons and two tables, they will be formatted as person-1, person-2,..., person-5 and table-1, table-2.

Provide a concise image caption that straightforwardly describes objects/things visible in the image, using provided relationships. Use the base captions for context understanding, but it is not mandatory to include them verbatim in the final description. Break down the description into shorter sentences. Importantly keep the names of the subject and object unchanged. Break down the description into shorter, clear sentences. Do not add any extra information. Keep it brief.

'person in truck', 'car parked on pavement-merged', 'truck parked on dirt-merged', 'house beside tree-merged', 'dirt-merged beside pavement-merged'



A large commercial truck is parked on a dirt area, with its lights on. A person is sitting inside this truck. Close by the truck, a car is parked on the pavement. There is a house nestled beside a tree that's near the dirt and pavement.

'person attached to dog', 'person on bed', 'dog on bed', 'dog attached to blanket', 'bed in front of wall', 'curtain attached to wall', 'cabinet attached to curtain'



The image depicts a person on a bed, lovingly hugging a dog. Both the person and the dog are on the bed, with the dog also attached to a blanket. Further aspects of the room include a curtain attached to a wall, and a cabinet attached to this curtain.

'person about to hit sports ball', 'person swinging tennis racket', 'person running on playingfield', 'banner attached to fence-merged', 'playingfield beside fence-merged'



A person is running on a playing field situated beside a fence. This individual is about to hit a sports ball, swinging a tennis racket in preparation. Attached to the fence is a banner.

(a) Samples from our GranD_f dataset: Illustrating the repurposing of the OpenPSG dataset for the GCG task.

Contains 214K image-grounded text pairs with 2.5K validation and 5K test samples

1). annotated subset. and 2). derived from open-source datasets (OpenPSG, RefCOCOg, Flickr-30K) using GPT4 with in-context learning

GranD-f Dataset Construction (Cont.)

Prompt: Create a concise image caption that straightforwardly describes things visible in the image, using provided phrases verbatim without altering them in any way. Do not infer actions or relationships not explicitly mentioned in the phrases. Use the base captions for contextual understanding, but it is not mandatory to include them verbatim in the final description. Preserve the exact wording of the provided phrases in the final caption. Break down the description into shorter, clear sentences, but without sacrificing the exact wording of the provided phrases. Construct a simple, coherent, and succinct caption that accurately uses the exact phrases, organizing the information into shorter sentences without assuming or inferring additional actions or relationships.

'a tupperware box filled with fruit', 'a plastic container with sliced radishes, green peppers, cucumbers and a sauce', 'slices of pizza are also visibly kept within a plastic container'



A tupperware box filled with fruit features prominently on the table, along with a plastic container with sliced radishes, green peppers, cucumbers and a sauce. Several slices of pizza are also visibly kept within a plastic container.

'a woman with glasses and a purple shirt is having a meal', 'a man with a mustache wearing a blue shirt with food on his plate', 'the man with the glasses on'



A woman with glasses and a purple shirt is having a meal at a table with a man with a mustache wearing a blue shirt with food on his plate and another man, the man with the glasses on.

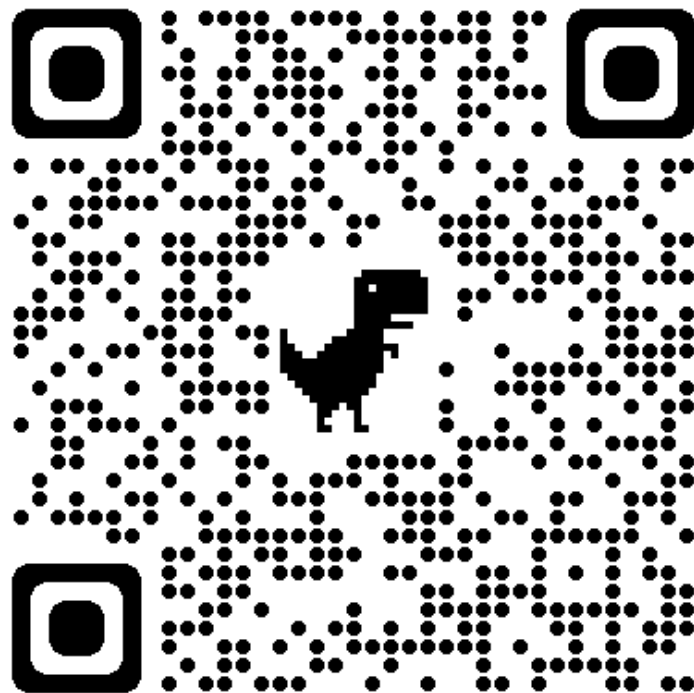
'a laptop computer sitting to the left of a red book', 'laptop on the right side of the table with a scrabble game open'



A laptop computer sitting to the left of a red book. There is also a laptop on the right side of the table with a scrabble game open.

(b) Samples from our GranD_f dataset: Illustrating the repurposing of the RefCOCO-g dataset for the GCG task.

GranD & GranD-f: Dataset Explorer



<https://grounding-anything.com/>

Training

- Use Vicuna LLM 7B model
- Region encoder is similar to GPT4RoI, and grounding image/pixel encoders are similar to LISA
- V-L and L-P layers are 2-leary MLPs
- Training loss
 - Text: auto-regressive cross-entropy loss
 - Segmentation: per-pixel binary cross entropy loss and DICE loss

GCG Benchmark

- **Evaluation Metrics**
 - METEOR, CIDEr for captioning
 - AP50, mIoU for phrase-to-mask
 - Mask Recall for region-specific grounding

Model	Test Set				
	METEOR	CIDEr	AP50	mIoU	Recall
BuboGPT	17.1	3.4	17.5	53.8	27.4
Kosmos-2	15.9	26.8	17.4	56.8	29.1
LISA*	13.1	33.0	25.1	61.6	36.1
GLaMM	15.8	43.5	29.2	65.6	40.8

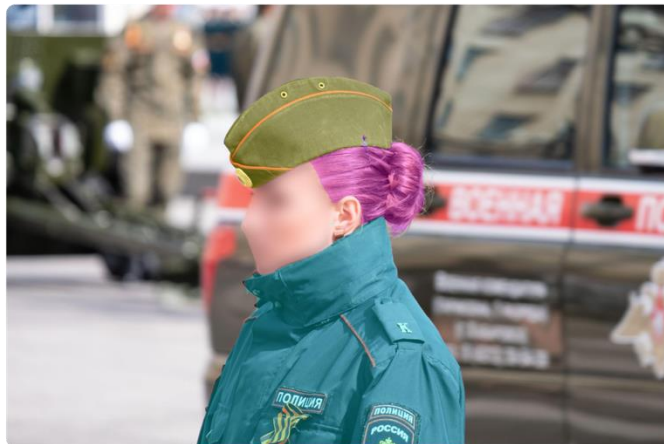
Training

- Use Vicuna LLM with 7B parameters
- V-L and L-P layers implemented with 2 layer MLP and GELU activation as in LLaVA
- Training losses:
 - Auto-regressive cross-entropy for text generation
 - binary cross-entropy loss and DICE for segmentation
- Pretraining on GranD and finetuning on downstream tasks

Qualitative Results



A man and a boy sit on a bench next to an old white car.



A woman in a navy blue jacket and hat with a hair ribbon in her hair.



A soccer player in a red uniform is about to kick the ball while a player in a white uniform is trying to block the shot.



The image features a large, old building with a roof, situated on a grassy field. A tree is also present on the grass. The sky is seen overarching the building.



A person in a black jacket and a straw hat is cooking some food.

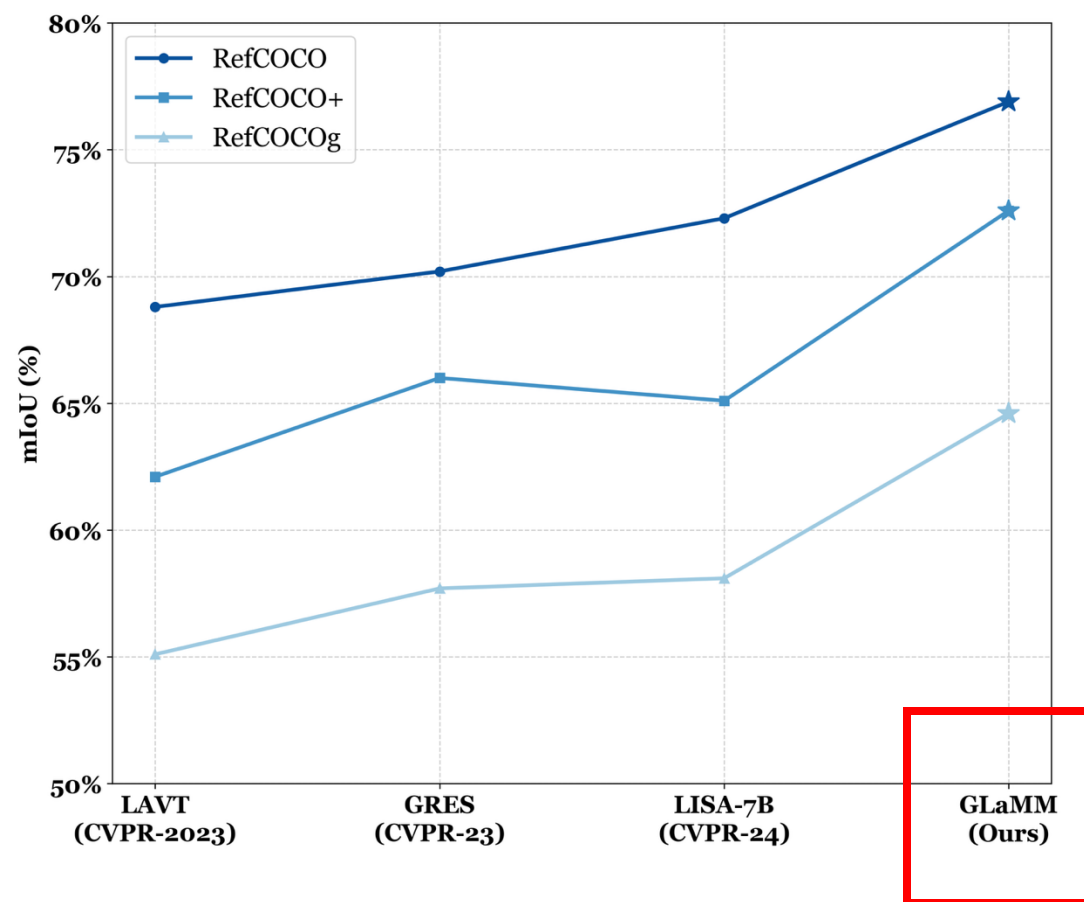


The image showcases a boat sailing on a river. The sky is overarching the boat, bridge, and the buildings.

Referring Expression Segmentation

The model outputs

- “Sure, it is <SEG>”
- <SEG> token is decoded to obtain mask.



Referring Expression Segmentation

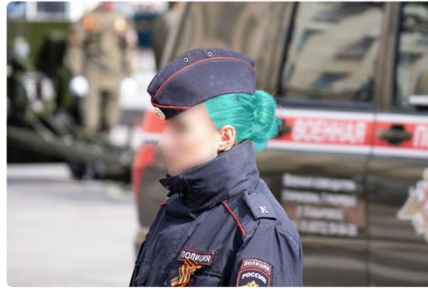
Can you segment Ramos in this image ?



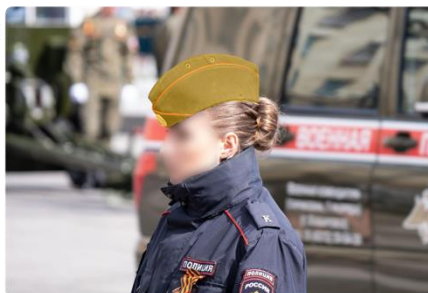
Now can you segment the football ?



Can you segment the woman's hair in this image



Now please segment the hat.



Can you segment what the person is using to ride ?



Now can you please segment the sea.



Region Level Captioning

- Given image and Region bbox.
- Model generates region-level caption.
- Region is encoded via GLaMM region encoder.

Model	refCOCOg		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
GRIT	15.2	71.6	17.1	142
Kosmos-2	14.1	62.3	-	-
GPT4RoI	-	-	17.4	145.2
GLaMM (FT)	16.2	106.0	19.7	180.5

Region Level Captioning

What is she doing in this image ?



She is using her cellphone while crossing the street.

Do you think this is dangerous ?

It could be potentially dangerous for the woman to be using her cellphone while crossing the street. She might not be paying full attention to the traffic and might be more susceptible to accidents.

Can you please describe the region ?



A man sitting on a motorcycle

Why does he need this ?



It is a helmet and he needs it for protection while riding his motorcycle.

What are they doing ?



They are wrestling in the sand.

Can you tell me which country this image is from, given the context ?

This is an image from India.

Emerging Properties Of GLaMM

Although, not trained on multi-turn grounding data, GLaMM can still engage in Multimodal Multi-turn conversation.

Could you provide me with a detailed analysis of this photo? Please output with interleaved segmentation masks for the corresponding parts of the answer.

A woman in a white hat and apron is presenting a bowl to a man in a red shirt.



Can you please describe this region ?



Why do we need so many bowls ?

This is a restaurant setting where the bowls are used to serve different dishes. It is a common practice in some cultures to serve a variety of small dishes or appetizers in bowls to customers.

Can you segment what he is using for cooking ?

Sure, the segmentation result is.



Could you please give me a detailed description of the image? Please respond with interleaved segmentation masks for the corresponding parts of the answer

A man in a fishing boat uses an oar to navigate the water.



What can you tell me about the region indicated in the image?

"A straw Hat"



Why does he need the hat ?

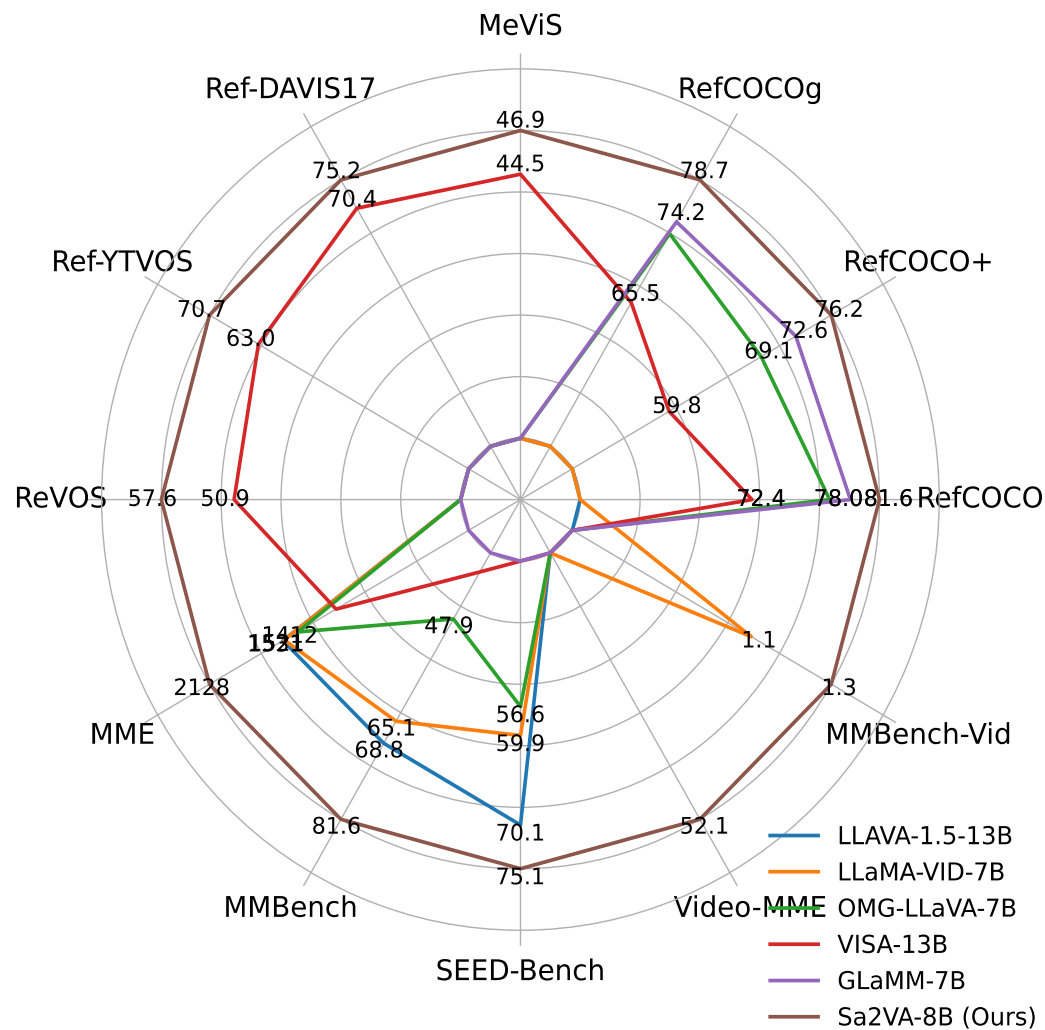
The man is wearing the hat to protect his head from the sun.

Can you segment the water body ?

Sure, here's the segmentation.



Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos



 Please segment the person walking on the street.



 Sure, it is [SEG].

Motivation

Existing models are mainly designed for specific **tasks** (image chat, referring segmentation) or limited to a single **modality** (image or video), our work aims to bridge that gap.



Video Caption by PLLaVA

The video depicts a woman walking down a street at night. She is wearing a black leather jacket and sunglasses, and she is carrying a black purse. The street is wet, suggesting that it has recently rained. There are other people in the background, but they are not the focus of the video. The woman appears to be the main character, and she is walking with purpose. The overall atmosphere of the video is dark and moody.

However, not capable of video object segmentation

Motivation

Existing models are mainly designed for specific **tasks** (image chat, referring segmentation) or limited to a single **modality** (image or video), our work aims to bridge that gap.

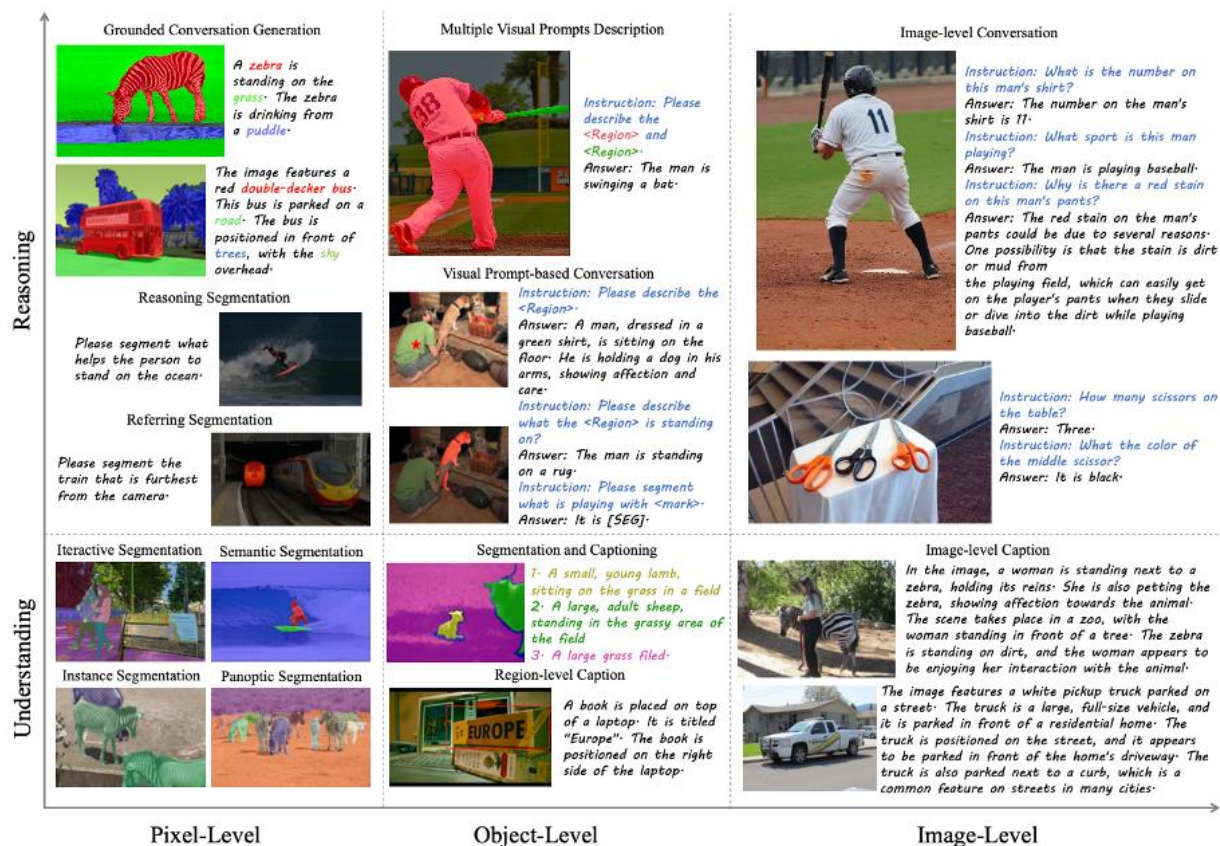


Image Referring Segmentation by OMG-LLaVA

However, conversational performance drop drastically and cannot handle videos

OMG-LLaVA: Bridging Image-level, Object-level, Pixel-level Reasoning and Understanding.

Sa2VA - Motivation

Goal: build a unified model for **dense, grounded understanding** across both **images** and **videos**.
See also VideoPoet, ICML 2024

Table 1. **Comparison of capabilities of different representative models.** Our method supports various tasks and modalities. Benefiting from these interactive features on video, Sa2VA can perform multiple promptable tasks in the video, as shown in Fig. 1 (a) and (b).

[illegible]

Sa2VA – Task Unification

Unify different tasks into a visual, textual, and mask modalities into a [shared token space](#), allowing the language model to generate multi-modal tokens for various tasks.

$$T_o, M_o = LLM(\{I_i, V_i, VP_i\}, T_i).$$

Referring Image / Video

Segmentation:

- - Input: Image / Video, Text
- - Output: Text, Masks

Image / Video Chat

- - Input: Image / Video, Text
- - Output: Text

Visual Prompt Understanding

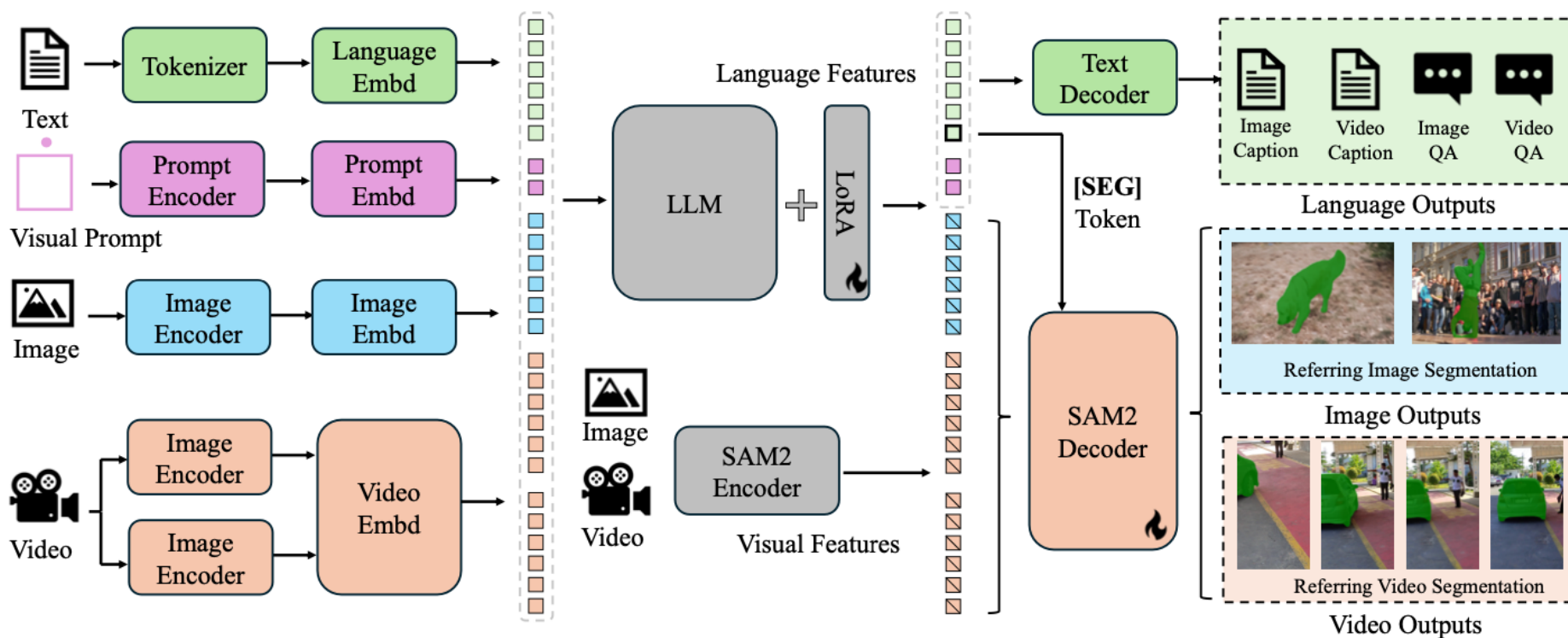
- - Input: Image / Video, Text, Visual Prompt
- - Output: Text, Masks

Sa2VA - Framework

Adopt a decoupled design and use [SEG] token to bridge SAM2 and the MLLM.

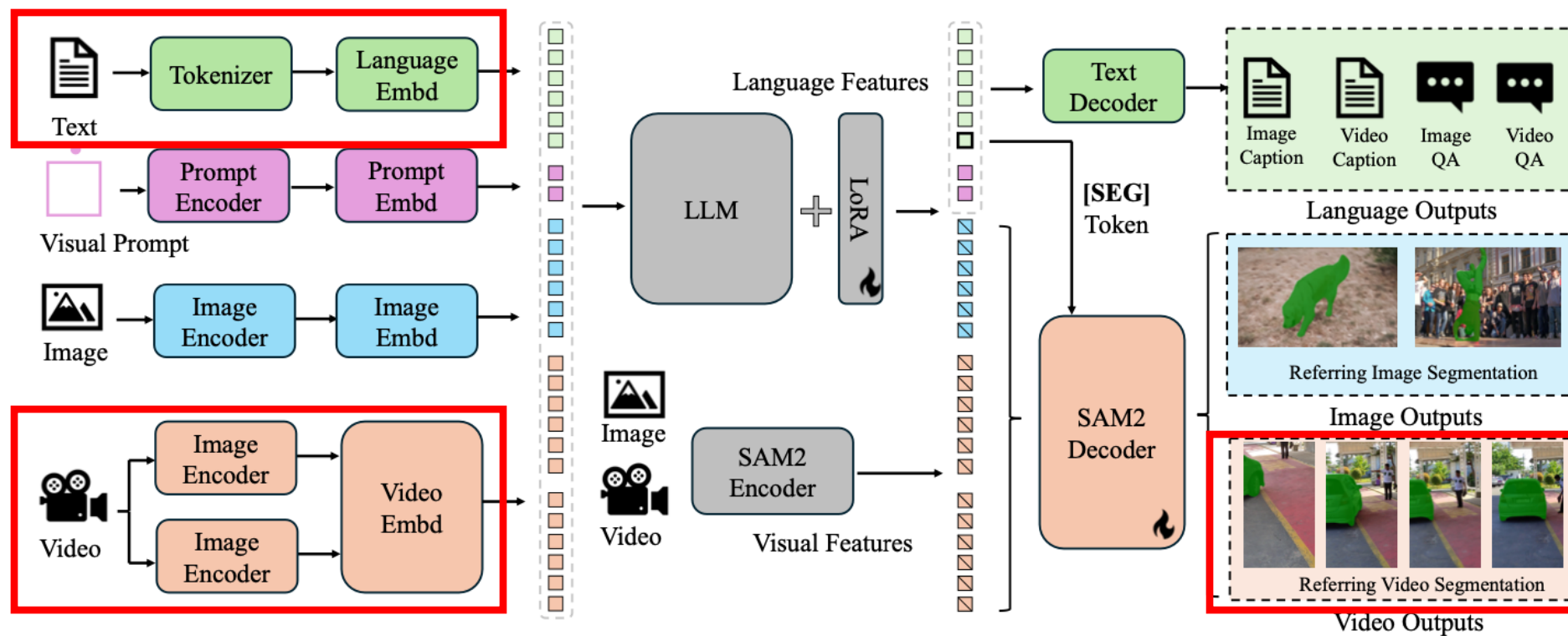
Tuning SAM2 decoder via a special [SEG] token as new prompt

Exploit LLM for multimodal reasoning and SAM2 for segmentation



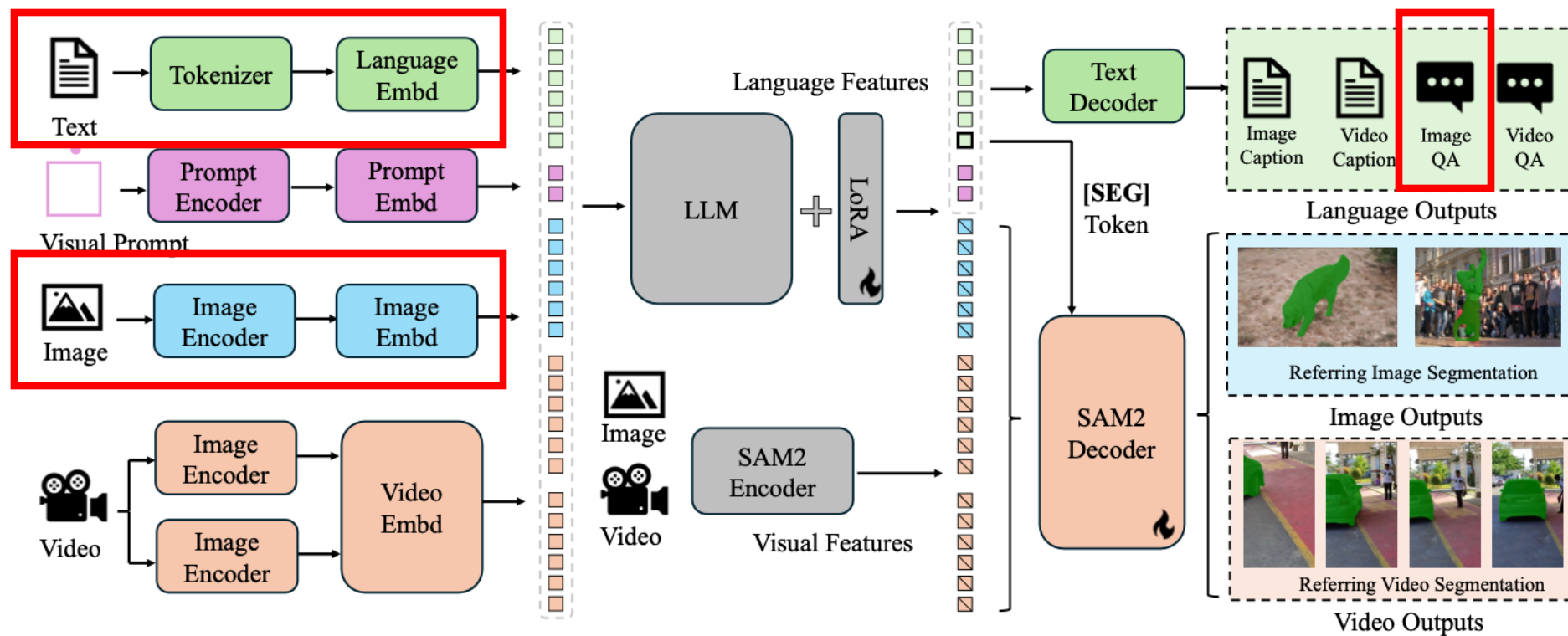
Sa2VA - Method

Video Segmentation



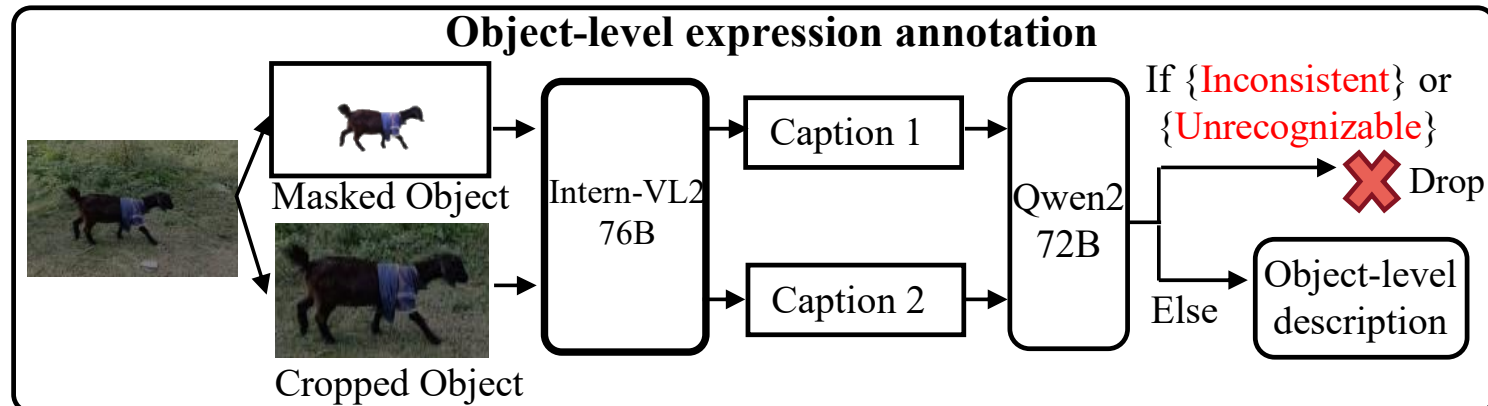
Sa2VA - Method

Image Q&A



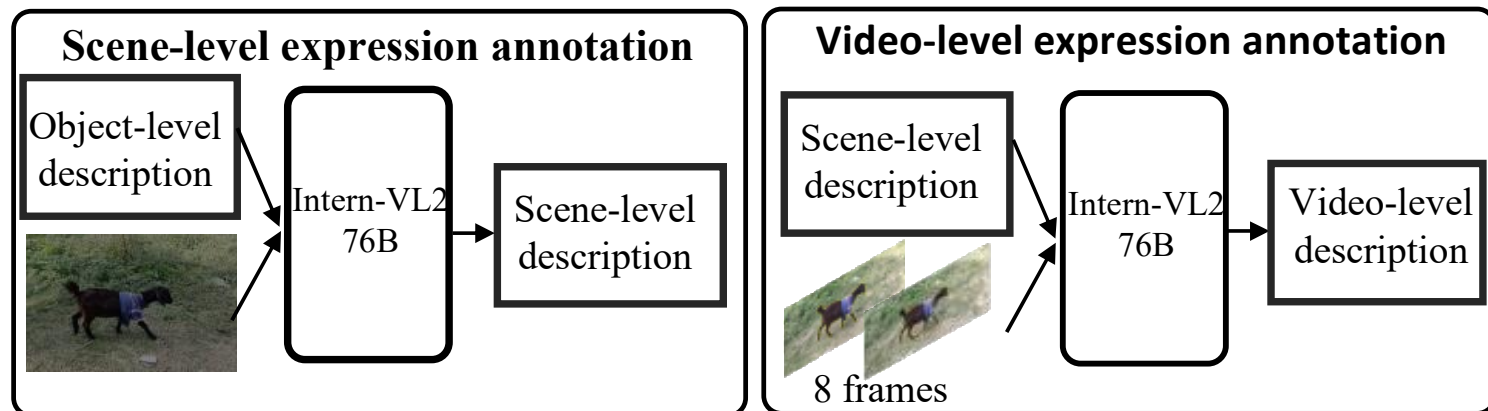
Sa2VA - Method

Data Engine



Introduce Ref-SAV, a large-scale dataset with over 72,000 automatically labeled object expressions in complex videos.

Manually validate 2,000 samples for robust benchmarking.



Final expression: The object is a black goat wearing a blue ... (Object-level) It is moving around in a natural outdoor environment ... (Scene-level) Throughout the video, the goat can be seen walking across ... (Video-level).

Sa2VA: Training and Testing

One for all co-training

$$\mathcal{L}_{instruction} = \mathcal{L}_{text} + \mathcal{L}_{mask}, \quad \mathcal{L}_{mask} = \mathcal{L}_{CE} + \mathcal{L}_{DICE}.$$

VQA: text regression

Segmentation: pixel-wise cross-entropy loss and DICE loss

Supervised one-shot fine-tuning

One for all testing

Encode text prompts, visual prompts, image/video features into tokens for LLM

Output tokens of LLM are decoded text responses and segmentation masks

Image/Video Referring Segmentation

Table 2: Experiment results on image/video referring segmentation benchmarks and image/video chat benchmarks. A / B means the perception (A) and cognition (B) scores, while C(+) means the total score (C = A + B).

Method	Image Segmentation			Video Segmentation				Image Chat			Video Chat		GCG
	RefCOCO [41]	RefCOCO+ [41]	RefCOCOg [102]	MeViS [17]	Ref-DAVIS17 [43]	Ref-YTVOS [78]	ReVOS [96]	MME [23]	MMBench [65]	SEED-Bench [47]	Video-MME [24]	MMBench-Video [21]	GCG [75]
LLaVA-1.5-13B [59]	-	-	-	-	-	-	-	1531(+)	68.8	70.1	-	-	-
Video-LLaVA-7B [57]	-	-	-	-	-	-	-	-	60.9	-	39.9	1.03	-
LLaMA-VID-7B [55]	-	-	-	-	-	-	-	1521(+)	65.1	59.9	-	1.08	-
mPLUG-Owl3-8B [100]	-	-	-	-	-	-	-	-	77.6	-	53.5	1.35	-
InternVL2-8B [13]	-	-	-	-	-	-	-	-	81.7	76.2	54.0	1.28	-
PixelLM-7B [77]	73.0	66.3	69.3	-	-	-	-	309/135	17.4	-	-	-	-
LaSagnA [86]	76.8	66.4	70.6	-	-	-	-	0/0	0.0	-	-	-	-
LISA-7B [45]	74.1	62.4	66.4	-	-	-	-	1/1	0.4	-	-	-	-
GLaMM-7B [75]	79.5	72.6	74.2	-	-	-	-	14/9	36.8	-	-	-	28.9
LLaVA-G-7B [107]	77.1	68.8	71.5	-	-	-	-	-	-	-	-	-	-
GSVA-13B [94]	79.2	70.3	75.7	-	-	-	-	-	-	-	-	-	-
OMG-LLaVA-7B [109]	78.0	69.1	72.9	-	-	-	-	1177/235	47.9	56.5	-	-	29.9
VideoLISA-3.8B [3]	73.8	63.4	68.3	44.4	68.8	63.7	-	-	-	-	-	-	-
VISA-13B [96]	72.4	59.8	65.5	44.5	70.4	63.0	50.9	-	-	-	-	-	-
Sa2VA-1B (Ours)	77.4	69.9	72.3	41.7	72.3	65.3	47.6	1381/405	68.3	64.8	39.9	1.07	23.8
Sa2VA-4B (Ours)	80.4	74.3	76.7	46.2	73.8	70.0	53.2	1553/540	76.8	72.6	50.4	1.23	28.2
Sa2VA-8B (Ours)	81.9	76.5	78.9	46.9	75.2	70.7	57.6	1651/578	82.4	75.5	52.1	1.34	31.0
Sa2VA-26B (Ours)	82.5	78.8	79.7	46.2	77.0	70.1	58.4	1691/538	83.7	76.8	52.6	1.45	33.5

Ref-SAV

Benchmark Results on Ref-SAV

Table 5: Ref-SAV validation sets. zs: zero-shot testing.

Method	Long			Short			Overall		
	J	F	J&F	J	F	J&F	J	F	J&F
UniRef++ [90] (zs)	14.1	10.8	12.5	9.0	8.2	8.6	11.6	9.5	10.5
UNINEXT [95] (zs)	11.7	8.3	10.0	5.8	4.4	5.1	8.8	6.4	7.6
MeVIS [17] (zs)	12.1	7.1	11.3	6.2	5.3	5.5	12.2	9.8	10.3
VISA [96] (zs)	16.1	12.2	14.1	12.3	9.6	9.2	13.2	11.3	11.8
Sa2VA-8b (zs)	47.7	50.9	49.3	31.5	35.0	33.3	39.6	43.0	41.3
Sa2VA-8b (Ours)	57.0	60.4	58.7	39.5	42.9	41.2	48.3	51.7	50.0

Sa2VA-8b (Ours) trained on Ref-SAV

Sa2VA - Experiments

Scalability of Sa2VA

Table 10: Ablation study on using more datasets.

Dataset	Size	RefCOCO	RefCOCOG	MMBench	MME	MeVIS(val_u)
baseline	1.2M	77.4	72.3	68.3	1381/405	50.8
Inifinity-MM [26]	1.2M+3M	77.1(-0.3)	72.6(+0.3)	70.4(+2.1)	1396/346(-44)	51.2(+0.4)
Ref-SAV	1.2M+37K	77.2(-0.2)	72.6(+0.3)	68.2(-0.1)	1384/418(+16)	52.5(+1.7)

Considering that Ref-SAV is designed for referring video segmentation, these improvements demonstrate great potential of further data scaling.

Sa2VA - Experiments

Co-Training Effect

Table 7: Ablation study on co-training effect on multiple datasets. We use Sa2VA-1B to test the performance.

Data	Image Segmentation			Video Segmentation		Image Chat			Video Chat	
	RefCOCO	RefCOCO+	RefCOCOg	MeViS (val_u)	Ref-DAVIS17	MME	MMBench	SEED-Bench	Video-MME	MMBench-Video
All Data	77.4	69.9	72.3	50.8	72.3	1381/405	68.3	64.8	39.9	1.07
w/o Image QA	78.0	70.1	72.2	48.3	73.0	1298/359	63.4	63.8	39.7	0.39
w/o Image Segmentation	20.2	20.6	23.2	38.0	48.8	1393/408	70.1	65.7	41.2	1.08
w/o Video QA	78.0	70.4	72.6	50.7	74.3	1370/402	69.1	65.0	41.3	0.71
w/o Video Segmentation	77.4	69.1	72.4	44.4	69.0	1403/398	67.8	64.9	40.4	1.04

Sa2VA - Visualization



 Please segment the girl wearing the yellow dress.

Sure, it is [SEG]. 

Sa2VA - Visualization



 Please segment the main character.

Sure, it is [SEG]. 

Sa2VA - Visualization



👤 What is the atmosphere of the scene?

🤖 The scene has a dark and mysterious atmosphere, with the men dressed in suits and ties, and the dimly lit room.

Sa2VA - Visualization



👤 Can you describe what this man holding the cat is doing and how he feels?

🤖 The man in the image is holding a cat in his arms while sitting in a chair. He appears to be looking at the cat with a smile on his face, suggesting that he is enjoying the company of the cat and feeling happy.

Sa2VA - Visualization



 Can you describe what this man holding the cat is doing and how he feels?

 ... It is possible that he is feeling overwhelmed, anxious, or even angry. Overall, the image conveys a sense of tension and unease, as the man struggles to cope with his emotions while holding the cat.

Sa2VA - Visualization



The object is a silver Honda Jazz car, parked on a paved area with its rear door open. It bears the license plate "AD 1632 TP" and features the Honda logo on the back, red taillights, and a rear windshield wiper. A person in a white shirt and black pants approaches the car from the right, interacting with it, possibly preparing to enter or exit. The person's actions suggest checking the car or getting ready to drive it. The car remains stationary throughout, ready for use.



The object is a person standing near a clothing stall in a shopping mall corridor. They are wearing a light-colored hoodie, khaki pants, and black shoes, with dark hair tied back. The individual is positioned close to a stall displaying various garments on racks and shelves. Throughout the video, they browse or stand still, occasionally shifting position while looking at the clothes. The person in the light-colored hoodie remains relatively stationary, primarily focused on the clothing stall.



Distilling Spectral Graph for Object Context Aware Open-Vocabulary Semantic Segmentation

Chanyoung Kim Dayun Ju Woojung Han Ming-Hsuan Yang Seong Jae Hwang

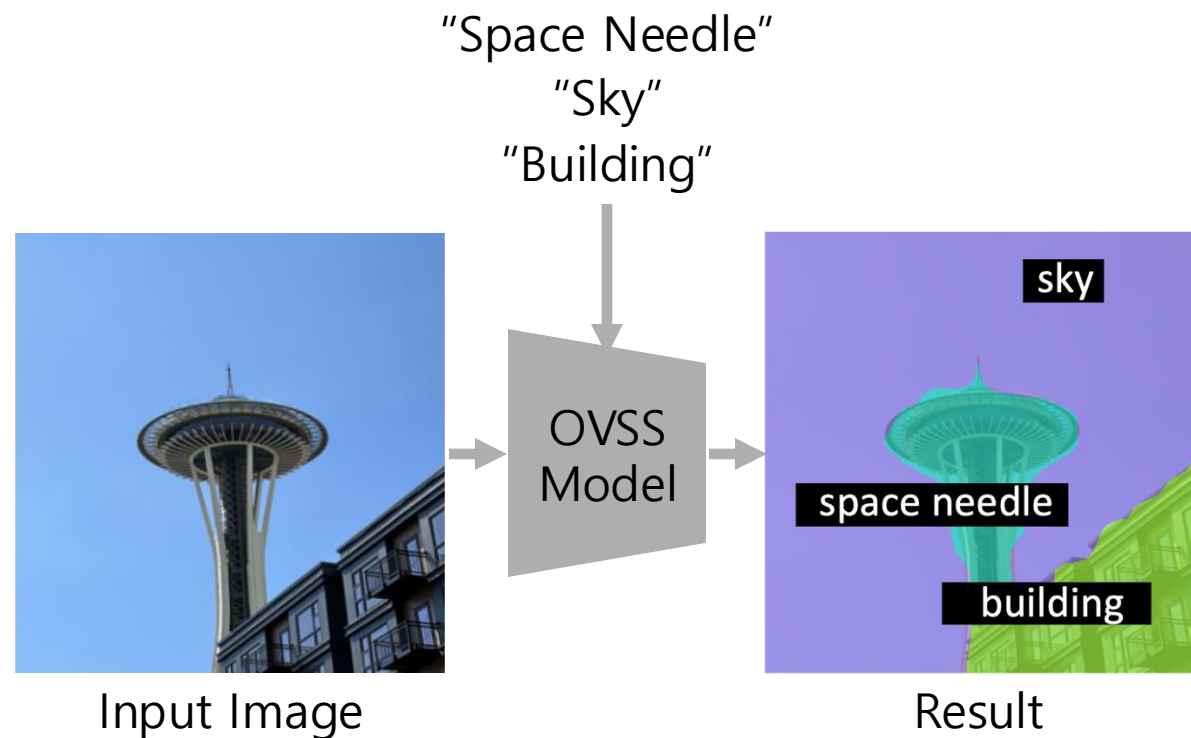
Yonsei University University of California, Merced

<https://micv-yonsei.github.io/cass/>



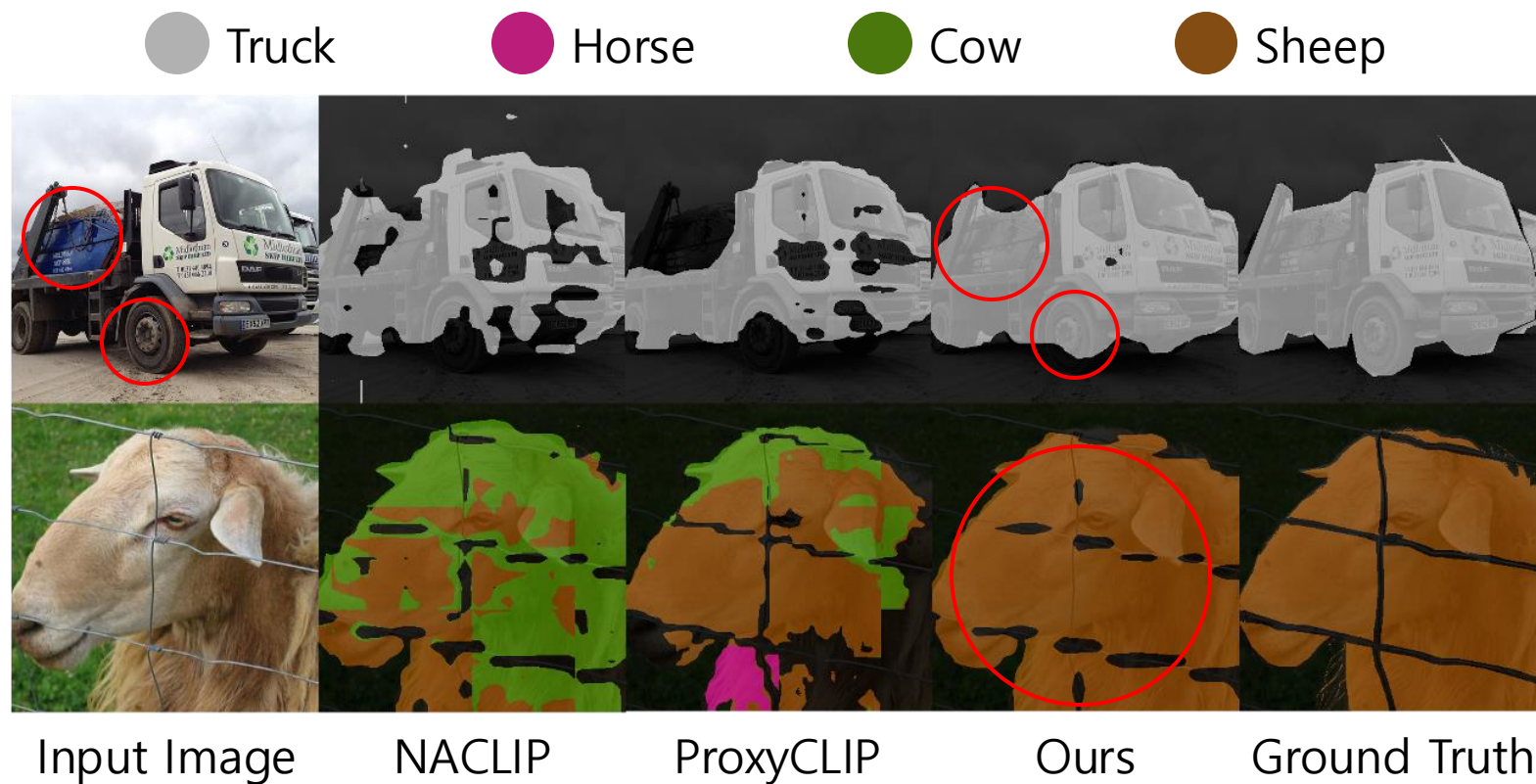
Motivation

- Open-Vocabulary Semantic Segmentation (OVSS)
 - Achieving adaptability across diverse domains by segmenting **unseen user given arbitrary classes**
- Advantages of Training-Free OVSS
 - Generalization
 - Real-World applicability



Existing OVSS Methods

- Rely on **patch-level** similarity without exploiting global structure
- Lack **object-level context**

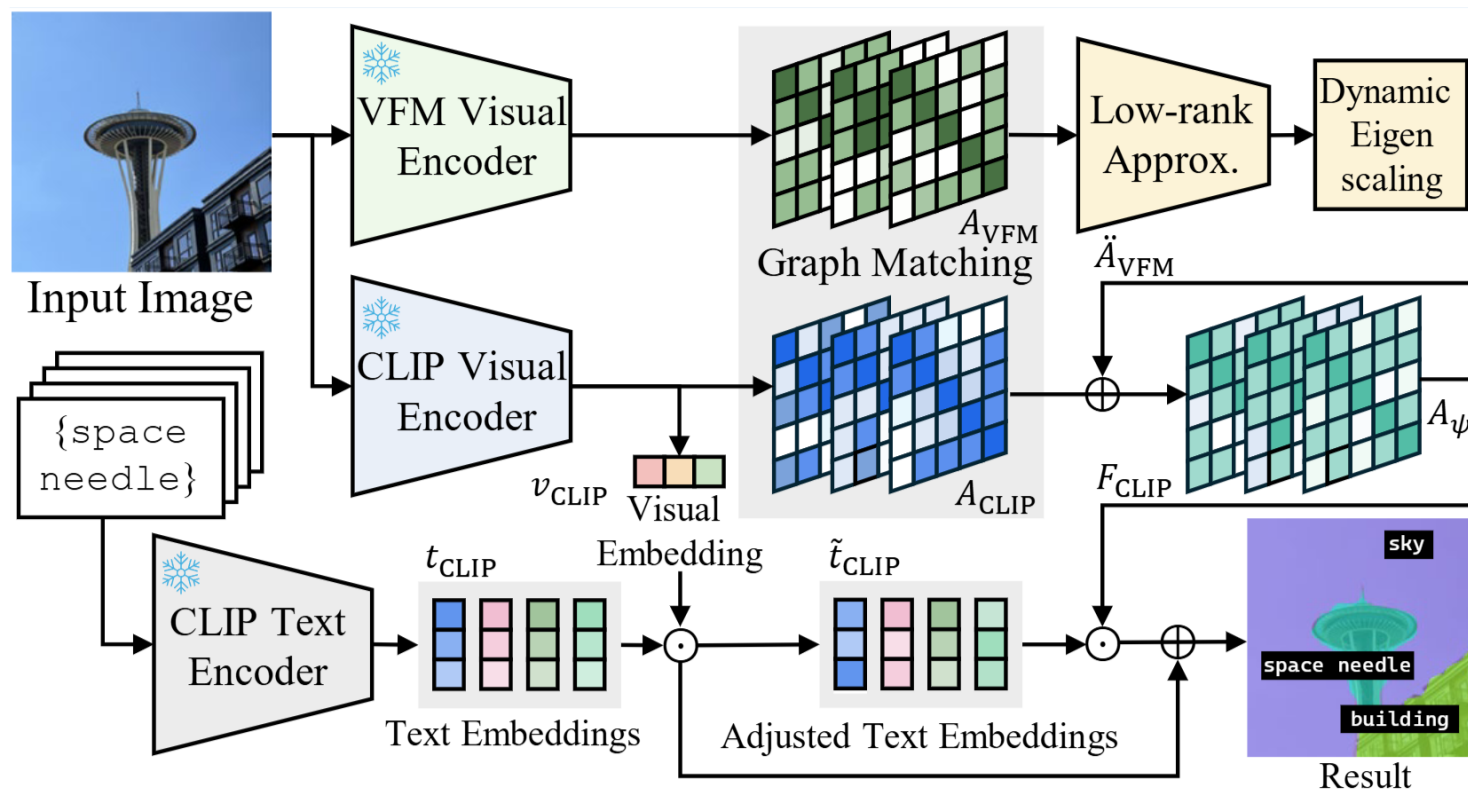


- Cargo area and wheels should be grouped as a truck.
- Components of a sheep should be unified under the correct user-given object class.

Overview

- Object-Level **C**ontext **A**ware training-free OV**SS**, CASS
 - Spectral Object-Level Context Distillation
 - Object Presence-Driven Object-Level Context

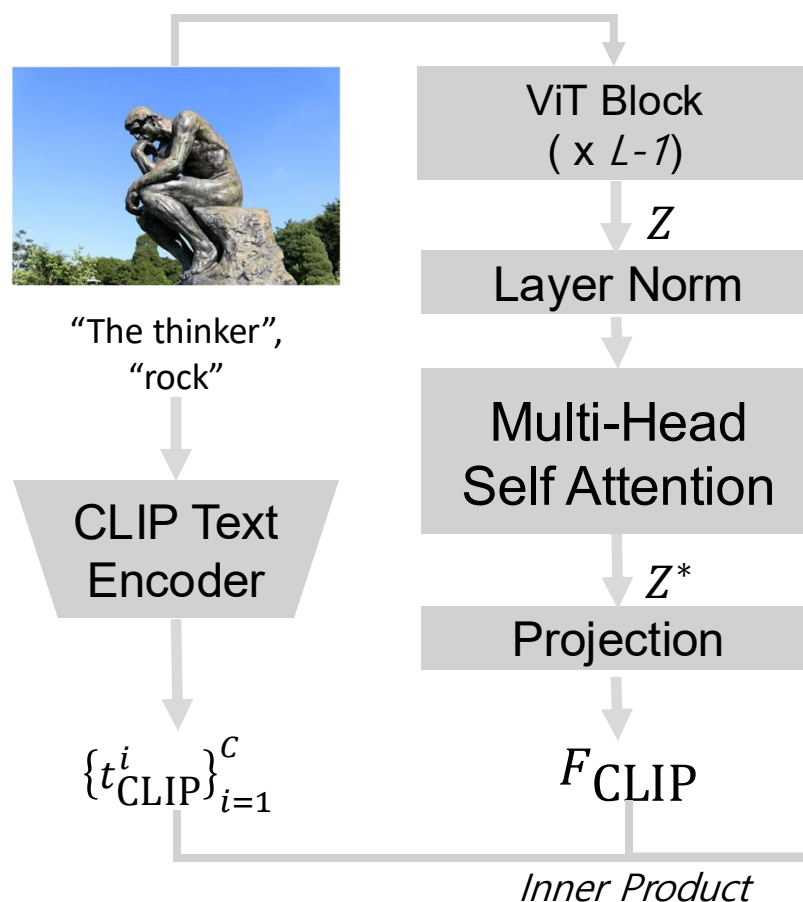
1. transfer dense object-level semantics from VFM into CLIP for more coherent and compact segmentation



2. prevent different parts of the same object from being assigned to semantically similar but incorrect categories

Preliminaries

- CLIP Visual Encoder: $\mathcal{F}_{\text{CLIP}}^v$
 - Modify the last block of CLIP visual encoder to make it suitable for dense prediction



$$Z^* = \text{Concat}[M_{\text{CLIP}}^1 V_{\text{CLIP}}^1, \dots, M_{\text{CLIP}}^h V_{\text{CLIP}}^h] W^O$$

$$M_{\text{CLIP}}^i = \text{softmax}(A_{\text{CLIP}}^i / \sqrt{D_h})$$

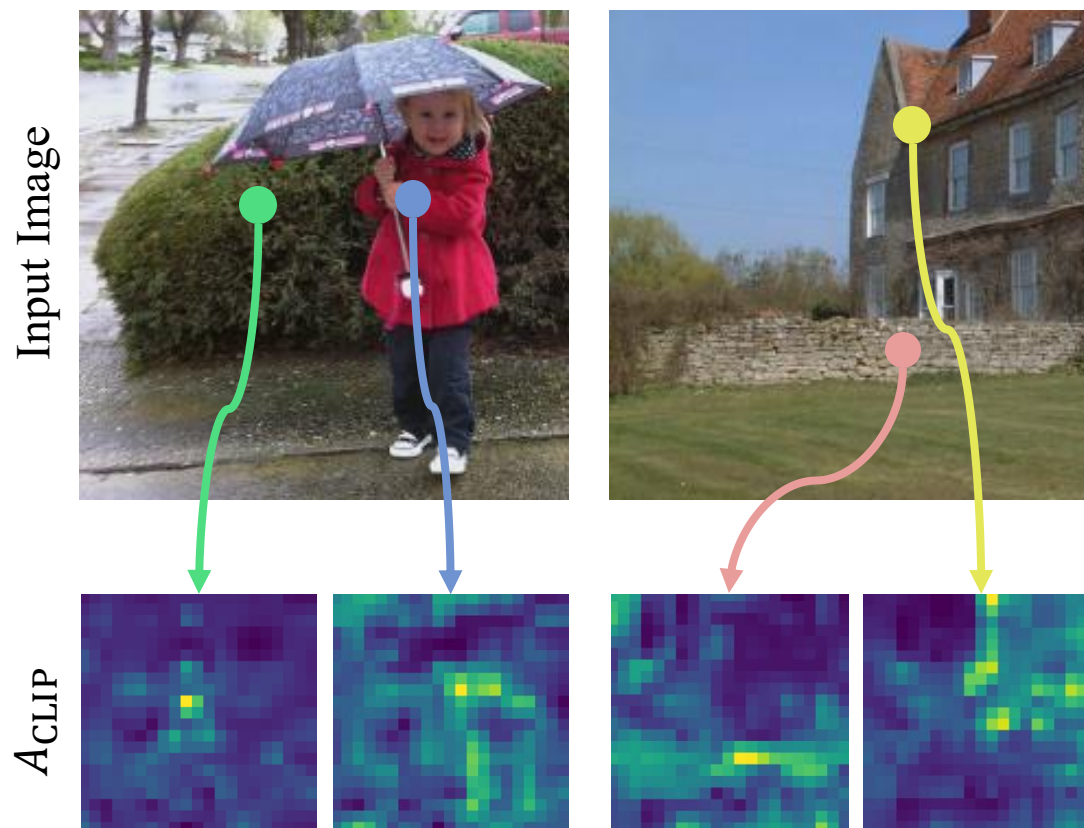
$$A_{\text{CLIP}}^i = K_{\text{CLIP}}^i K_{\text{CLIP}}^{i\top}, \forall i \in \{1, \dots, h\}$$

Interpolation
argmax



Method

- Spectral Object-Level Context Distillation
 - Since CLIP neglects object-level context, distill VFM (DINO) into CLIP to enable object-level contextual understanding.



CLIP's attention performs poorly in highlighting objects.

- Spectral Object-Level Context Distillation

- Since CLIP neglects object-level context, distill VFM (DINO) into CLIP to facilitate object-level contextual understanding.
- First transform VFM attention graph's **patch-level** representation into **object-level** representation.
- Then distill VFM attention's **object-level** representation into CLIP's attention graph, enhancing intra-object consistency.

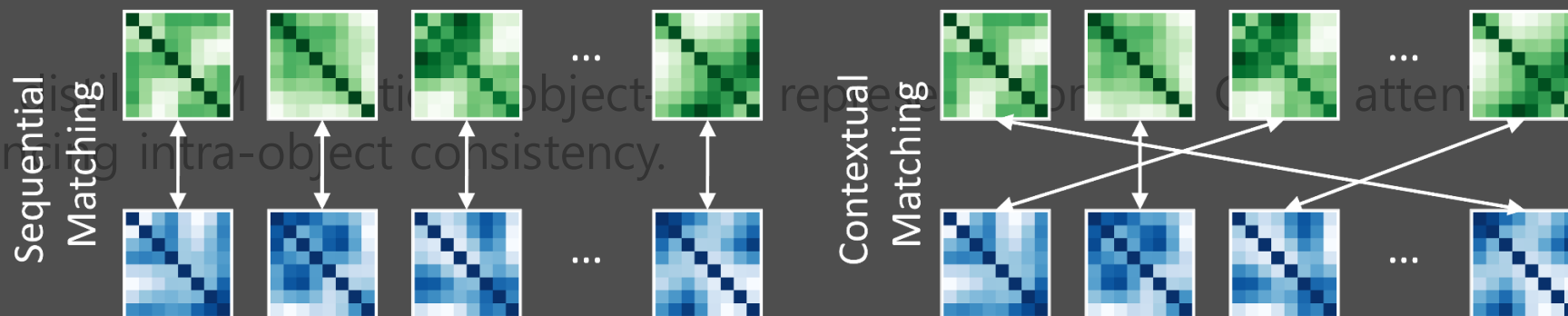
Method

- Spectral Object-Level Context Distillation

- Q: How to account for the **multi-head attention mechanisms** in both VFM (DINO) and CLIP?

- We first transform VFM attention graph's patch-level representation into object-level representation.

- Then distill multi-head attention graph, enhancing intra-object consistency.



Method

- Different attention heads focus on distinct parts of an image
 - Complementary Spectral Graph Matching
 - Propose matching attention graphs with *contrasting structures*, enabling VFM to supplement object-level contextual knowledge that CLIP alone struggles to capture.

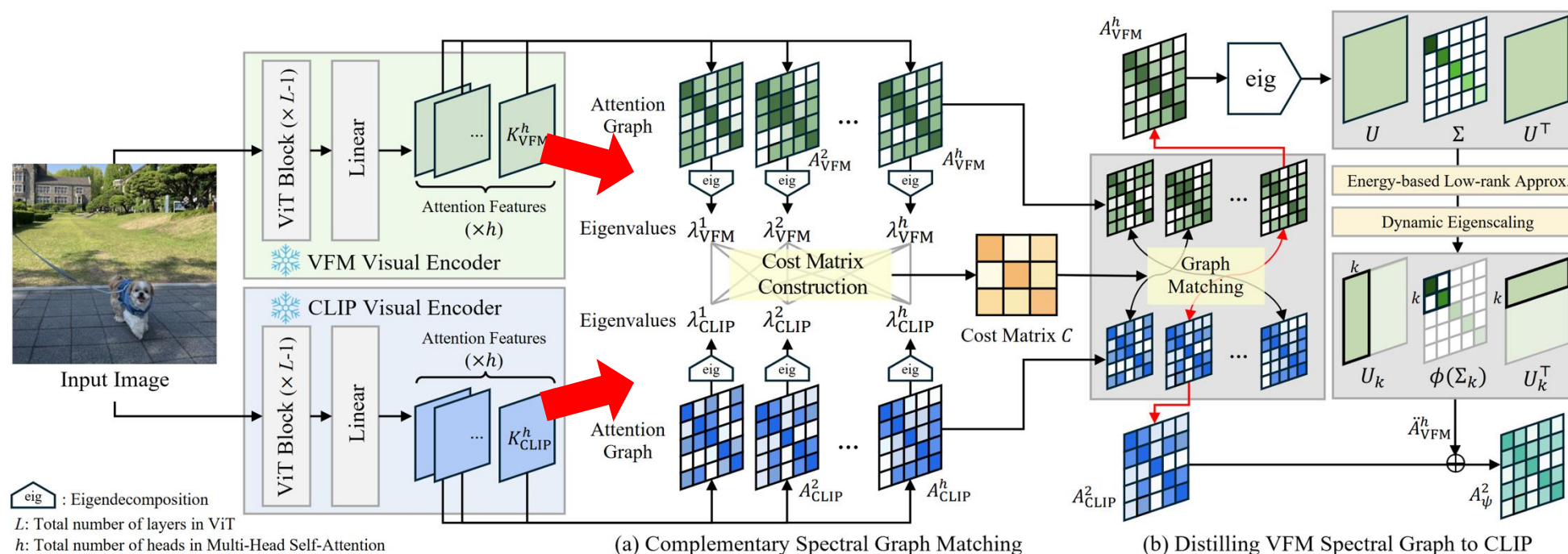
Method

• Spectral Object-Level Context Distillation

• Complementary Spectral Graph Matching

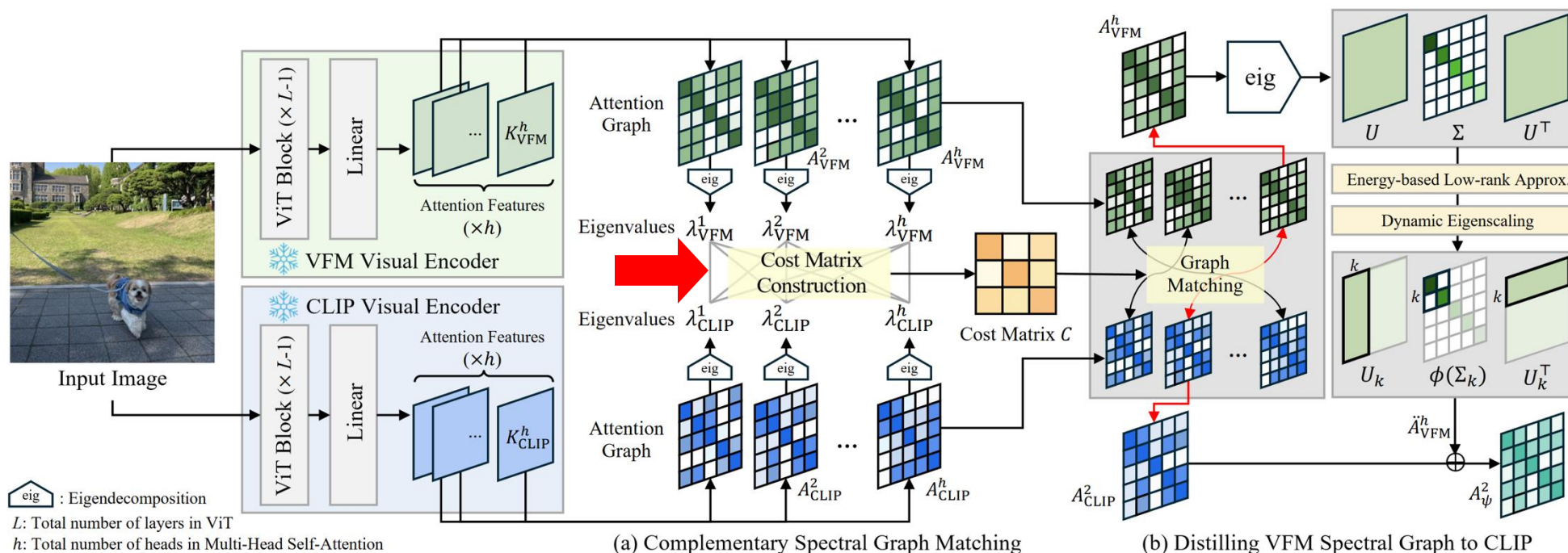
1) Eigenvalue as key structure signature

- To analyze the unique structural features of each attention graph, perform **eigendecomposition** on the graph for each attention head.



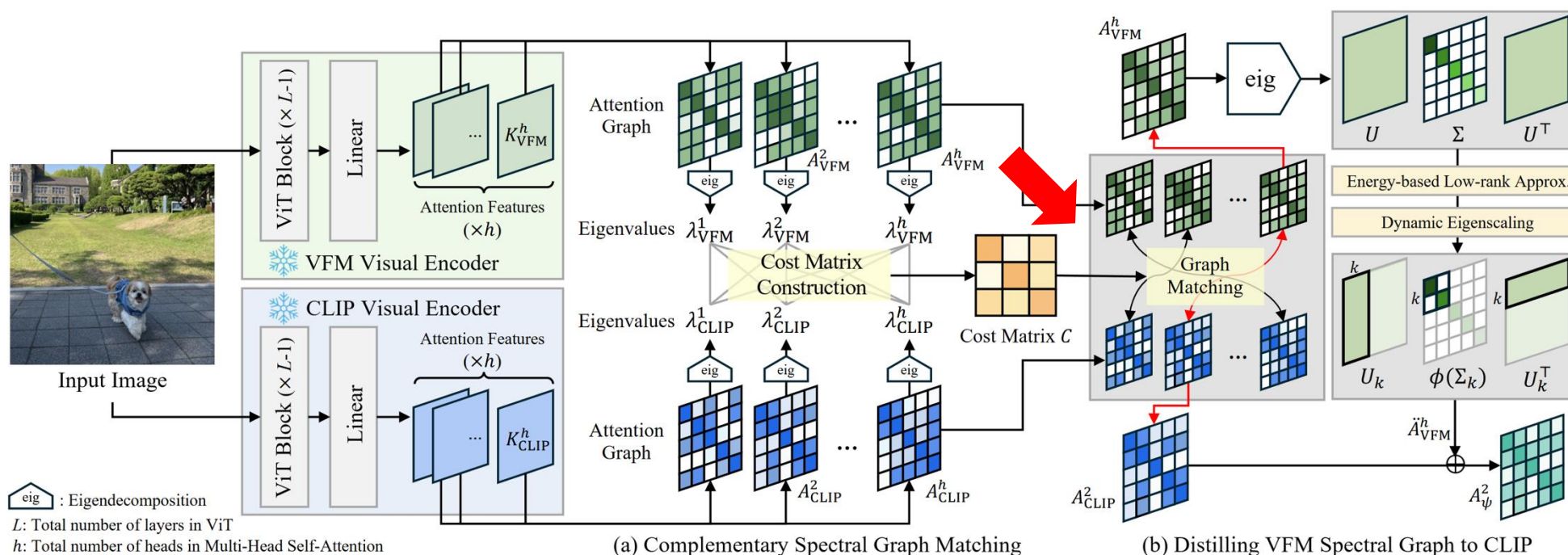
Method

- Spectral Object-Level Context Distillation
 - Complementary Spectral Graph Matching
 - 2) Graph Matching via Spectral Distribution
 - Construct a **cost matrix** that quantifies structural differences between VFM and CLIP attention graphs using Wassertein distance (based on eigenvalues)



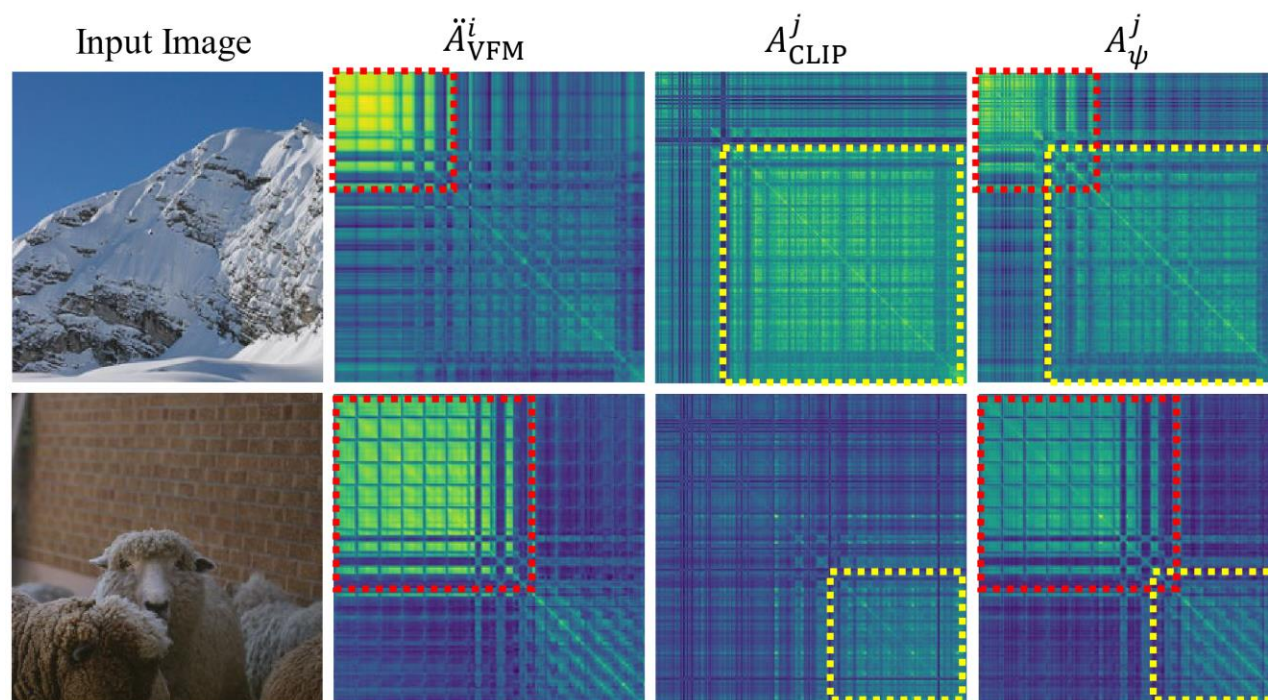
Method

- Spectral Object-Level Context Distillation
 - Complementary Spectral Graph Matching
 - 2) Graph Matching via Spectral Distribution
 - Use this cost matrix, match graph heads (using Hungarian algorithm) to enable **complementary knowledge transfer** for object-level context.



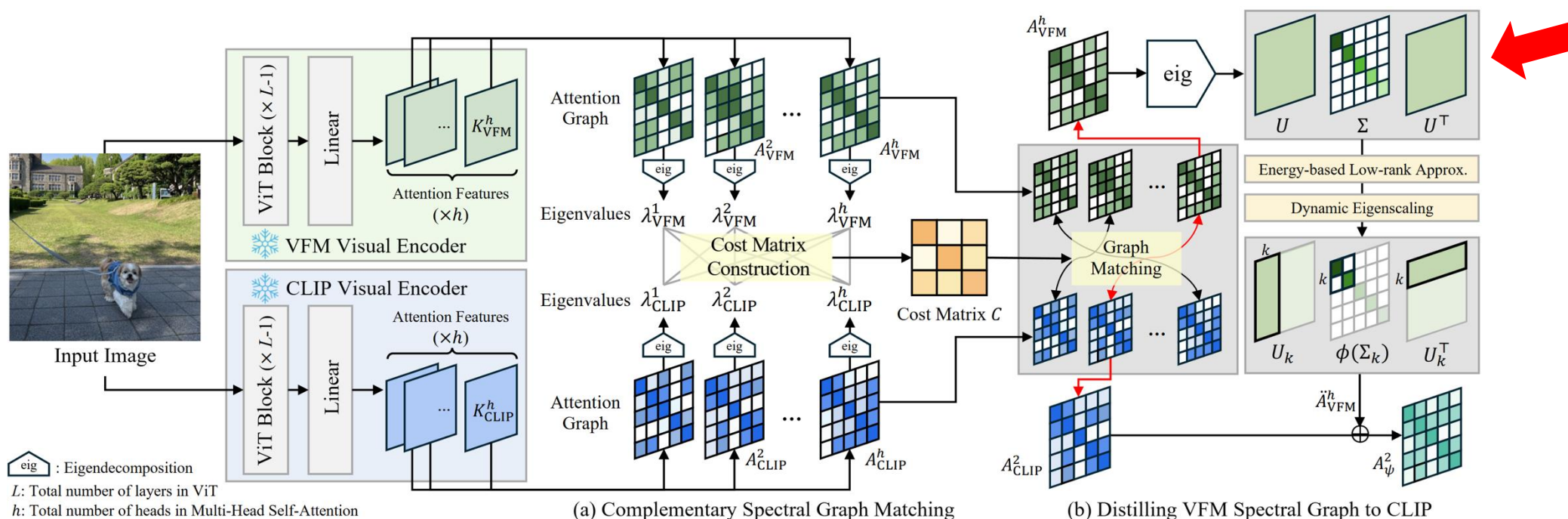
Method

- Spectral Object-Level Context Distillation
 - Complementary Spectral Graph Matching
 - 2) Graph Matching via Spectral Distribution
 - Use this cost matrix, we match graph heads to enable **complementary knowledge transfer** for object-level context.



Method

- Transform VFM **patch-level** to object-level representations
 - Distilling VFM Spectral Graph to CLIP
 - 1) Low-Rank Dynamic Eigenscaling
 - Perform low-rank approximation to capture essential **object-level** contextual features while discarding noise and irrelevant details.



Method

- Spectral Object-Level Context Distillation
 - Distilling VFM Spectral Graph to CLIP
 - 1) Low-Rank Dynamic Eigenscaling
 - Perform energy-based a low-rank approximation, capturing essential object-level contextual features while discarding noise and irrelevant details.

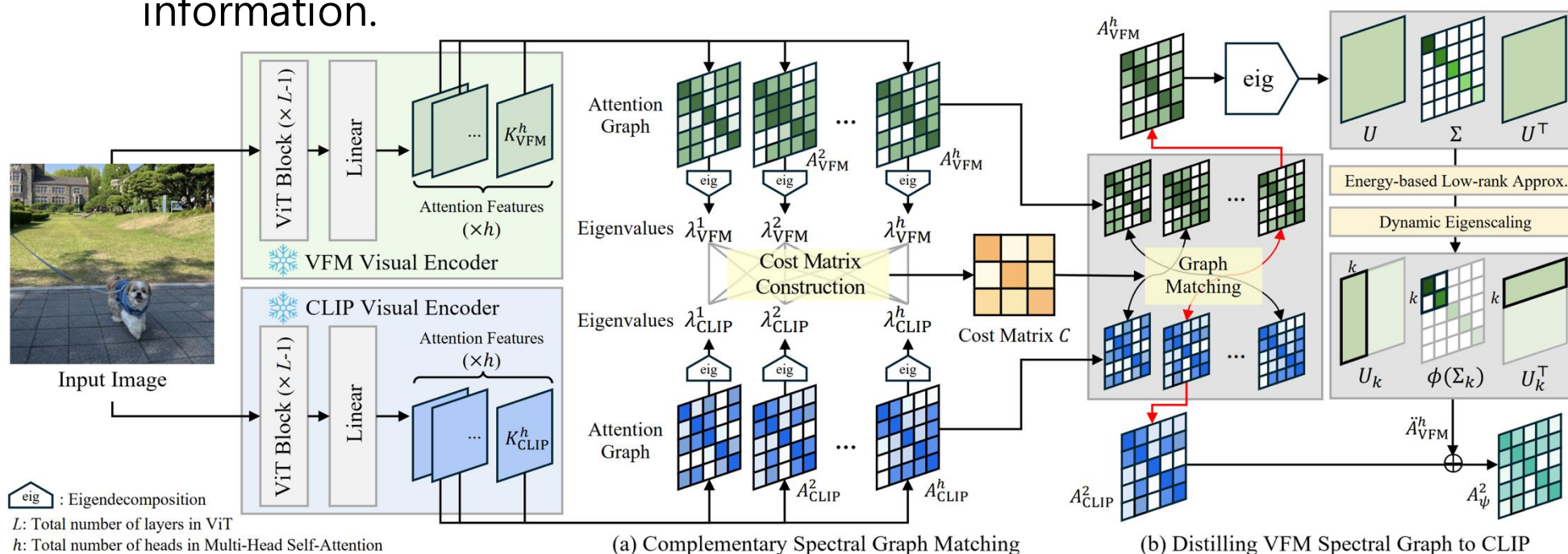
Method

• Spectral Object-Level Context Distillation

• Distilling VFM Spectral Graph to CLIP

1) Low-Rank Dynamic Eigenscaling

- Refine low-rank components using a **dynamic scaling function** ϕ , amplifying significant eigenvalues and suppressing smaller ones to emphasize key structural information.



Method

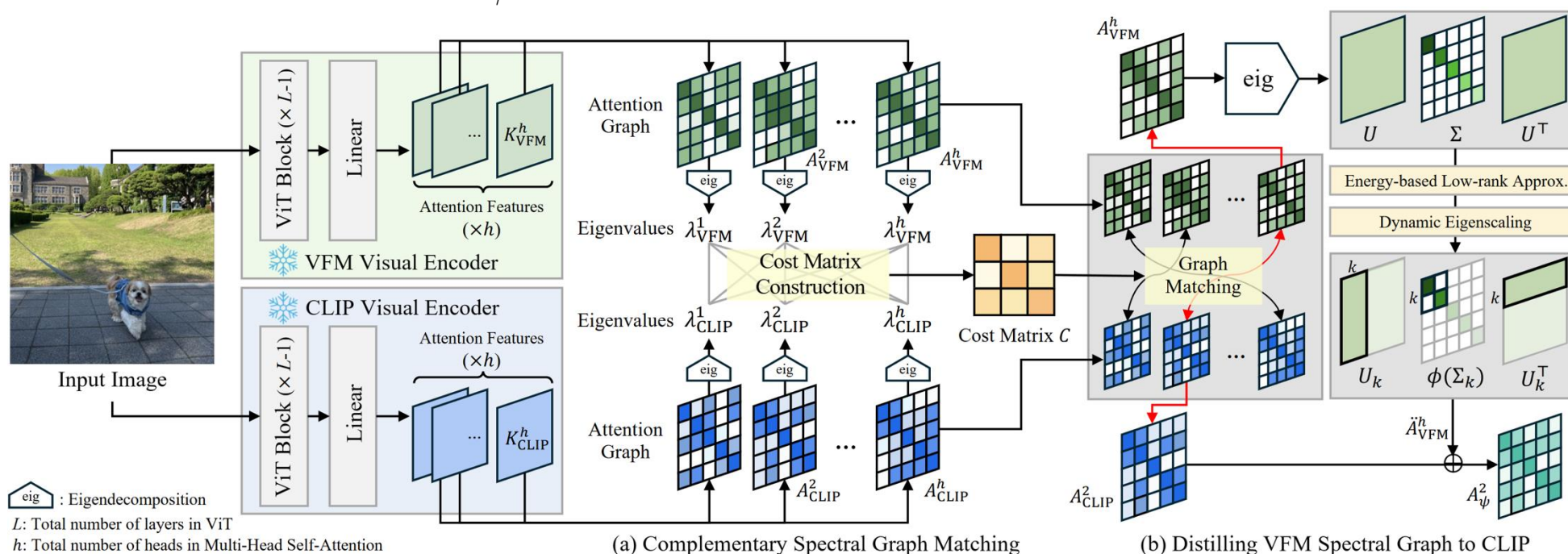
• Spectral Object-Level Context Distillation

• Distilling VFM Spectral Graph to CLIP

2) VFM Graph Distillation

- Aggregate VFM and CLIP attention structures using spectral (Wasserstein) distance

$$A_{\psi}^j = (w_{ij} \ddot{A}_{\text{VFM}}^i + A_{\text{CLIP}}^j) / (w_{ij} + 1)$$



Method

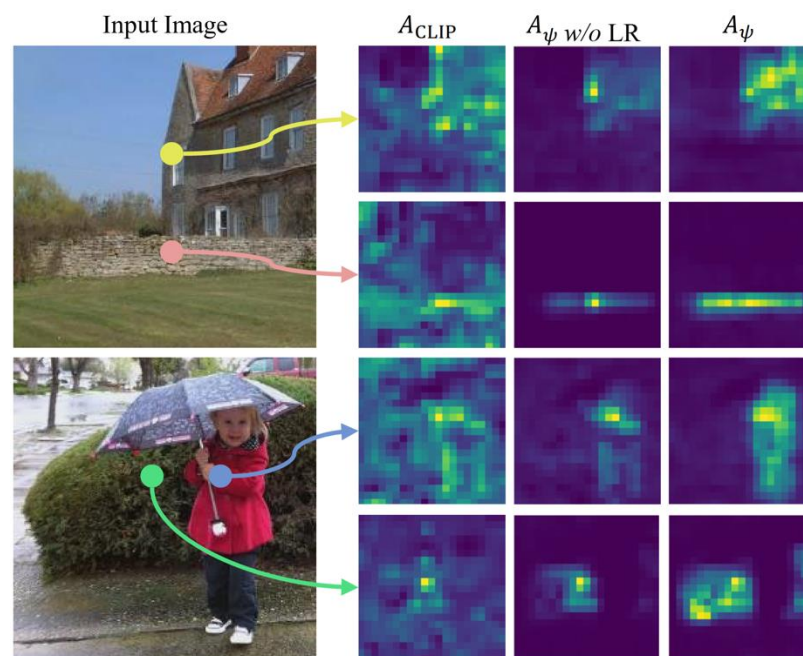
- Spectral Object-Level Context Distillation

- Distilling VFM Spectral Graph to CLIP

- 2) VFM Graph Distillation

- Aggregate VFM and CLIP attention structures using spectral distance.

$$A_{\psi}^j = (w_{ij} \ddot{A}_{\text{VFM}}^i + A_{\text{CLIP}}^j) / (w_{ij} + 1)$$



Method

- Object Presence-Driven Object-Level Context

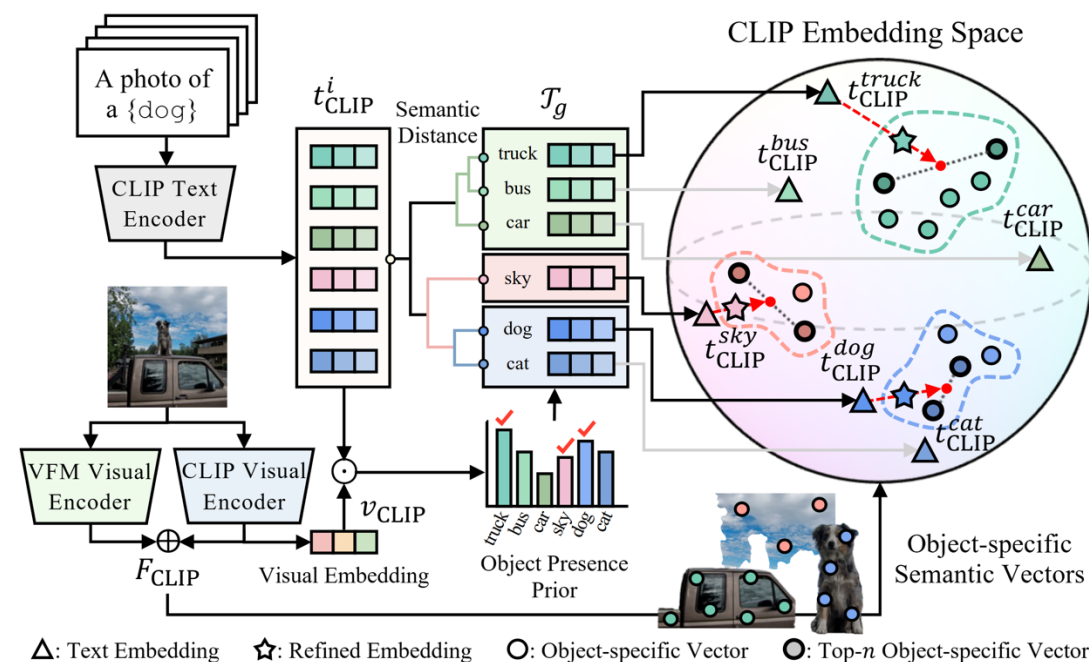
- Due to nature of OVSS, users provide arbitrary text prompts
 - Parts of the same objects may be assigned to different closely related categories
- Further enhance object-level context using **zero-shot object classification capability** (i.e., object presence prior) of CLIP
- Object presence prior: $P(i) = \{t_{\text{CLIP}}^i\}_{i=1}^C v_{\text{CLIP}}$
 - Object-Guided Text Embedding Adjustment
 - Object Perspective Patch-Text Similarity

Method

• Object Presence-Driven Object-Level Context

• Object-Guided Text Embedding Adjustment

- ① Grouping semantically adjacent classes: Hierarchical clustering of CLIP text embedding's semantics.
- ② Identifying objects that is likely to be present in the image: Choose classes that is most likely to be presented in the image using Object presence prior.
- ③ Adjust text embeddings towards object-specific image vectors



Method

- Object Presence-Driven Object-Level Context

- Object Perspective Patch-Text Similarity

Integrate object-presence prior with patch-text similarity to produce segmentation map from an object-perspective view.

$$\hat{\mathcal{S}}^* = (1 - \gamma) \cdot \hat{\mathcal{S}} + \gamma \cdot \{t_{\text{CLIP}}^i\}_{i=1}^C v_{\text{CLIP}}$$

Experiments

- Experiment Setups
 - Datasets

	Without Background			With Background				
Dataset	PASCAL VOC (V21)	PASCAL Context (PC60)	COCO-Object (C-Obj)	PASCAL VOC (V20)	PASCAL Context (PC59)	COCO-Stuff (C-Stf)	Cityscapes (City)	ADE20K (ADE)
Number of Classes	21	60	81	20	59	171	19	150

- Models
 - Vision-language Model: CLIP ViT-B/16
 - Vision Foundation Model: DINO ViT-B/8

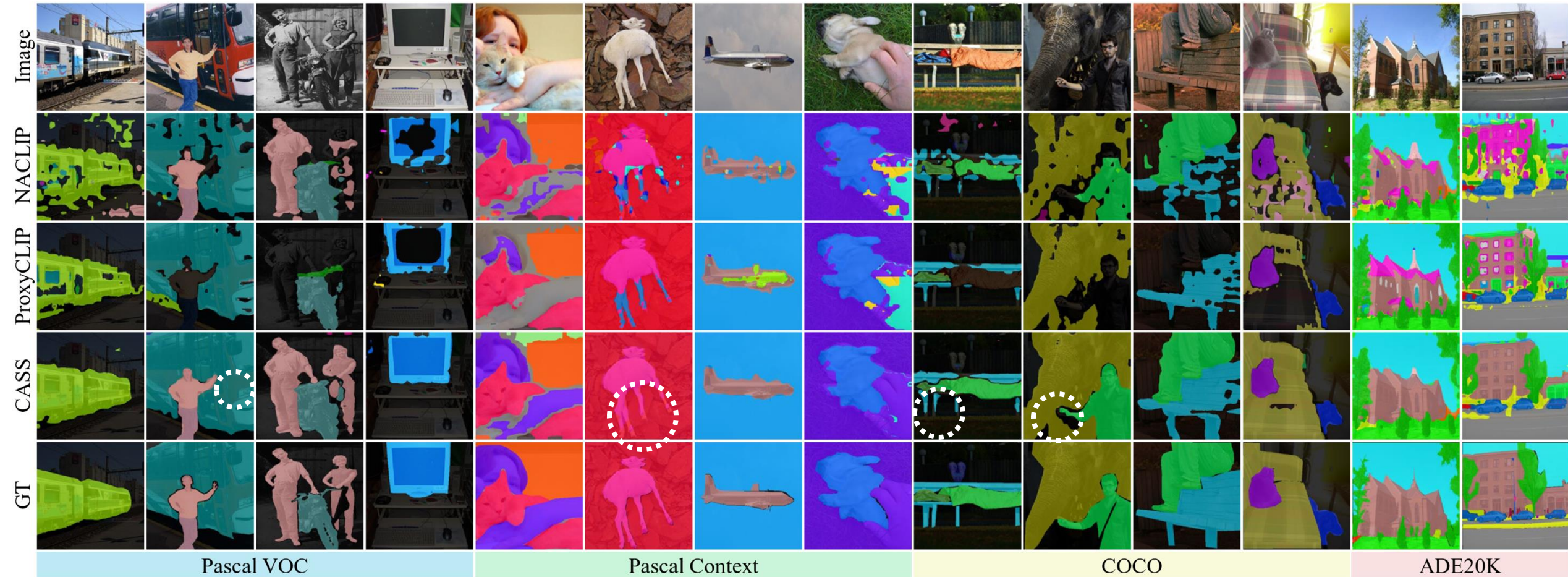
Experiments

- Quantitative Evaluation: mIoU

Table 1. Quantitative results with state-of-the-art unsupervised open-vocabulary semantic segmentation models on eight datasets.

Model		Supporting Dataset	Extra-Training	Fair	V21	PC60	C-Obj	V20	PC59	C-Stf	City	ADE	Avg.
GroupViT [49]	CVPR'22	CC12M+RedCaps	✓	✗	50.4	18.7	27.5	79.7	23.4	15.3	11.1	9.2	29.4
TCL [9]	CVPR'23	CC3M+CC12M	✓	✗	55.0	30.4	31.6	83.2	33.9	22.4	24.0	17.1	37.2
CoDe [46]	CVPR'24	CC3M+RedCaps	✓	✗	57.5	30.5	32.3	-	-	23.9	28.9	17.7	-
CLIP-DINOiser [47]	ECCV'24	ImageNet1k	✓	✗	62.1	32.4	34.8	80.9	35.9	24.6	31.7	20.0	40.3
ReCo [40]	NeurIPS'22	ImageNet1k	✗	✗	25.1	19.9	15.7	57.7	22.3	14.8	21.6	11.2	23.5
FOSSIL [3]	WACV'24	COCO Captions	✗	✗	-	-	-	-	35.8	24.8	23.2	18.8	-
FreeDa [4]	CVPR'24	COCO Captions	✗	✗	-	-	-	85.6	43.1	27.8	36.7	22.4	-
CLIP [34]	ICML'21	✗	✗	✓	18.6	7.8	6.5	49.1	11.2	7.2	6.7	3.2	13.8
MaskCLIP [53]	ECCV'22	✗	✗	✓	38.3	23.6	20.6	74.9	26.4	16.4	12.6	9.8	27.9
GEM [6]	CVPR'24	✗	✗	✓	46.2	-	-	-	32.6	-	-	15.7	-
CaR [42]	CVPR'24	✗	✗	✓	48.6	13.6	15.4	73.7	18.4	-	-	5.4	-
PnP-OVSS [42]	CVPR'24	✗	✗	✓	-	-	36.2	51.3	28.0	17.9	-	14.2	-
CLIPtrase [39]	ECCV'24	✗	✗	✓	50.9	29.9	43.6	81.0	33.8	22.8	21.3	16.4	32.7
ClearCLIP [24]	ECCV'24	✗	✗	✓	51.8	32.6	33.0	80.9	35.9	23.9	30.0	16.7	38.1
SCLIP [43]	ECCV'24	✗	✗	✓	59.1	30.4	30.5	80.4	34.1	22.4	32.2	16.1	38.2
LaVG [20]	ECCV'24	✗	✗	✓	62.1	31.6	34.2	82.5	34.7	23.2	26.2	15.8	38.8
ProxyCLIP [25]	ECCV'24	✗	✗	✓	59.1	35.2	36.2	78.2	38.8	26.2	38.1	19.6	41.4
NACLIP [16]	WACV'25	✗	✗	✓	58.9	32.2	33.2	79.7	35.2	23.3	35.5	17.4	39.4
CASS		✗	✗	✓	65.8	36.7	37.8	87.8	40.2	26.7	39.4	20.4	44.4

Qualitative Evaluation



Ablations

- Different CLIP Backbones

Table 4. Ablation results on different CLIP backbones.

	Model	V21		PC59		C-Stf		Avg. mIoU	Avg. pAcc
		mIoU	pAcc	mIoU	pAcc	mIoU	pAcc		
ViT-B/32	SCLIP [50]	50.6	81.0	28.7	49.6	20.0	33.4	33.1	54.7
	LaVG [22]	54.8	84.4	29.0	50.1	20.5	33.7	34.8	56.1
	ProxyCLIP [30]	57.9	85.6	35.2	59.7	23.6	40.2	38.9	61.8
	NACLIP [17]	51.1	82.8	32.4	56.8	21.2	37.1	34.9	58.9
	CASS	58.2	86.2	36.5	61.0	24.4	40.8	39.7	62.7
ViT-L/14	SCLIP [50]	44.4	78.1	25.2	46.6	17.6	29.6	29.1	51.4
	LaVG [22]	51.5	84.1	27.5	50.3	19.4	32.3	32.8	55.6
	ProxyCLIP [30]	59.8	85.7	38.3	61.0	26.2	43.5	41.4	63.4
	NACLIP [17]	52.2	83.1	32.1	52.8	21.4	35.7	35.2	57.2
	CASS	62.1	88.0	39.1	62.1	26.3	43.4	42.5	64.5

Ablations

- Different VFM (DINO) Backbones

Table 6. Ablation results with different VFM backbones. For the CLIP visual encoder, ViT-B/16 is used.

Model			V21		PC59		C-Stf		Avg. mIoU	Avg. pAcc
			mIoU	pAcc	mIoU	pAcc	mIoU	pAcc		
DINOv1 ViT-B/8	LaVG [22]	ECCV'24	62.1	89.3	34.7	58.9	23.2	39.1	40.0	62.4
	ProxyCLIP [30]	ECCV'24	59.1	86.6	38.8	63.4	26.2	43.4	41.4	64.5
	CASS		65.8	90.1	40.2	65.0	26.7	43.6	44.2	66.2
DINOv1 ViT-B/16	LaVG [22]	ECCV'24	61.5	88.9	34.6	58.9	22.8	38.8	39.6	62.2
	ProxyCLIP [30]	ECCV'24	56.6	85.8	37.4	62.0	25.1	42.2	39.7	63.3
	CASS		64.3	89.1	38.9	63.8	26.1	43.1	43.1	65.3
DINOv2 ViT-B/14	LaVG [22]	ECCV'24	25.3	75.3	25.1	49.3	17.3	34.4	22.6	53.0
	ProxyCLIP [30]	ECCV'24	57.1	85.2	37.3	61.4	25.3	42.3	39.9	63.0
	CASS		63.0	88.6	38.1	62.2	25.3	42.0	42.1	64.3

Ablations

- Distance Metric for Graph Matching

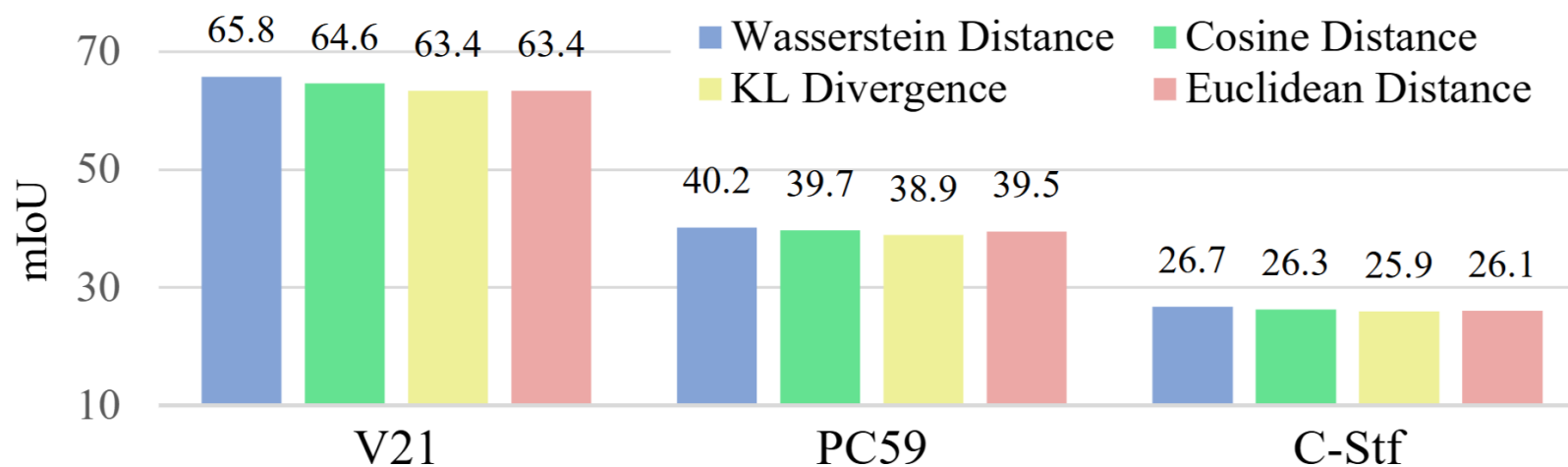


Figure 8. Performance evaluation of different distance metrics used in graph matching and graph distillation.

Ablations

- Graph Matching Strategy

Table 8. Ablation results on different graph matching strategies.

Method	V21		PC59		C-Stf		Avg. mIoU	Avg. pAcc
	mIoU	pAcc	mIoU	pAcc	mIoU	pAcc		
Sequential	64.9	89.6	39.6	64.5	26.2	43.1	43.5	65.7
Similar	65.5	90.0	40.2	65.1	26.6	43.6	44.1	66.2
Complementary	65.8	90.1	40.2	65.0	26.7	43.6	44.2	66.2

Ablations

- Number of Eigenvalues for Low-Rank Components

Table 9. Ablation results on rank selection methods for low-rank components.

Method	Adaptive	V21		PC59		C-Stf		Avg. mIoU	Avg. pAcc
		mIoU	pAcc	mIoU	pAcc	mIoU	pAcc		
$k = 3$	✗	62.7	89.3	39.2	64.0	25.9	42.6	42.6	65.3
$k = 4$	✗	64.5	89.8	39.8	64.6	26.4	43.1	43.6	65.8
Eigengap	✓	64.9	89.6	39.8	64.6	26.3	43.3	43.7	65.8
Energy-based	✓	65.8	90.1	40.2	65.0	26.7	43.6	44.2	66.2

Experiments

- Real-World Application

