Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.1: New Research Directions

Mehul Agarwal, Soham Dinesh Tiwari, Haofei Yu and LP Morency

** Fall 2021, 2022 and 2023 co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk. Spring 2023 edition taught by Yonatan and Daniel Fried*

Administrative Stuff

Last Reading Assignment!

- Four main steps for the reading assignments
  - Monday 8pm: Official start of the assignment
  - **Wednesday 8pm: Select your paper**
  - **Friday 8pm:** Post your summary
  - **Monday 8pm:** Post your extra comments (5 posts)

Final Project Report (Due Sunday 12/10 at 8pm)

Main goals:

1. Produce a research paper which will motivate your research problem, describe the prior work, present your research contributions, explain the details of your experiments, and discuss your results.
2. Novel research ideas (N-1 new ideas for N students)
    - Novel algorithm
    - Novel application
3. Incorporate feedback from previous milestones
4. Compare to multimodal baselines from midterm report
    1. Did the proposed ideas solve the errors highlighted in error analysis?
    2. Broader implications of proposed ideas.

Final Poster Presentations (Tuesday 12/5 and Thursday 12/7)

Main objective:

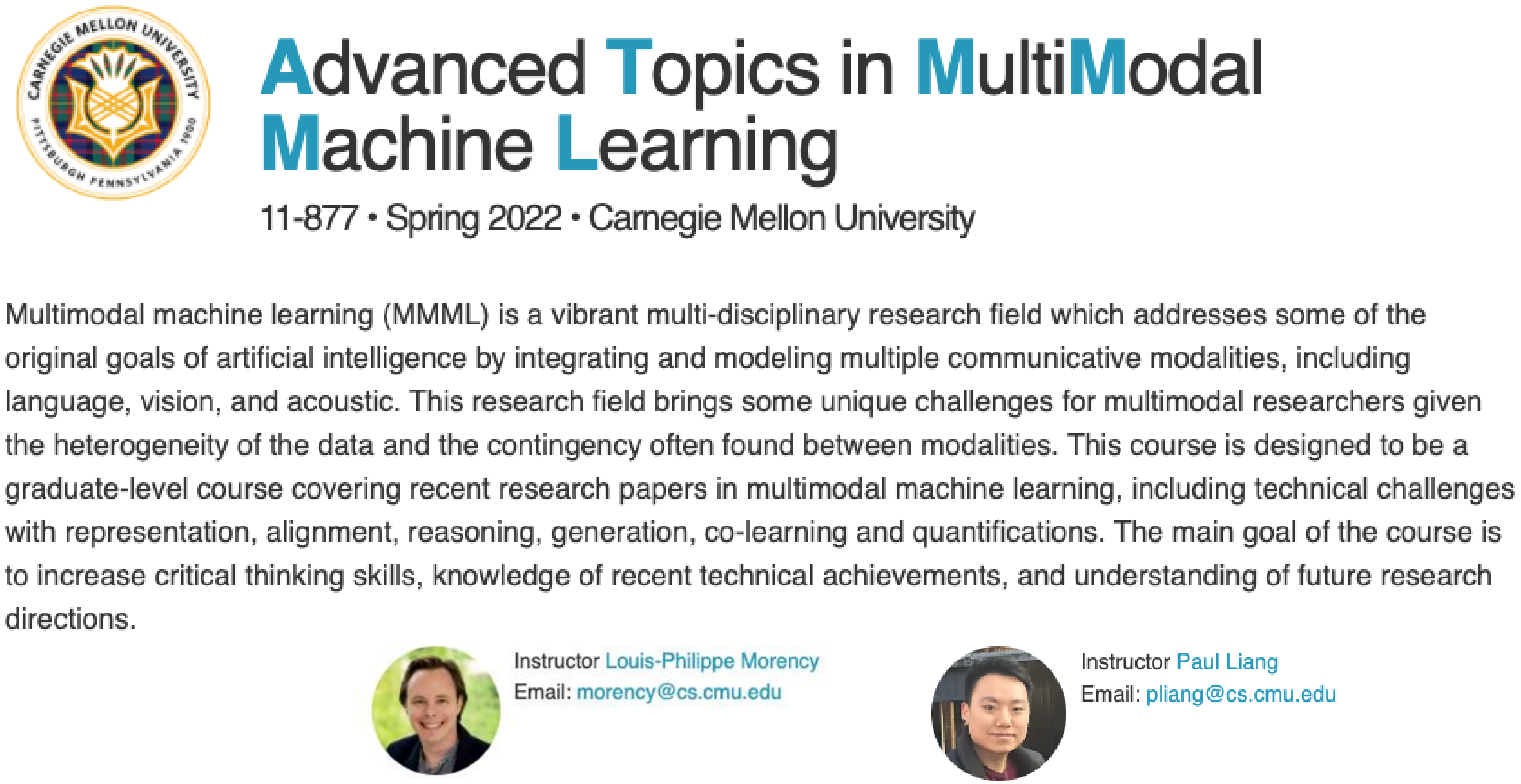
- Focus on only one of your new research ideas
- All students should present and answer questions
- Be sure to be on time! We have many presentations each day ☺
- All presentations are in person (no remote presentations)

Presentation length:

- 30-seconds elevator pitch
- 4-minute full presentation – all students should present

- Following each presentation, audience can ask questions

11-877 Next Semester!



Advanced Topics in MultiModal Machine Learning

11-877 • Spring 2022 • Carnegie Mellon University

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including language, vision, and acoustic. This research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities. This course is designed to be a graduate-level course covering recent research papers in multimodal machine learning, including technical challenges with representation, alignment, reasoning, generation, co-learning and quantifications. The main goal of the course is to increase critical thinking skills, knowledge of recent technical achievements, and understanding of future research directions.

Instructor Louis-Philippe Morency
Email: morency@cs.cmu.edu

Instructor Paul Liang
Email: pliang@cs.cmu.edu

https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

Do you want to be TA for Multimodal course?
Email me!

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.1: New Research Directions

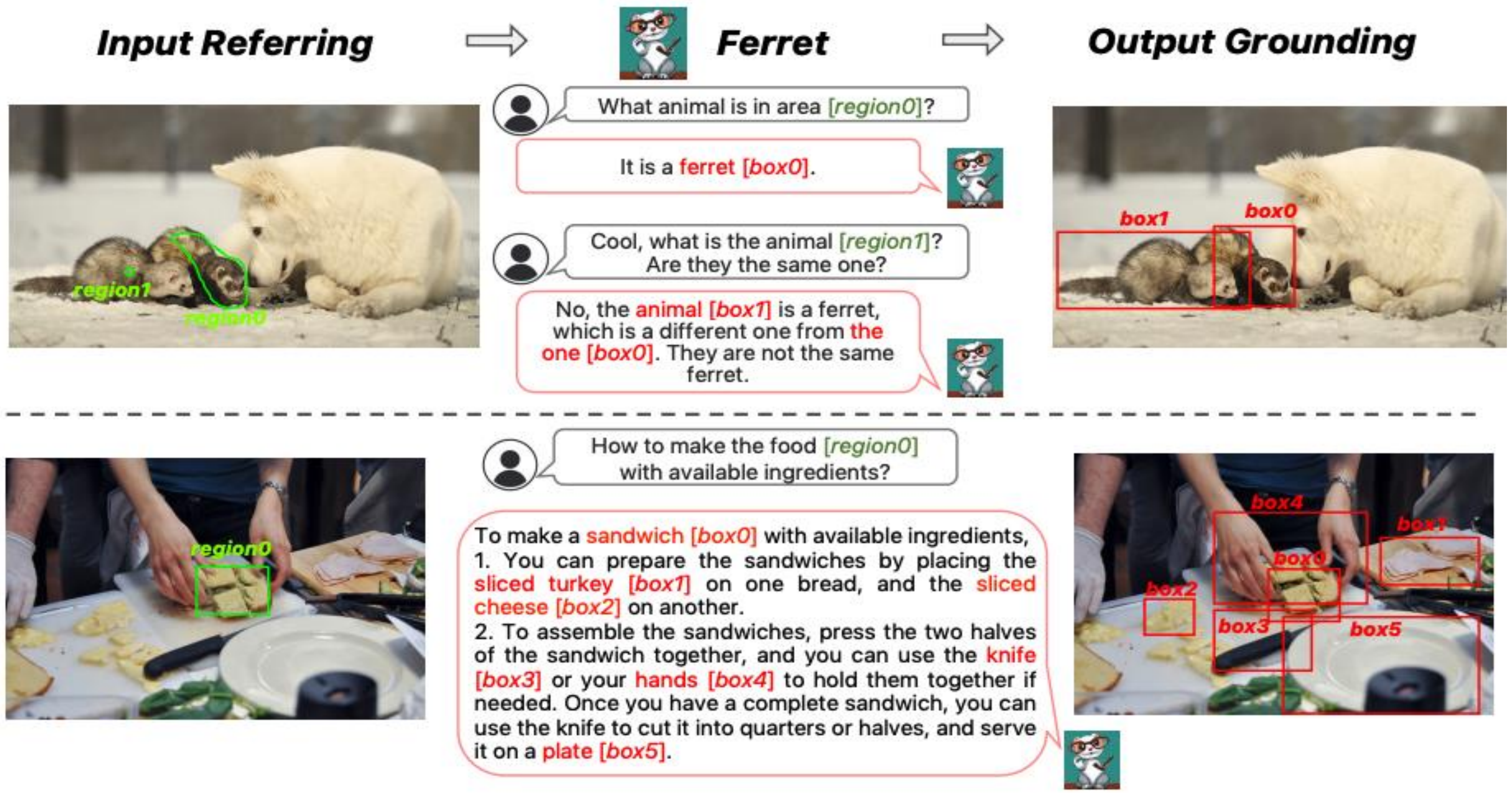
Mehul Agarwal, Soham Dinesh Tiwari, Haofei Yu and LP Morency

** Fall 2021, 2022 and 2023 co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk. Spring 2023 edition taught by Yonatan and Daniel Fried*

Multimodal Grounding

FERRET: Refer and Ground Anything Anywhere at Any Granularity

Can we point to some part of the image and ask typical questions?
Yes! FERRET allows multimodal models to understand free-form input referring and generate grounded output



What animal is in area [region0]?

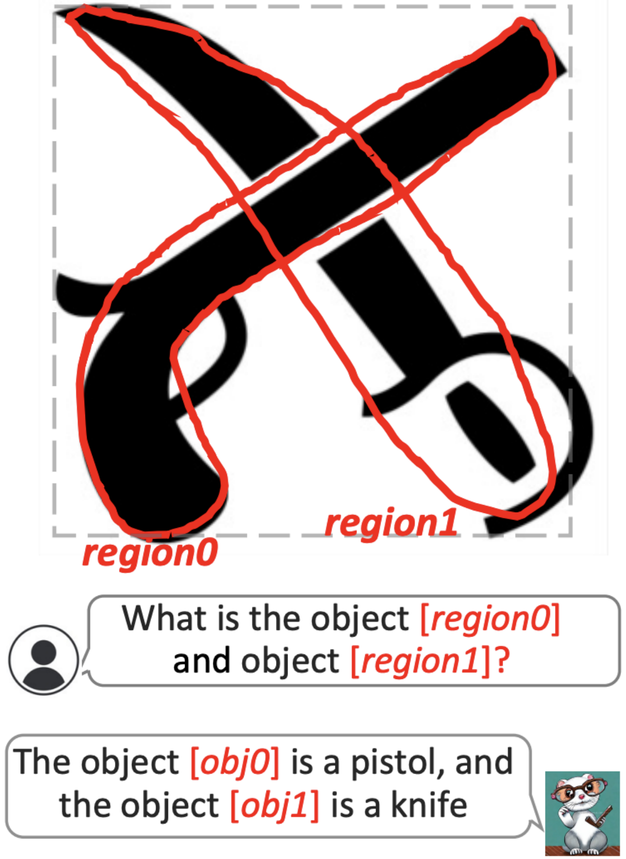
It is a ferret [box0].

Cool, what is the animal [region1]? Are they the same one?

No, the animal [box1] is a ferret, which is a different one from the one [box0]. They are not the same ferret.

How to make the food [region0] with available ingredients?

To make a sandwich [box0] with available ingredients,
1. You can prepare the sandwiches by placing the sliced turkey [box1] on one bread, and the sliced cheese [box2] on another.
2. To assemble the sandwiches, press the two halves of the sandwich together, and you can use the knife [box3] or your hands [box4] to hold them together if needed. Once you have a complete sandwich, you can use the knife to cut it into quarters or halves, and serve it on a plate [box5].

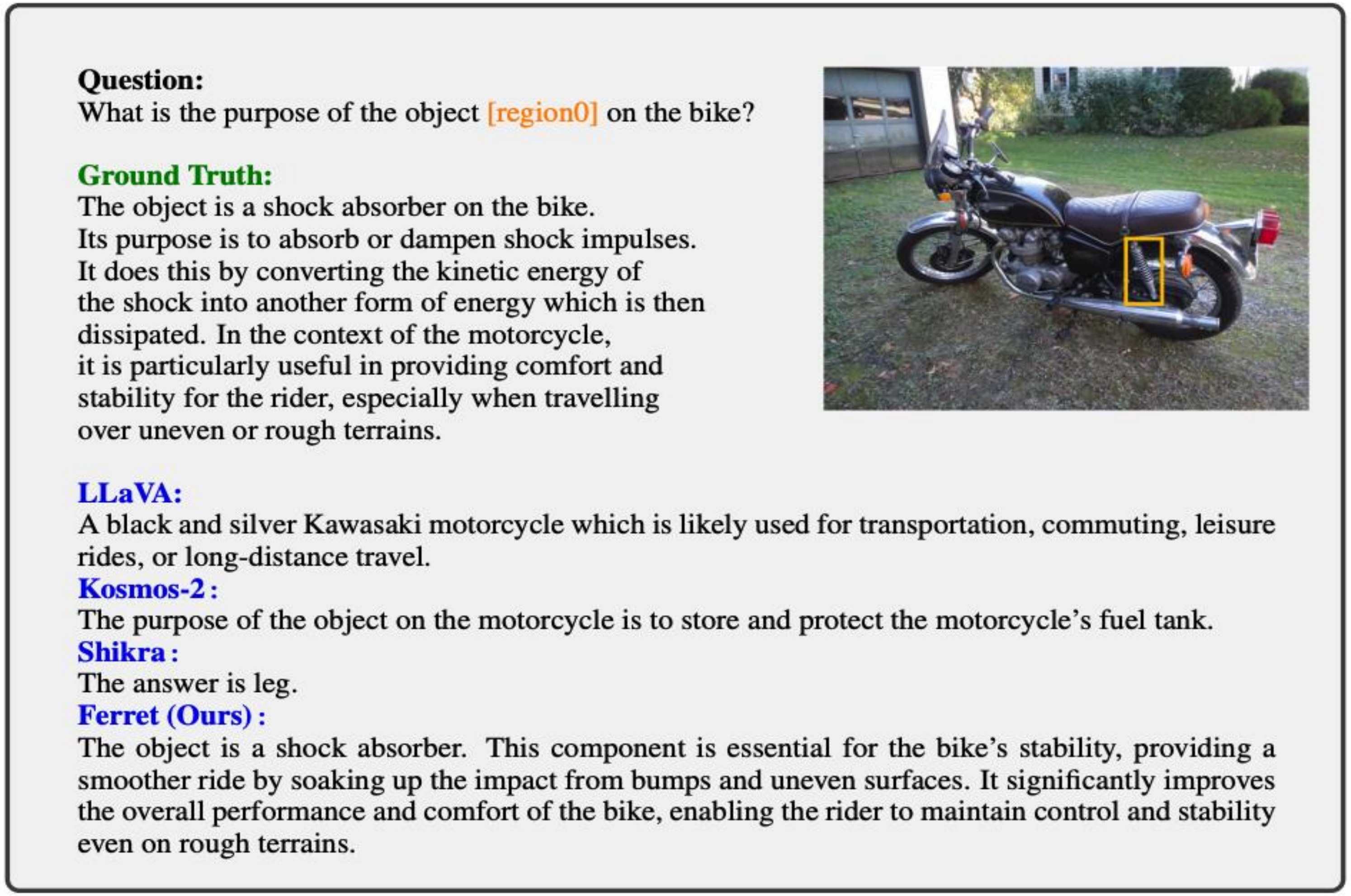


What is the object [region0] and object [region1]?

The object [obj0] is a pistol, and the object [obj1] is a knife

FERRET: Refer and Ground Anything Anywhere at Any Granularity

Currently, other multimodal model has no good ability to understand free-form referring.



Question:
What is the purpose of the object [region0] on the bike?

Ground Truth:
The object is a shock absorber on the bike. Its purpose is to absorb or dampen shock impulses. It does this by converting the kinetic energy of the shock into another form of energy which is then dissipated. In the context of the motorcycle, it is particularly useful in providing comfort and stability for the rider, especially when travelling over uneven or rough terrains.

LLaVA:
A black and silver Kawasaki motorcycle which is likely used for transportation, commuting, leisure rides, or long-distance travel.

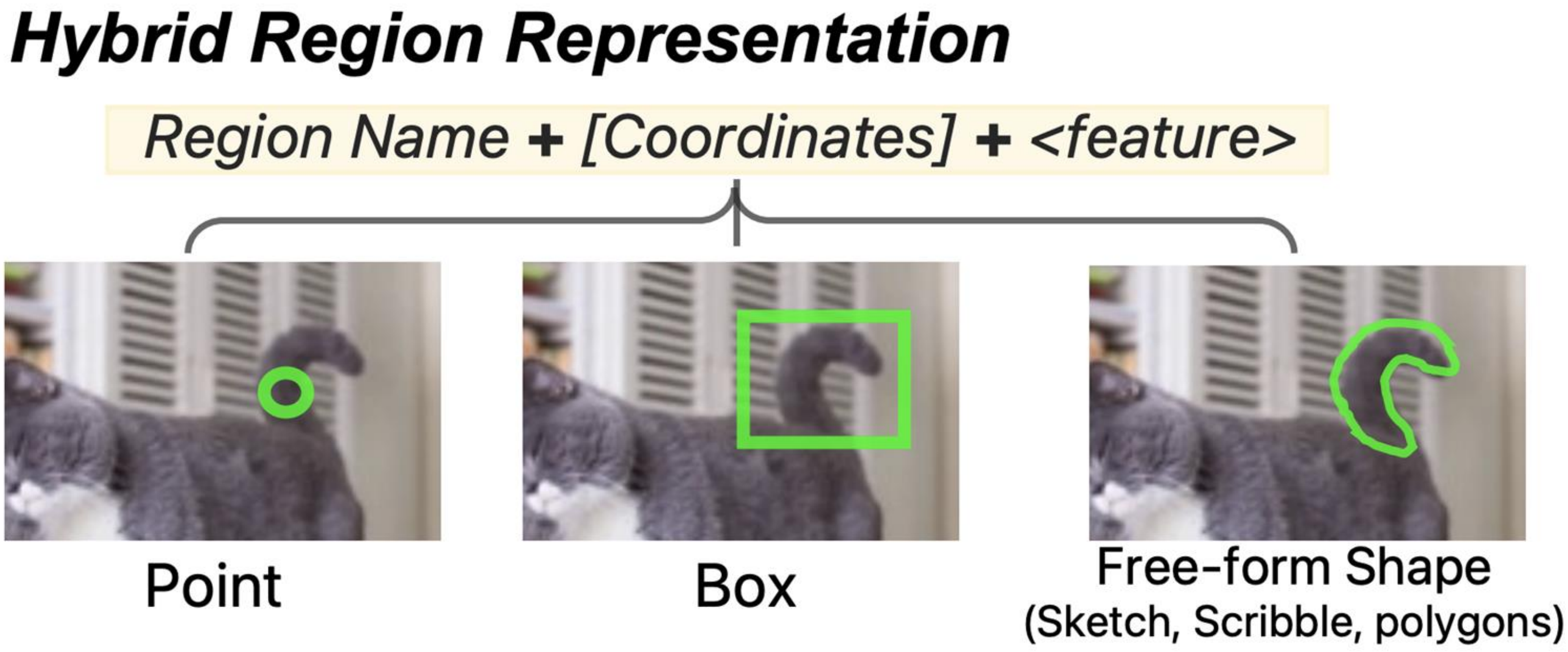
Kosmos-2 :
The purpose of the object on the motorcycle is to store and protect the motorcycle's fuel tank.

Shikra :
The answer is leg.

Ferret (Ours) :
The object is a shock absorber. This component is essential for the bike's stability, providing a smoother ride by soaking up the impact from bumps and uneven surfaces. It significantly improves the overall performance and comfort of the bike, enabling the rider to maintain control and stability even on rough terrains.

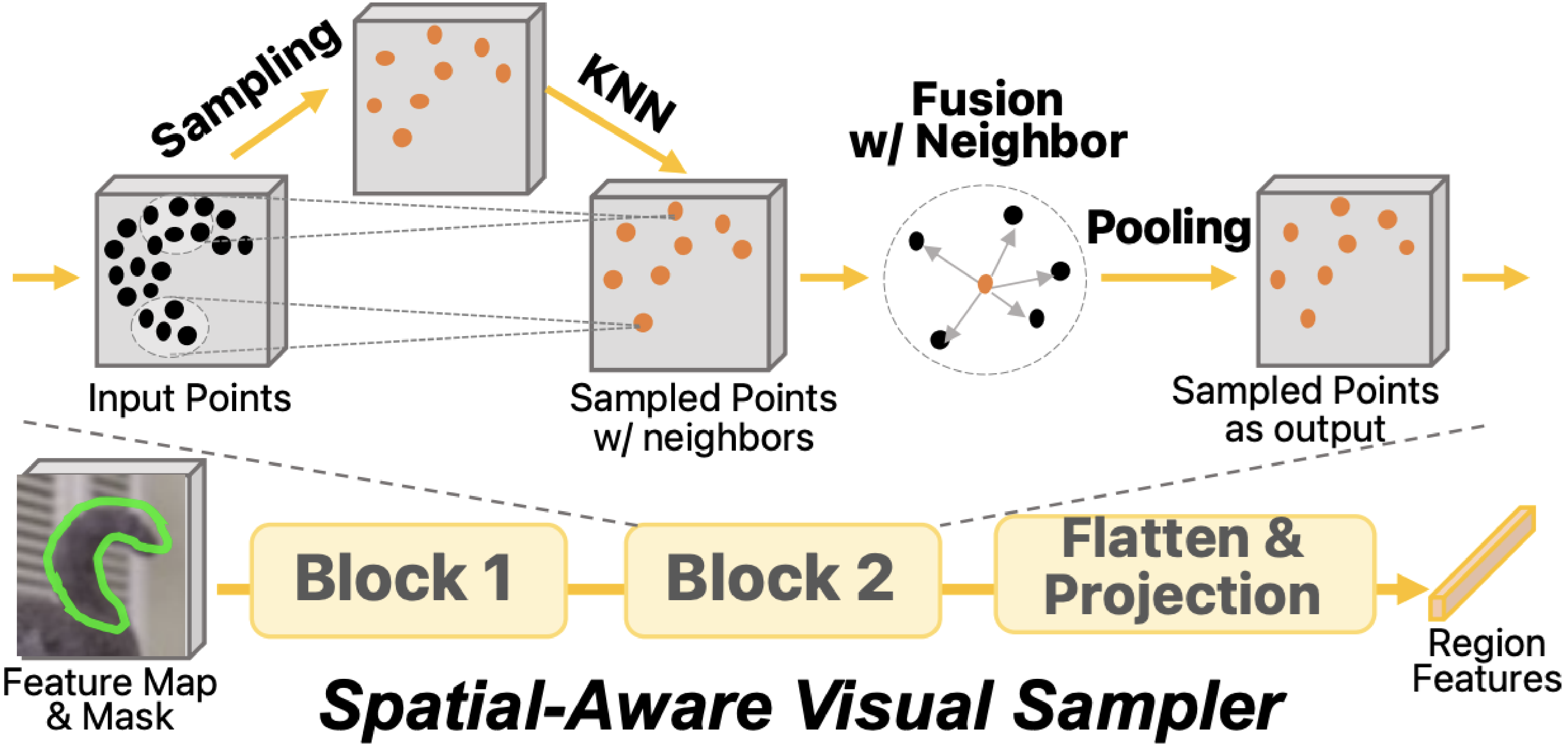
FERRET: Refer and Ground Anything Anywhere at Any Granularity

How to design unified representations for three types of regions: point / box / free-form shape


Hybrid Region Representation
Region Name + [Coordinates] + <feature>
Point
Box
Free-form Shape
(Sketch, Scribble, polygons)

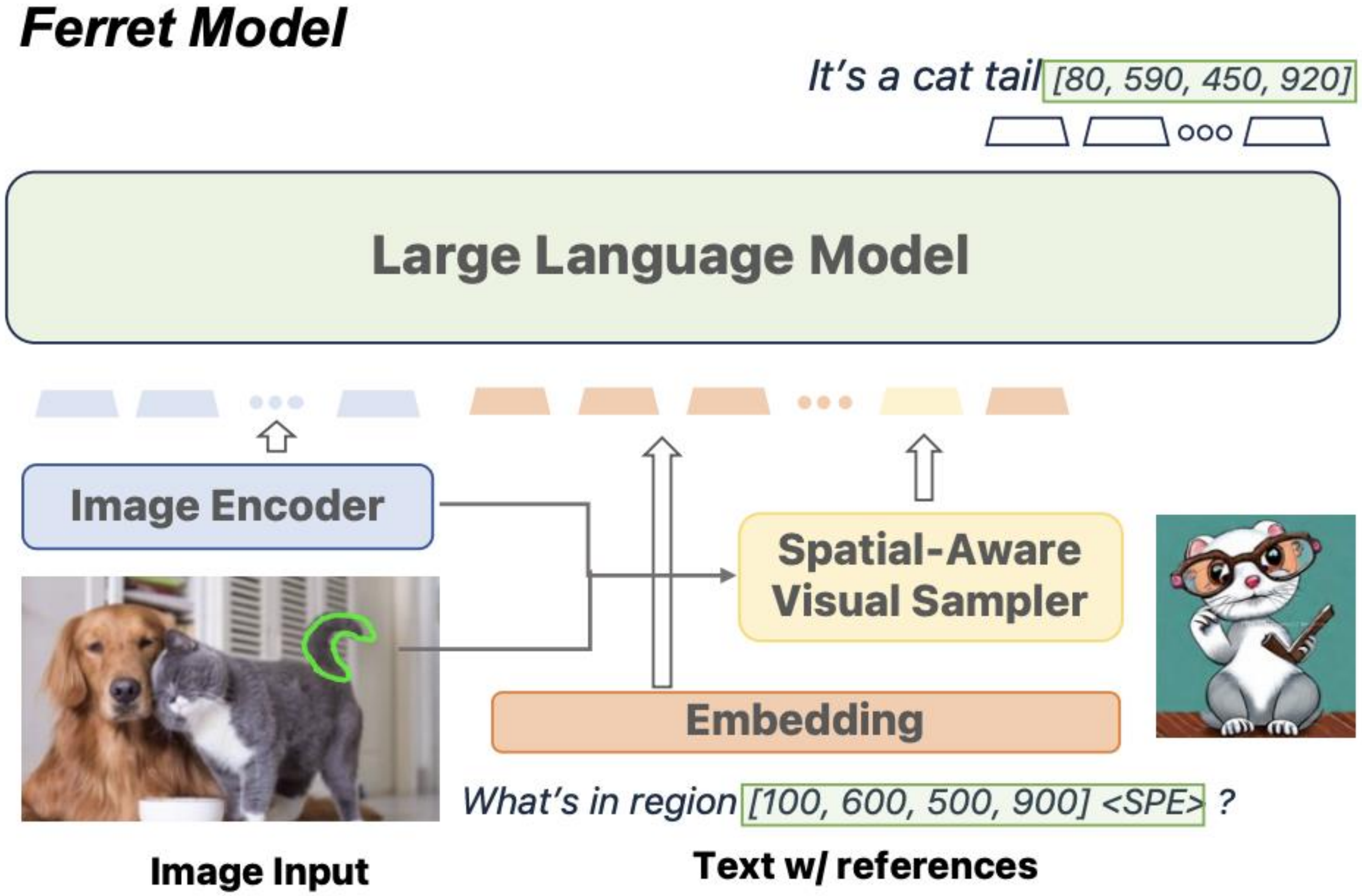
FERRET: Refer and Ground Anything Anywhere at Any Granularity

For the feature part of the free-form regions, use spatial-aware visual sampler to extract features.



FERRET: Refer and Ground Anything Anywhere at Any Granularity

Now the hardest part has been solved. How to train a LLM-based model to learn from those representations?


Ferret Model
It's a cat tail [80, 590, 450, 920]
Large Language Model
Image Encoder
Spatial-Aware Visual Sampler
Embedding
What's in region [100, 600, 500, 900] <SPE> ?
Image Input
Text w/ references

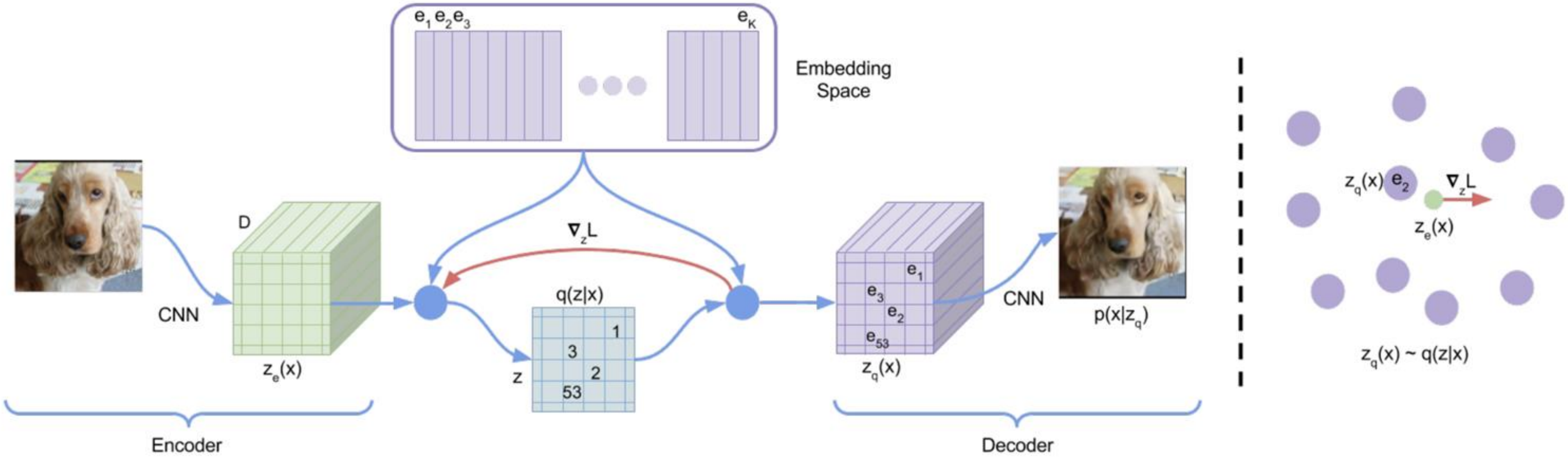
Multimodal Generation

SPAE: Semantic Pyramid AutoEncoder for Multimodal Generation with Frozen LLMs

Motivation for VQ-VAE:
Reduce the dimension size and do auto-regressive PixelCNN generation

Important points about VQ-VAE:

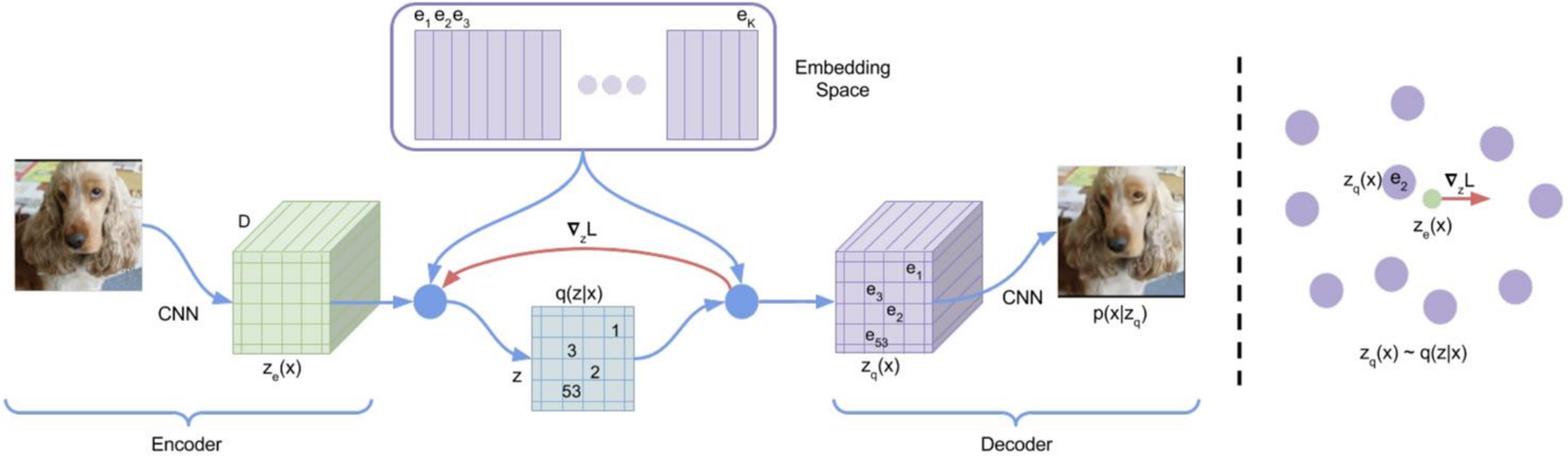
1. It is actually an auto-encoder
2. It uses discrete quantizer.



SPAE: Semantic Pyramid AutoEncoder for Multimodal Generation with Frozen LLMs

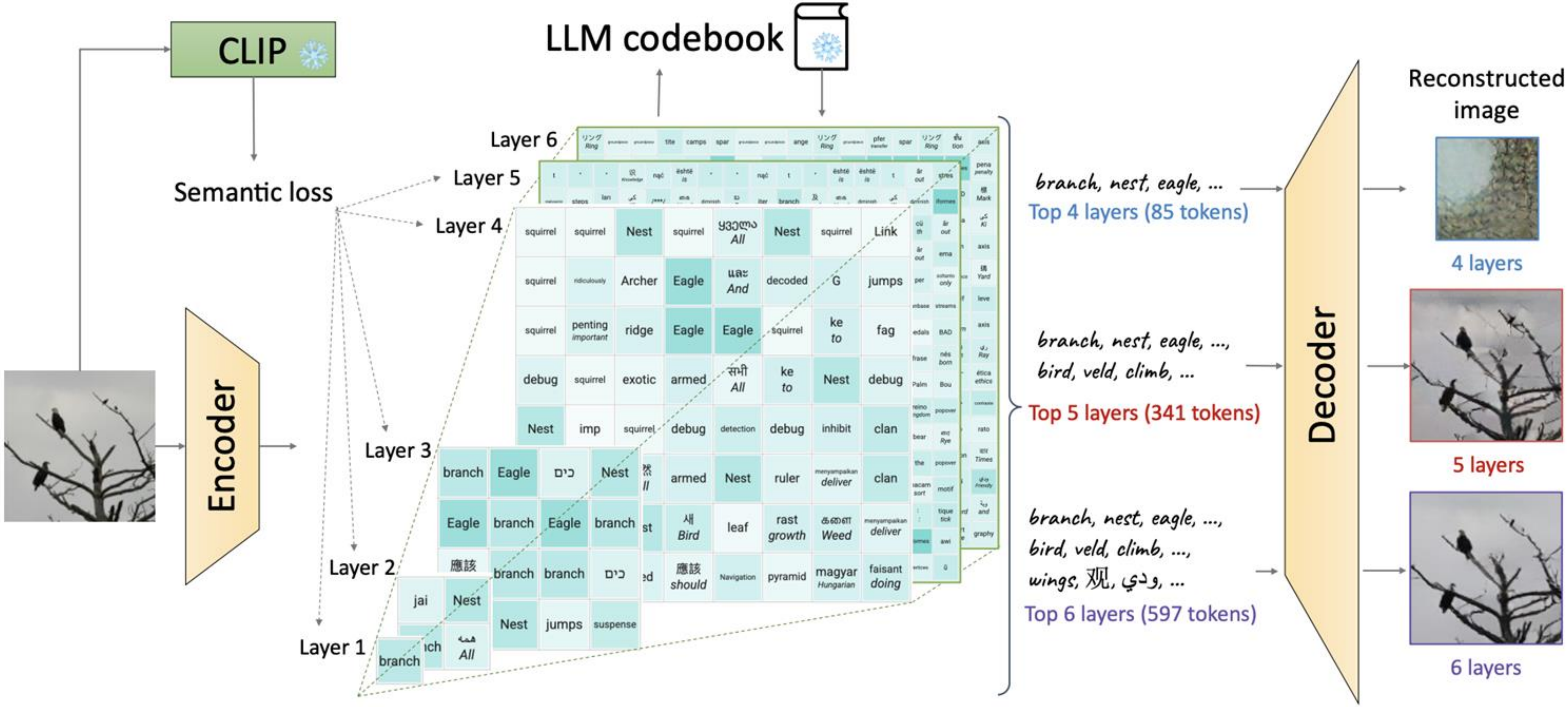
Codebook from VQ-VAE:

It pre-defines an embedding space for quantization with size K.
Discret quantization is a must for PixelCNN to generate the final output.



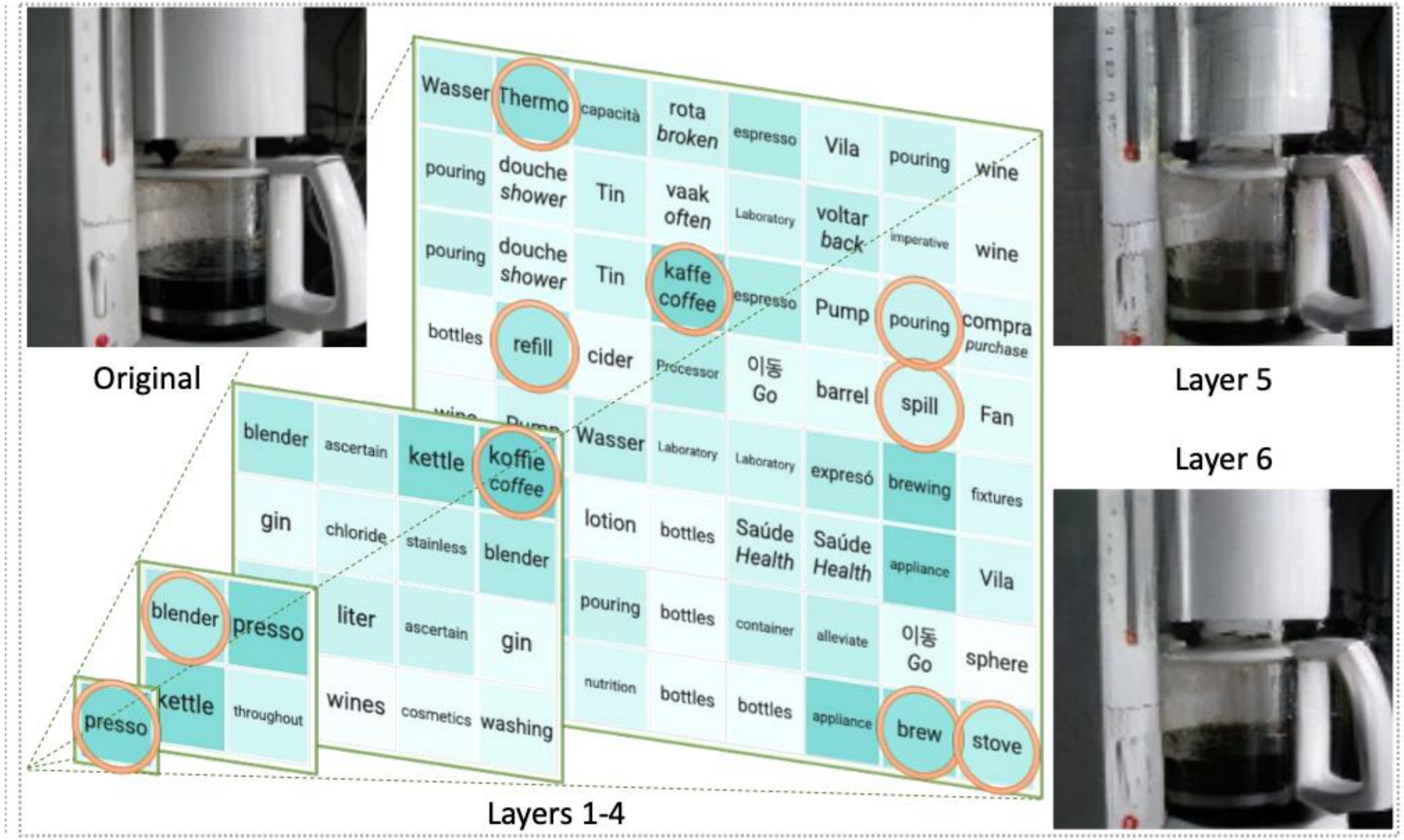
SPAE: Semantic Pyramid AutoEncoder for Multimodal Generation with Frozen LLMs

SPAE extends based on VQ-VAE to use a frozen LLM to be the quantizer.



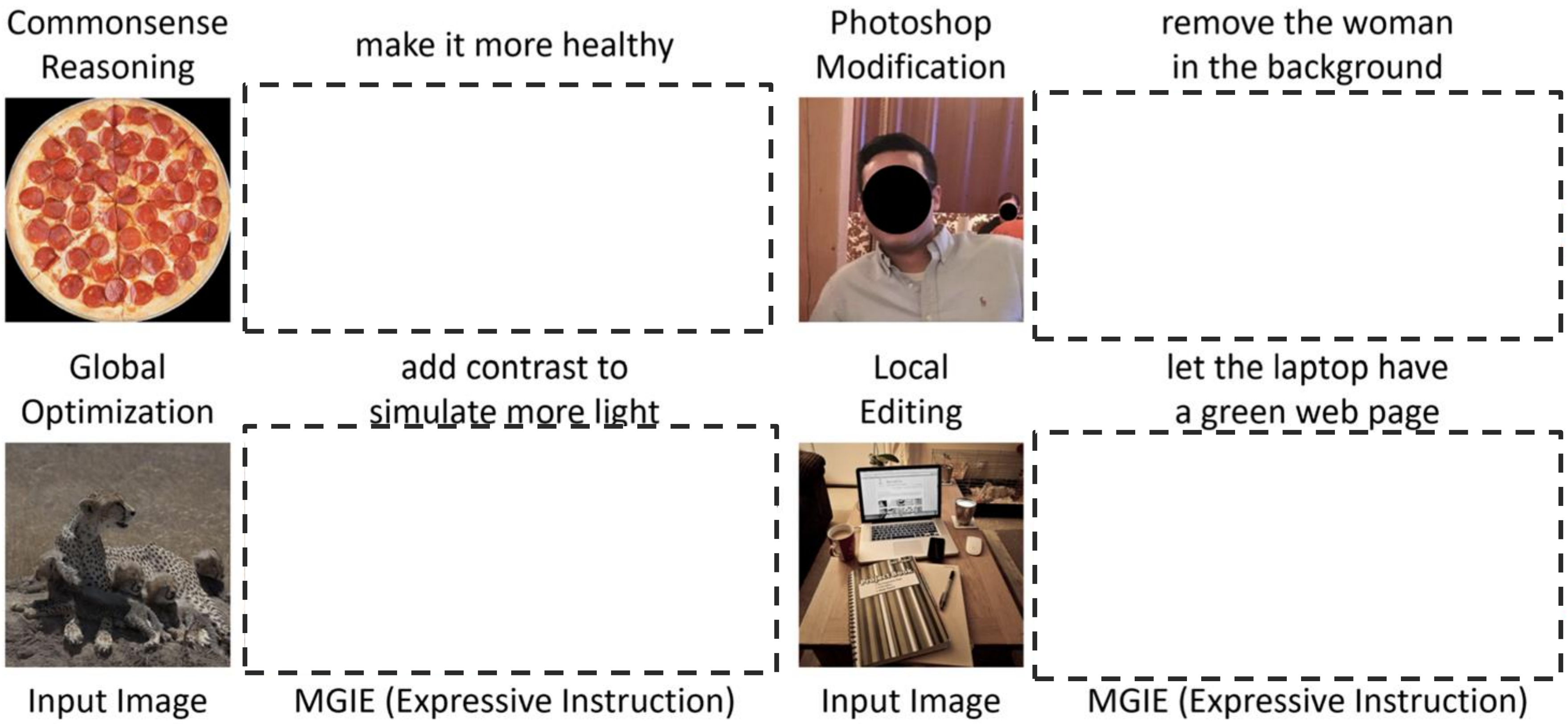
SPAE: Semantic Pyramid AutoEncoder for Multimodal Generation with Frozen LLMs

Semantic Pyramid AutoEncoder: allows for representing semantic concepts with notably fewer tokens


Original
Wasser Thermo capacità rota broken espresso Vila pouring wine
pouring douche shower Tin vaak often Laboratory voltar back imperative wine
pouring douche shower Tin kaffe coffee espresso Pump pouring compra purchase
bottles refill cider Processor 이동 Go barrel spill Fan
wine Pump Wasser Laboratory Laboratory expresó brewing fixtures
lotion bottles Saúde Health Saúde Health appliance Vila
pouring bottles container alleviate 이동 Go sphere
nutrition bottles bottles appliance brew stove
blender ascertain kettle koffie coffee
gin chloride stainless blender
liter ascertain gin
wines cosmetics washing
blender presso
kettle throughout
presso
Layers 1-4
Layer 5
Layer 6

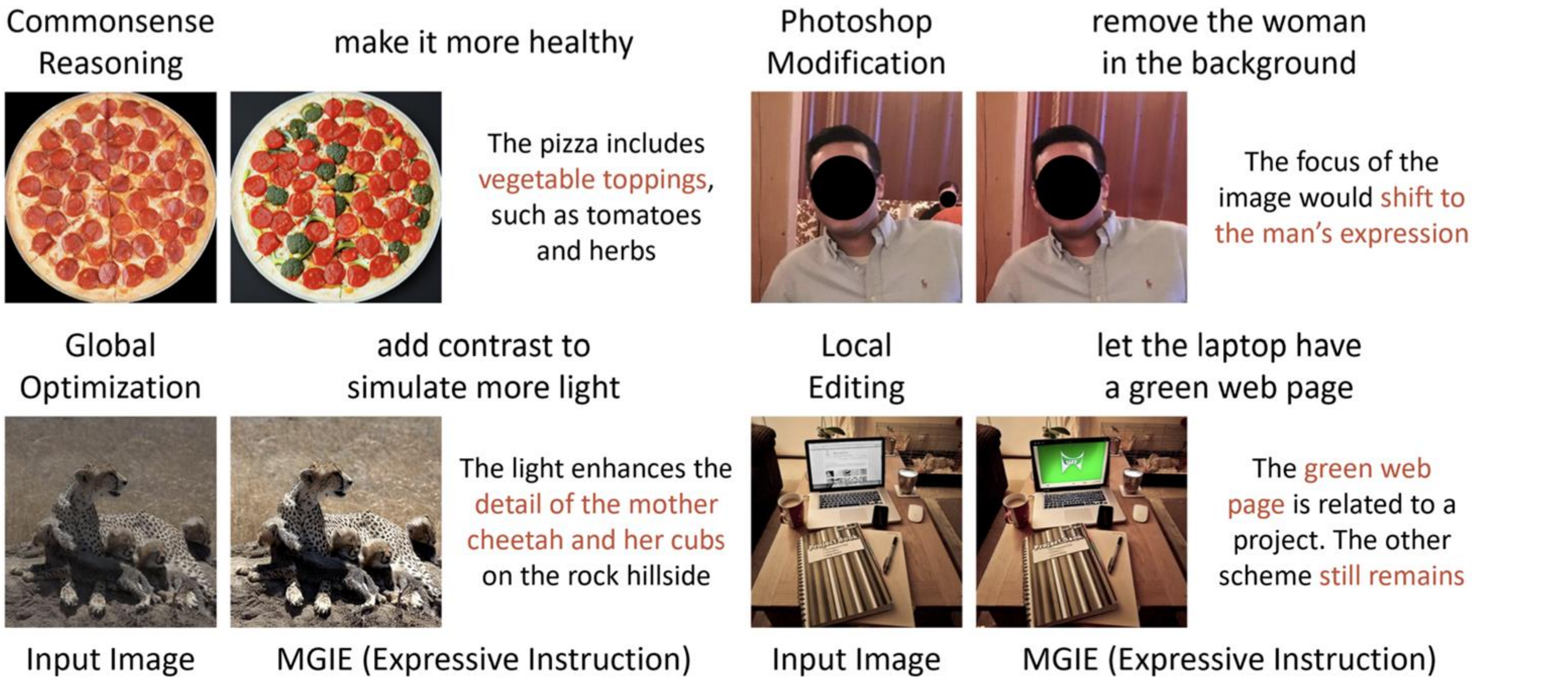
MLLM-guided Image Editing

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS


Commonsense Reasoning
make it more healthy
Photoshop Modification
remove the woman in the background
Global Optimization
add contrast to simulate more light
Local Editing
let the laptop have a green web page
Input Image
MGIE (Expressive Instruction)
Input Image
MGIE (Expressive Instruction)

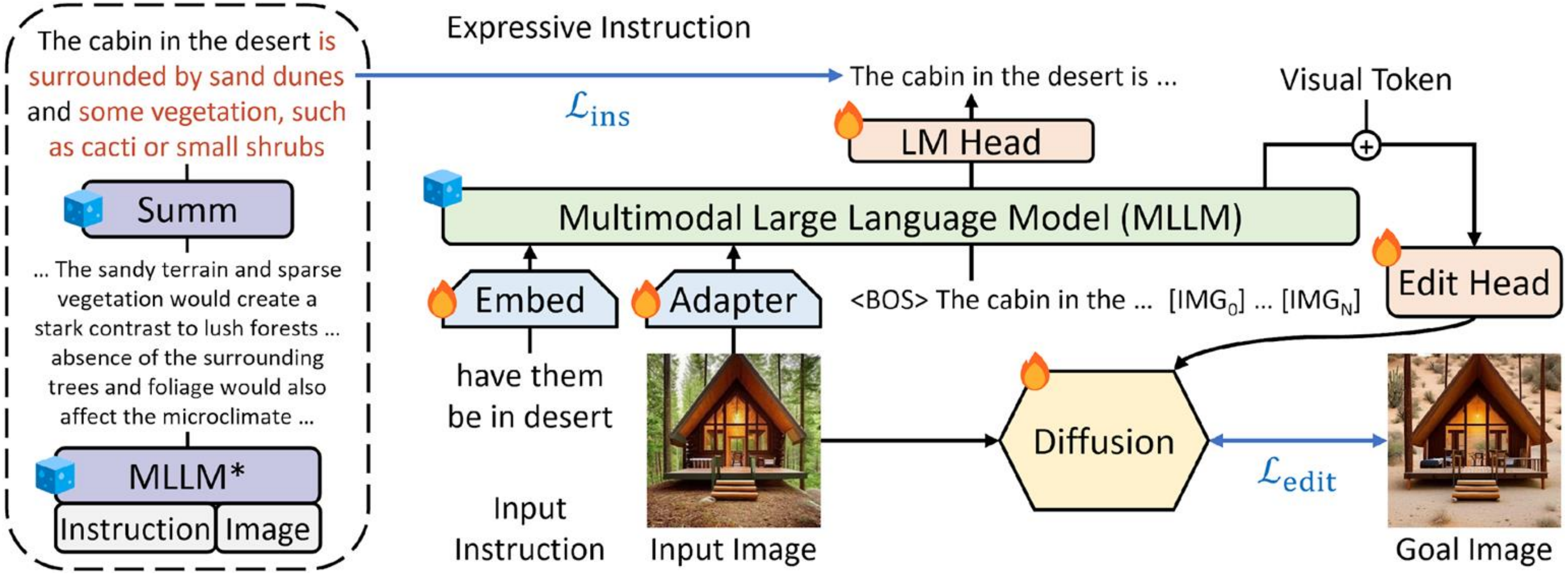
Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS


Commonsense Reasoning
make it more healthy
The pizza includes vegetable toppings, such as tomatoes and herbs
Photoshop Modification
remove the woman in the background
The focus of the image would shift to the man's expression
Global Optimization
add contrast to simulate more light
The light enhances the detail of the mother cheetah and her cubs on the rock hillside
Local Editing
let the laptop have a green web page
The green web page is related to a project. The other scheme still remains
Input Image
MGIE (Expressive Instruction)
Input Image
MGIE (Expressive Instruction)

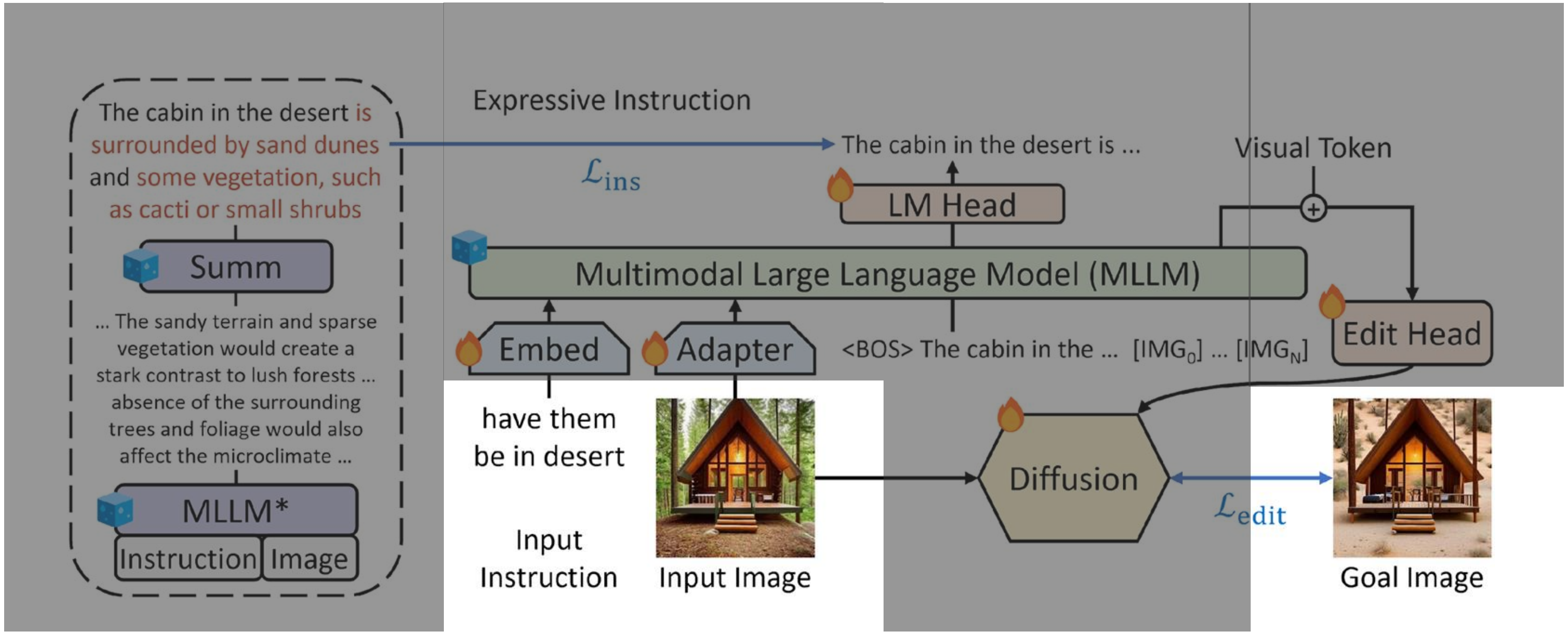
Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS


The cabin in the desert is surrounded by sand dunes and some vegetation, such as cacti or small shrubs
Summ
... The sandy terrain and sparse vegetation would create a stark contrast to lush forests ... absence of the surrounding trees and foliage would also affect the microclimate ...
MLLM*
Instruction
Image
Expressive Instruction
$\mathcal{L}_{\text{ins}}$
The cabin in the desert is ...
Visual Token
LM Head
Multimodal Large Language Model (MLLM)
Embed
Adapter
<BOS> The cabin in the ... [IMG$_0$] ... [IMG$_N$]
Edit Head
have them be in desert
Input Instruction
Input Image
Diffusion
$\mathcal{L}_{\text{edit}}$
Goal Image

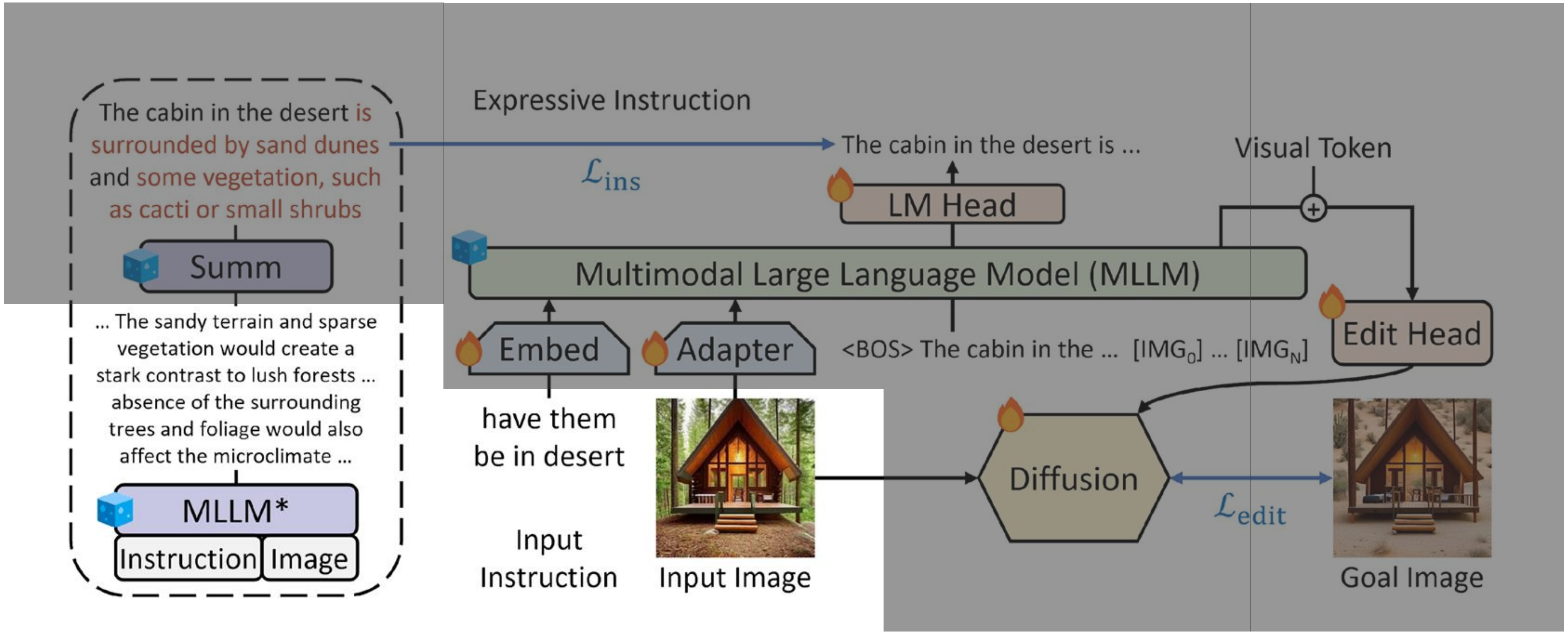
Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS


The cabin in the desert is surrounded by sand dunes and some vegetation, such as cacti or small shrubs
Summ
... The sandy terrain and sparse vegetation would create a stark contrast to lush forests ... absence of the surrounding trees and foliage would also affect the microclimate ...
MLLM*
Instruction
Image
Expressive Instruction
$\mathcal{L}_{\text{ins}}$
The cabin in the desert is ...
LM Head
Visual Token
Multimodal Large Language Model (MLLM)
Embed
Adapter
<BOS> The cabin in the ... [IMG$_0$] ... [IMG$_N$]
Edit Head
have them be in desert
Input Instruction
Input Image
Diffusion
$\mathcal{L}_{\text{edit}}$
Goal Image

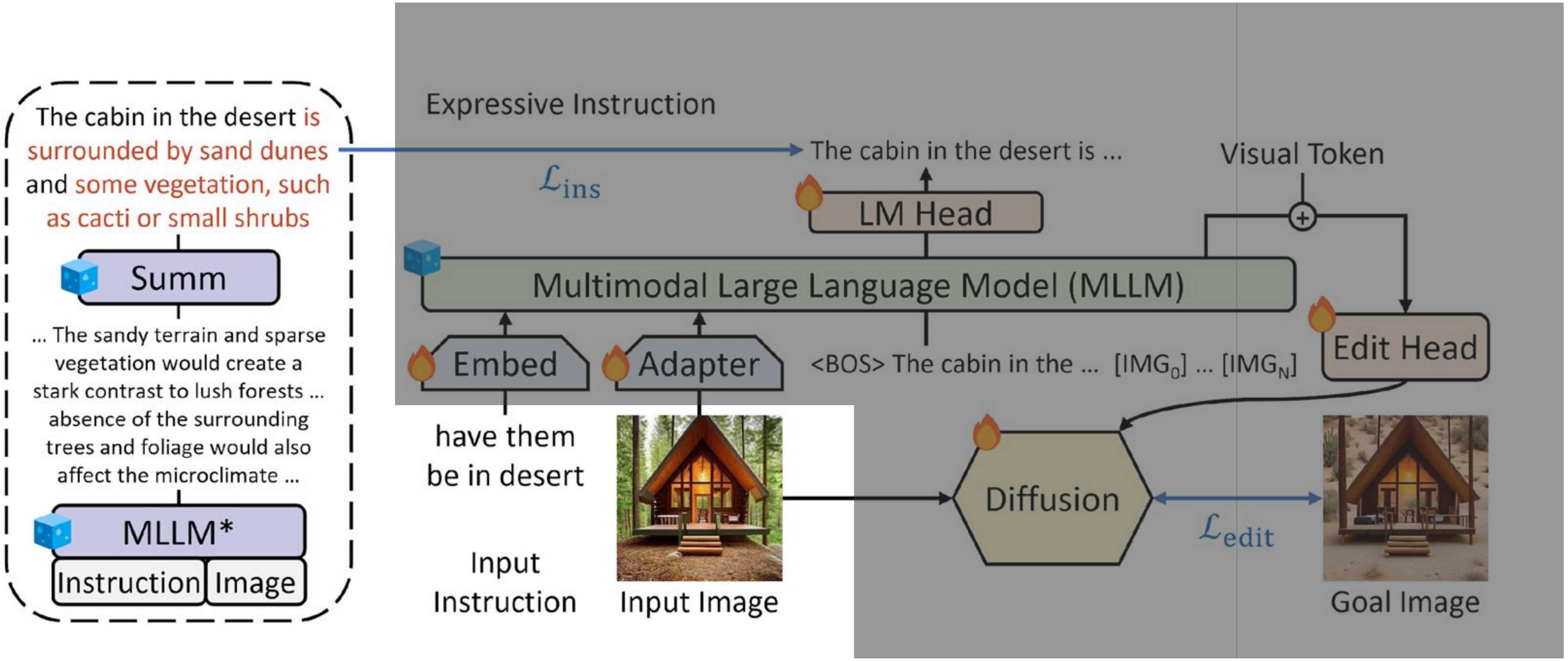
Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS


The cabin in the desert is surrounded by sand dunes and some vegetation, such as cacti or small shrubs
Summ
... The sandy terrain and sparse vegetation would create a stark contrast to lush forests ... absence of the surrounding trees and foliage would also affect the microclimate ...
MLLM*
Instruction
Image
Expressive Instruction
$\mathcal{L}_{\text{ins}}$
The cabin in the desert is ...
LM Head
Visual Token
Multimodal Large Language Model (MLLM)
Embed
Adapter
<BOS> The cabin in the ... [IMG$_0$] ... [IMG$_N$]
Edit Head
have them be in desert
Input Instruction
Input Image
Diffusion
$\mathcal{L}_{\text{edit}}$
Goal Image

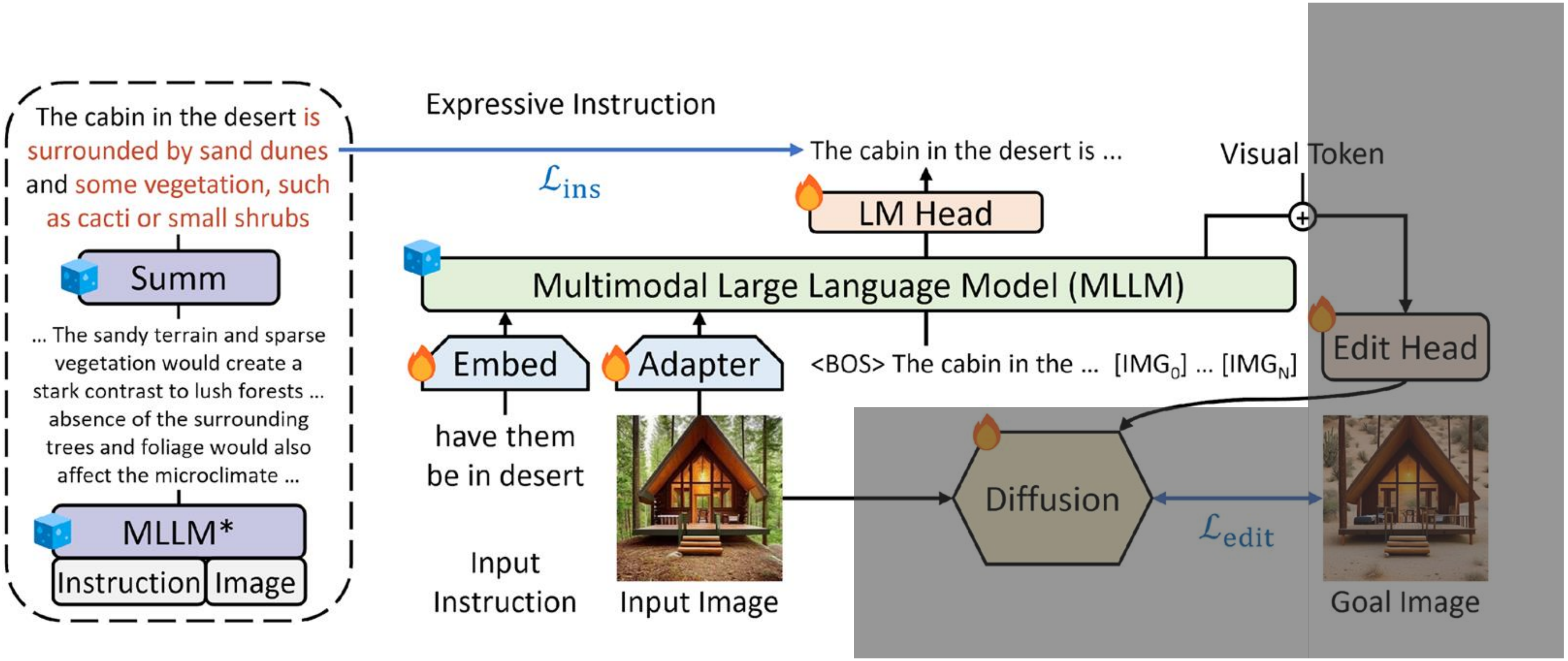
Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS


The cabin in the desert is surrounded by sand dunes and some vegetation, such as cacti or small shrubs
Summ
... The sandy terrain and sparse vegetation would create a stark contrast to lush forests ... absence of the surrounding trees and foliage would also affect the microclimate ...
MLLM*
Instruction
Image
Expressive Instruction
$\mathcal{L}_{\text{ins}}$
The cabin in the desert is ...
LM Head
Visual Token
Multimodal Large Language Model (MLLM)
Embed
Adapter
<BOS> The cabin in the ... [IMG$_0$] ... [IMG$_N$]
Edit Head
have them be in desert
Input Instruction
Input Image
Diffusion
$\mathcal{L}_{\text{edit}}$
Goal Image

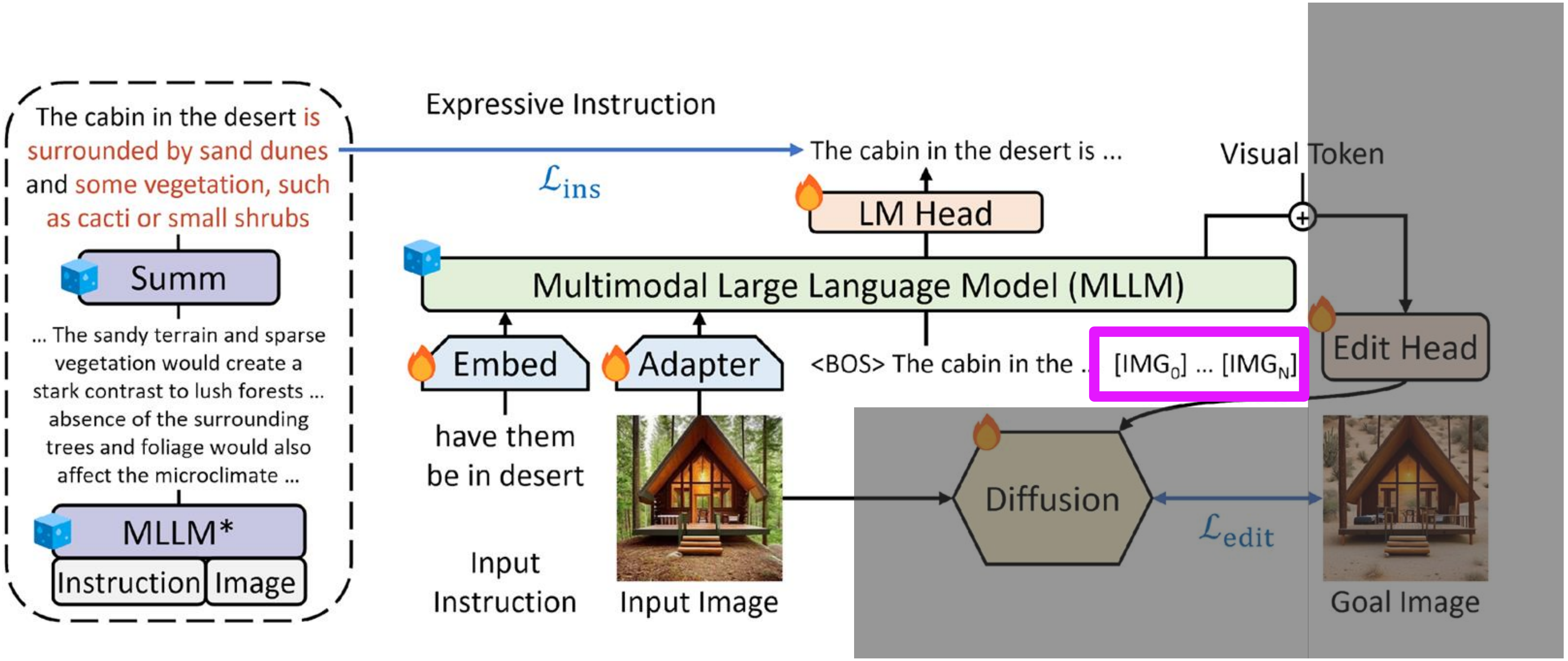
Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS


The cabin in the desert is surrounded by sand dunes and some vegetation, such as cacti or small shrubs
Summ
... The sandy terrain and sparse vegetation would create a stark contrast to lush forests ... absence of the surrounding trees and foliage would also affect the microclimate ...
MLLM*
Instruction
Image
Expressive Instruction
$\mathcal{L}_{ins}$
The cabin in the desert is ...
LM Head
Visual Token
Multimodal Large Language Model (MLLM)
Embed
Adapter
<BOS> The cabin in the ... [IMG$_0$] ... [IMG$_N$]
Edit Head
have them be in desert
Input Instruction
Input Image
Diffusion
$\mathcal{L}_{edit}$
Goal Image

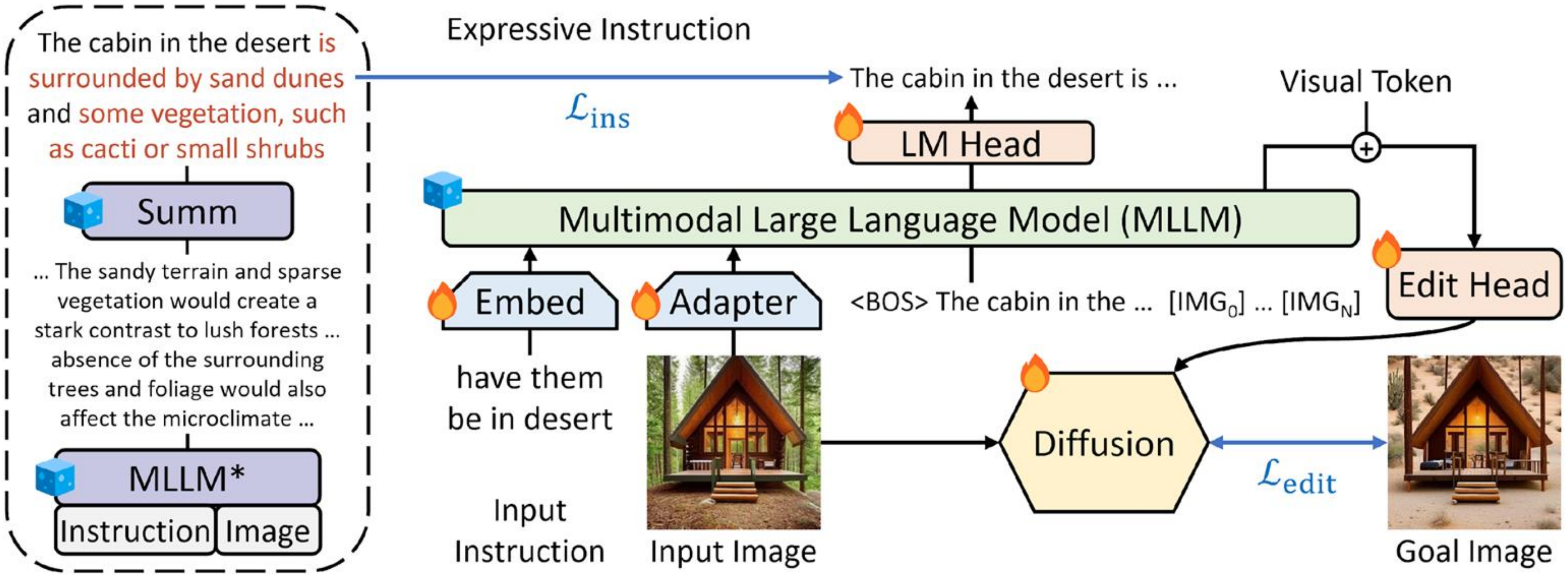
Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS


The cabin in the desert is surrounded by sand dunes and some vegetation, such as cacti or small shrubs
Summ
... The sandy terrain and sparse vegetation would create a stark contrast to lush forests ... absence of the surrounding trees and foliage would also affect the microclimate ...
MLLM*
Instruction
Image
Expressive Instruction
$\mathcal{L}_{ins}$
The cabin in the desert is ...
LM Head
Visual Token
Multimodal Large Language Model (MLLM)
Embed
Adapter
<BOS> The cabin in the ...
[IMG$_0$] ... [IMG$_N$]
Edit Head
have them be in desert
Input Instruction
Input Image
Diffusion
$\mathcal{L}_{edit}$
Goal Image

Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS



The cabin in the desert is surrounded by sand dunes and some vegetation, such as cacti or small shrubs
Summ
... The sandy terrain and sparse vegetation would create a stark contrast to lush forests ... absence of the surrounding trees and foliage would also affect the microclimate ...
MLLM*
Instruction
Image
Expressive Instruction
$\mathcal{L}_{\text{ins}}$
The cabin in the desert is ...
LM Head
Visual Token
Multimodal Large Language Model (MLLM)
Embed
Adapter
<BOS> The cabin in the ... [IMG$_0$] ... [IMG$_N$]
Edit Head
have them be in desert
Input Instruction
Input Image
Diffusion
$\mathcal{L}_{\text{edit}}$
Goal Image

Fu, Tsu-Jui, et al. "Guiding Instruction-based Image Editing via Multimodal Large Language Models." arXiv preprint arXiv:2309.17102 (2023)

LLMs for Speech

SALMONN: TOWARDS GENERIC HEARING ABILITIES FOR LARGE LANGUAGE MODELS

- **SALMONN**, a **S**peech **A**udio **L**anguage **M**usic **O**pen **N**eural **N**etwork
- Integrating a pre-trained text-based large language model (LLM) with speech and audio encoders into a single multimodal model.
- Many prior Audio-Speech-Text LLMs - such as SpeechGPT and AudioPaLM.
- How is SALMONN different?

Tang, Changli, et al. "SALMONN: Towards Generic Hearing Abilities for Large Language Models." arXiv preprint arXiv:2310.13289 (2023).

SALMONN: TOWARDS GENERIC HEARING ABILITIES FOR LARGE LANGUAGE MODELS

How is SALMONN different?

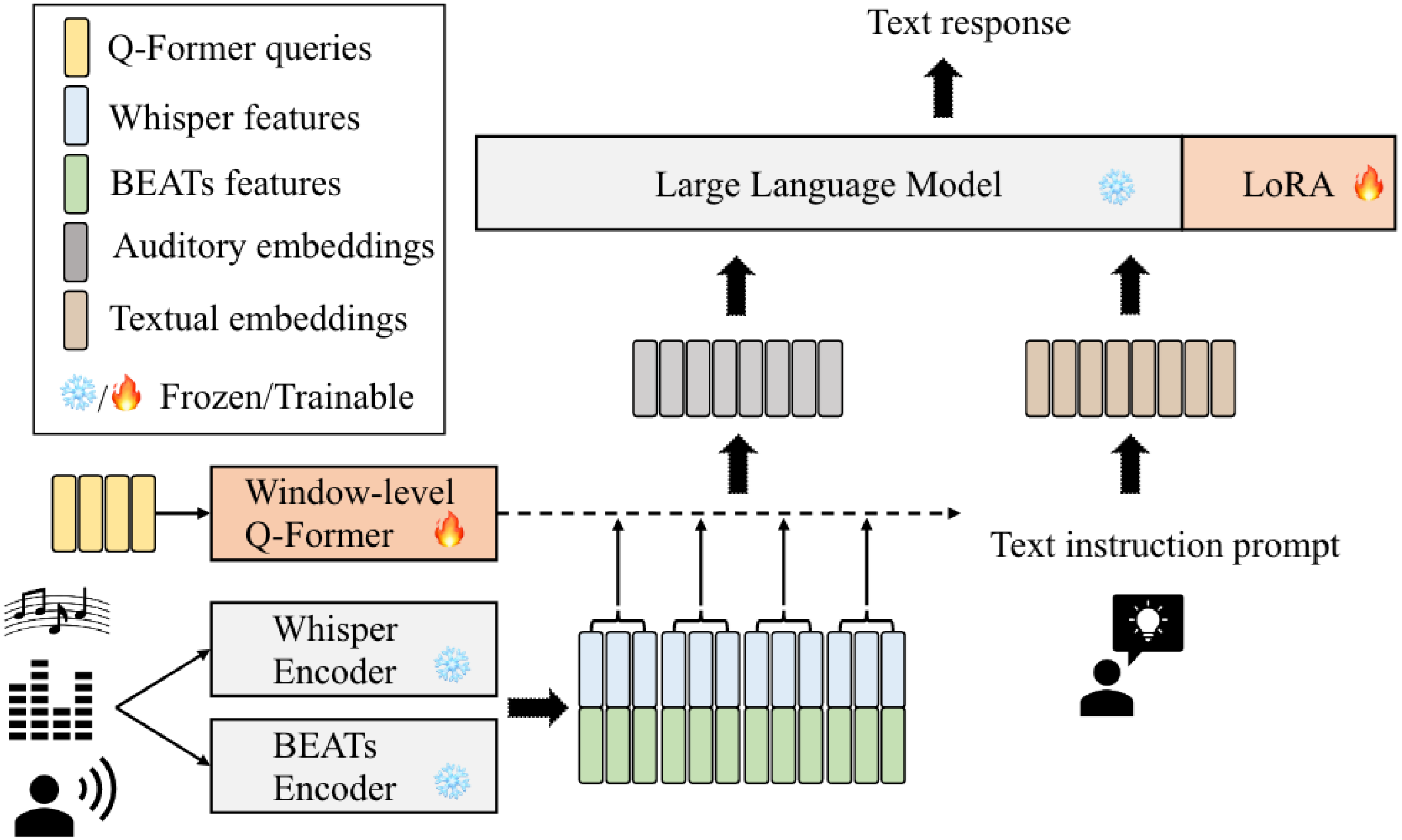
- competitive performances on training tasks
- ASR and translation**,** auditory information-based question answering, emotion recognition, speaker verification, and music and audio captioning etc.

- SALMONN also has diverse **emergent abilities** unseen in training,
- speech translation to untrained languages, speech-based slot filling, **spoken-query-based question answering**, **audio-based storytelling**, and **speech audio co-reasoning** etc.

- A novel few-shot activation, by tuning LoRA scaling factor proposed to activate **cross-modal emergent** abilities of SALMONN.

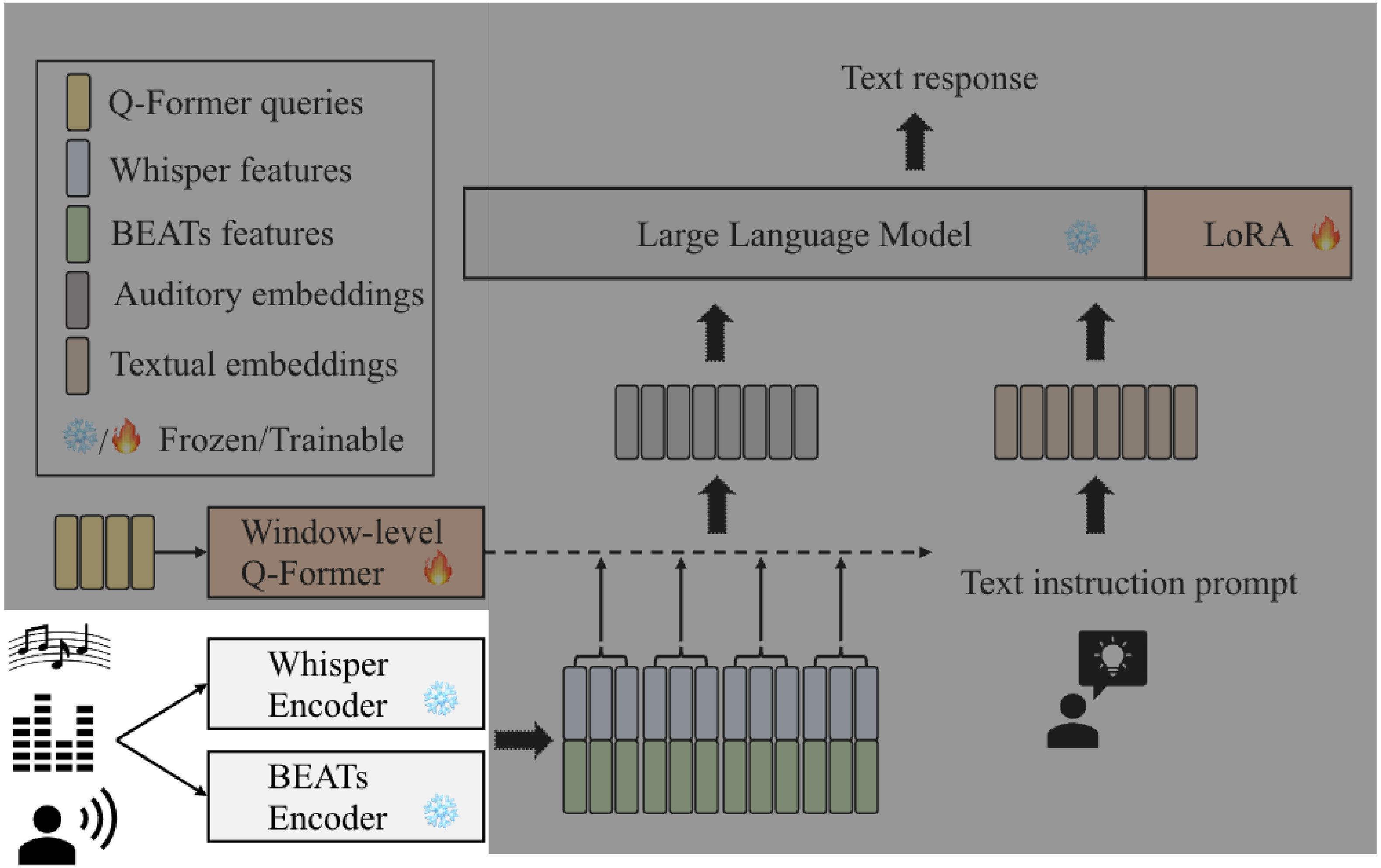
Tang, Changli, et al. "SALMONN: Towards Generic Hearing Abilities for Large Language Models." arXiv preprint arXiv:2310.13289 (2023).

SALMONN: TOWARDS GENERIC HEARING ABILITIES FOR LARGE LANGUAGE MODELS



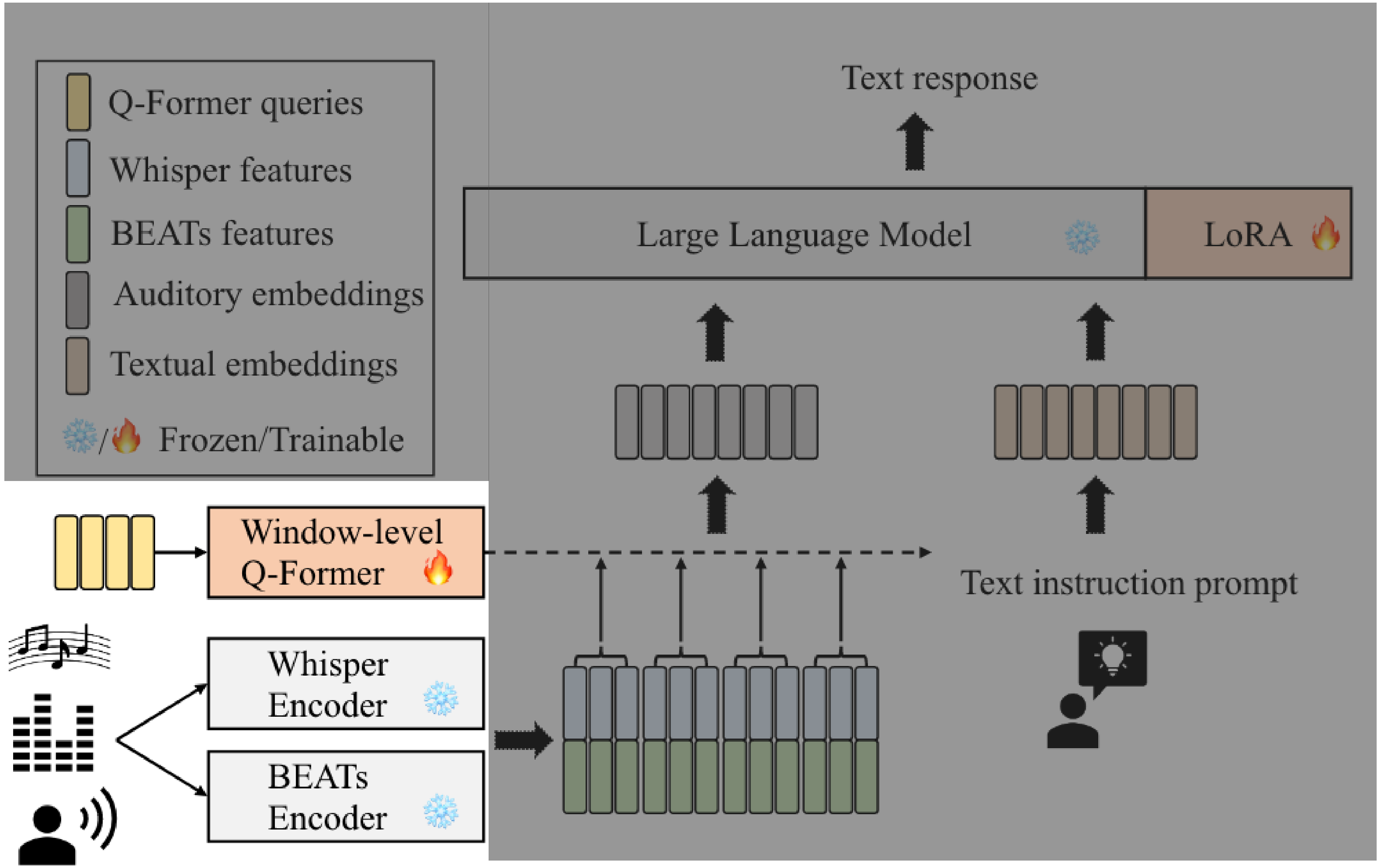
Tang, Changli, et al. "SALMONN: Towards Generic Hearing Abilities for Large Language Models." arXiv preprint arXiv:2310.13289 (2023).

SALMONN: TOWARDS GENERIC HEARING ABILITIES FOR LARGE LANGUAGE MODELS



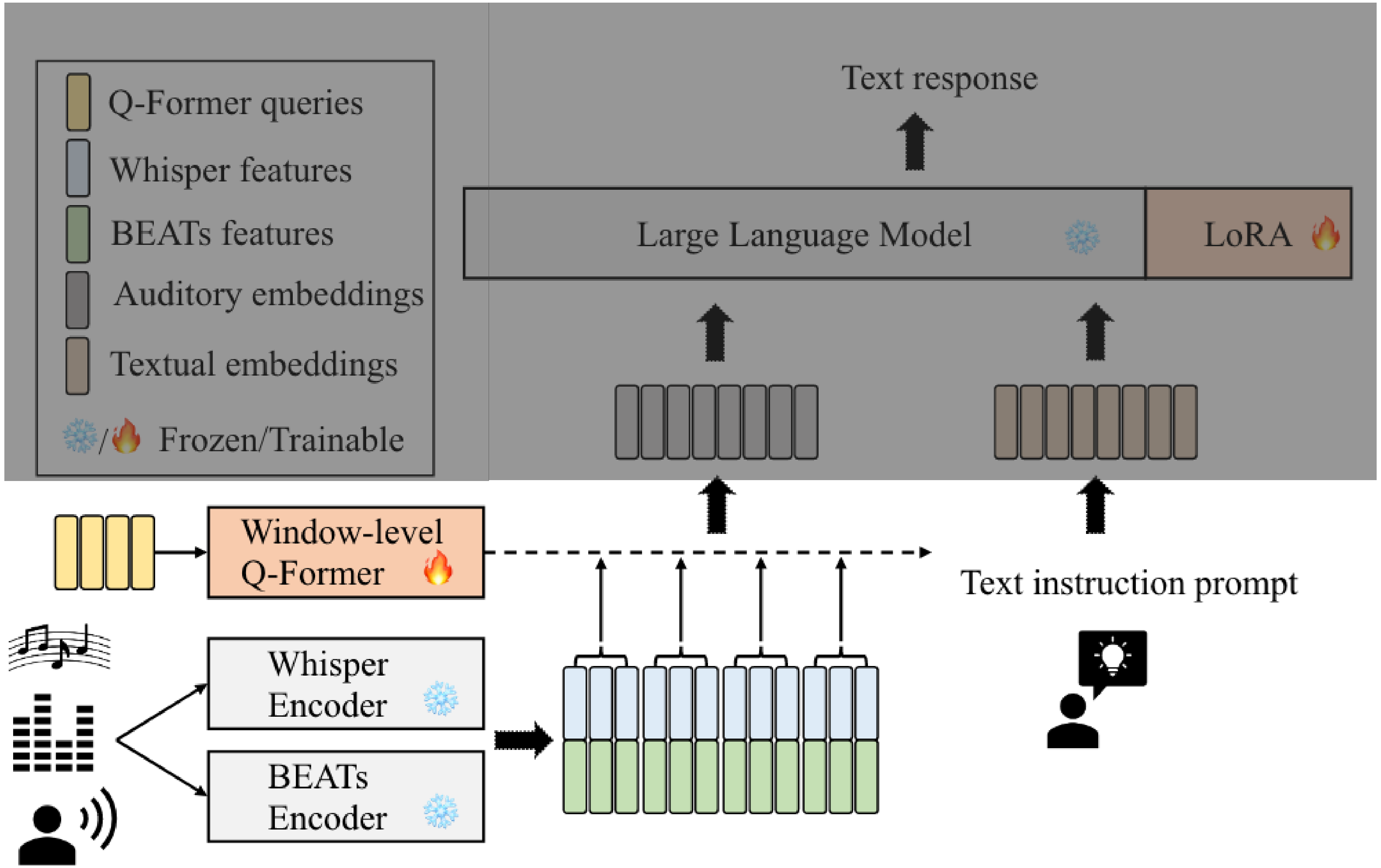
Tang, Changli, et al. "SALMONN: Towards Generic Hearing Abilities for Large Language Models." arXiv preprint arXiv:2310.13289 (2023).

SALMONN: TOWARDS GENERIC HEARING ABILITIES FOR LARGE LANGUAGE MODELS



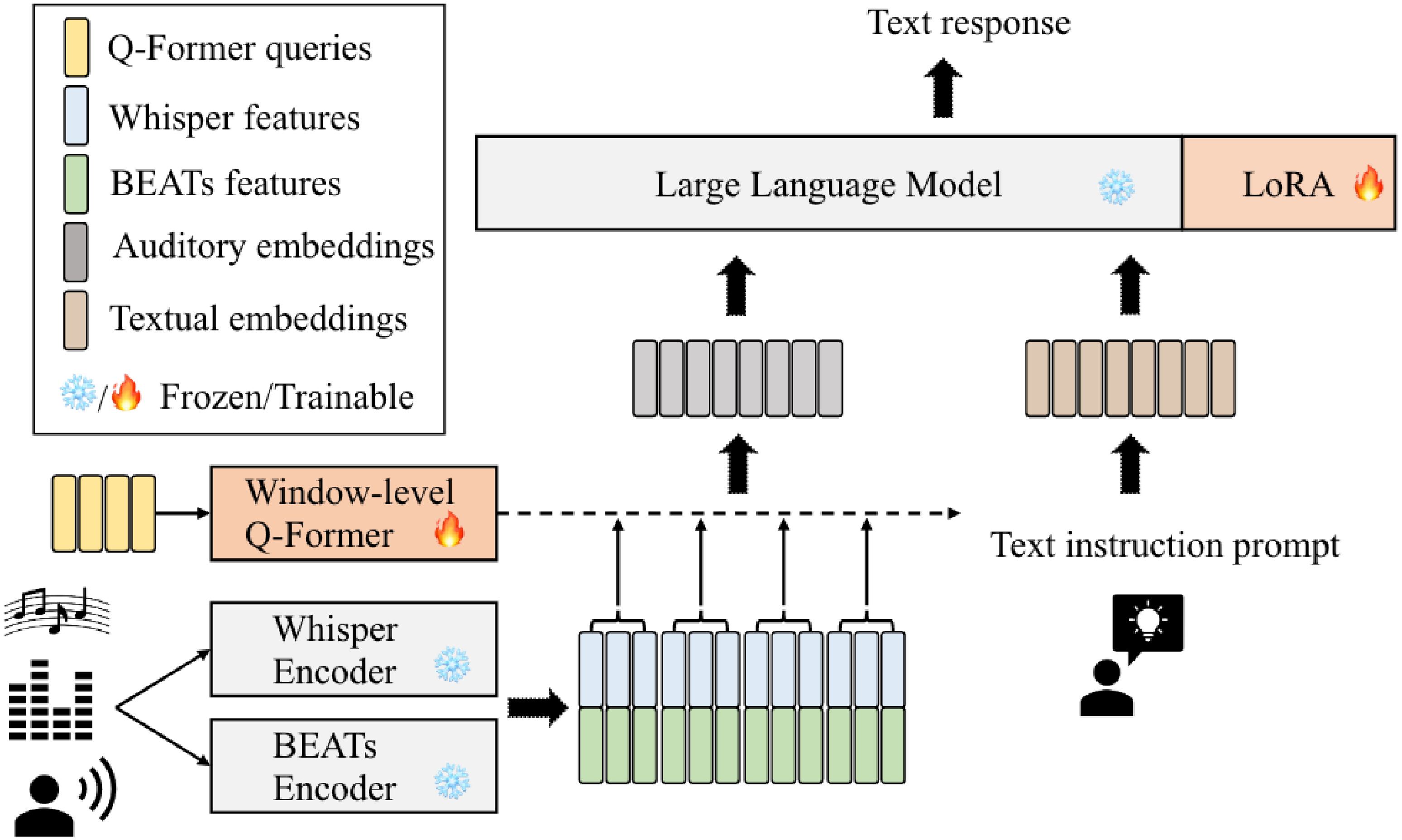
Tang, Changli, et al. "SALMONN: Towards Generic Hearing Abilities for Large Language Models." arXiv preprint arXiv:2310.13289 (2023).

SALMONN: TOWARDS GENERIC HEARING ABILITIES FOR LARGE LANGUAGE MODELS



Tang, Changli, et al. "SALMONN: Towards Generic Hearing Abilities for Large Language Models." arXiv preprint arXiv:2310.13289 (2023).

SALMONN: TOWARDS GENERIC HEARING ABILITIES FOR LARGE LANGUAGE MODELS



Tang, Changli, et al. "SALMONN: Towards Generic Hearing Abilities for Large Language Models." arXiv preprint arXiv:2310.13289 (2023).

Generation from unique modalities

DreamDiffusion: Generating High-Quality Images from Brain EEG Signals

Motivation:

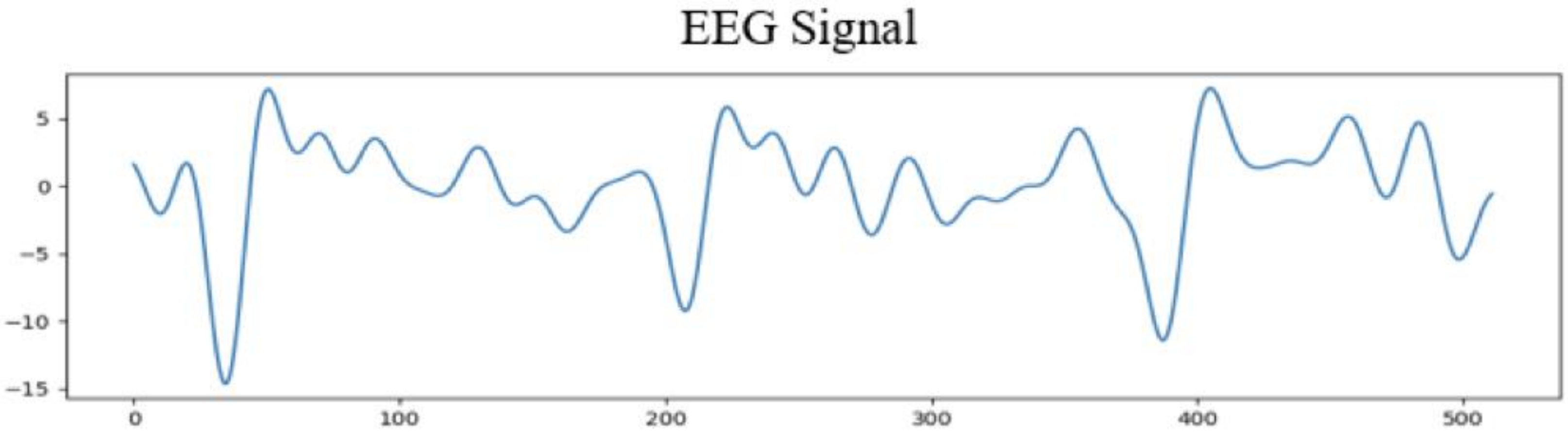
- Image generation has seen advancements, especially in text-to-image methods
- What about "thoughts-to-images"?
- Challenges and Opportunities:
    - Current methods of image reconstruction rely on fMRI, but it's non-portable and costly.
    - EEG is non-invasive, low-cost, and has potential applications in art, dreams visualization, and therapy

Challenges in EEG-Based Image Generation

- Inherent Noise in EEG Signals: High temporal resolution, low spatial resolution

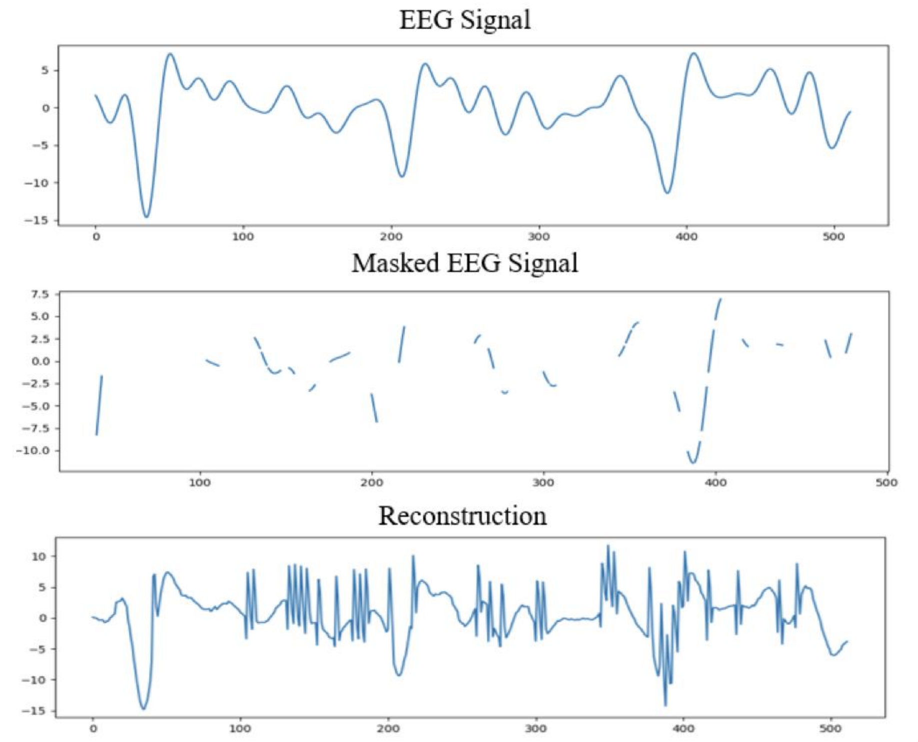
- Limited Information and Individual Differences in EEG data

- EEG space differs significantly from text and image spaces



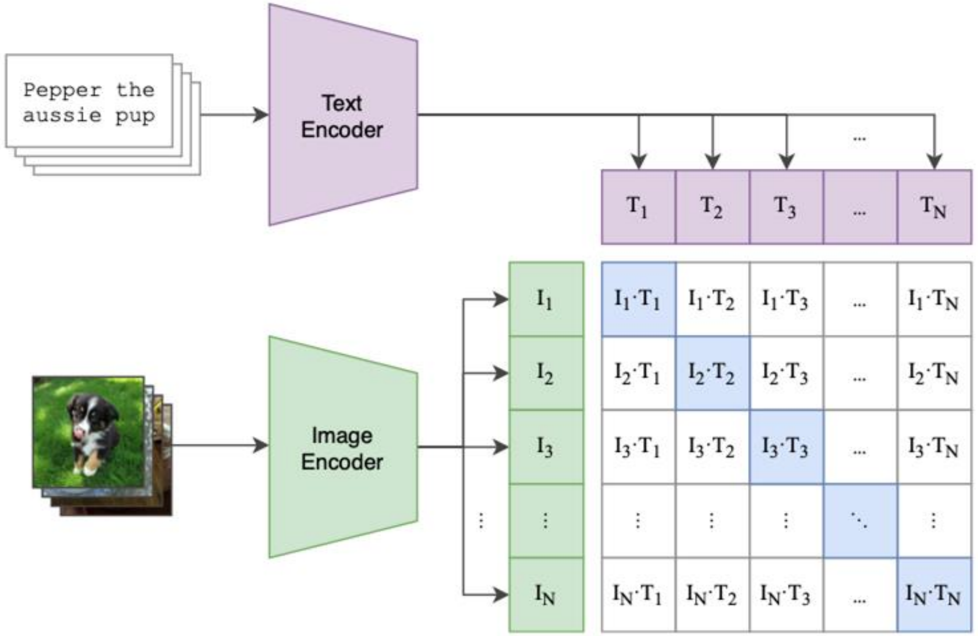
Addressing Noise and Limited Information

- Objective: Train EEG representations using abundant EEG data
- Method: Temporal masked signal modeling to predict missing tokens
- Uniqueness: Focus on temporal characteristics, deepening understanding across diverse brain activitie



Aligning EEG, Text, and Image Spaces

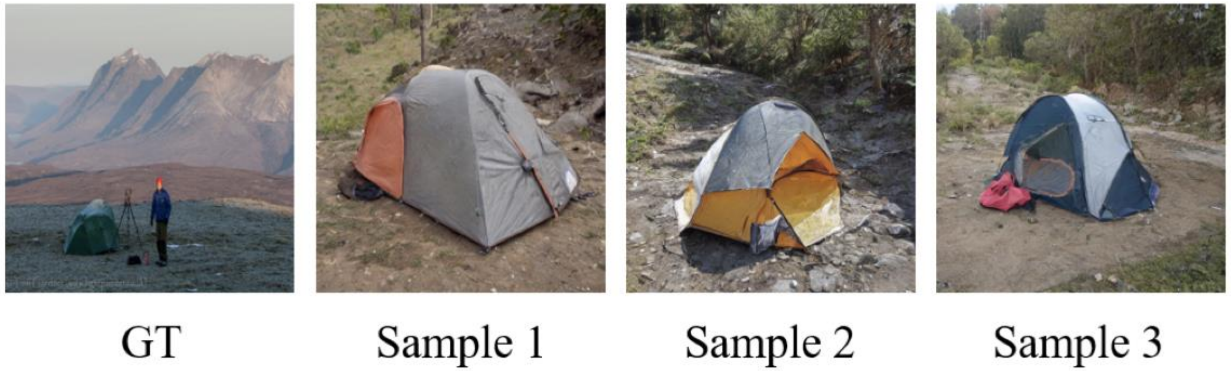
- Previous methods' limitation: Fine-tuning Stable Diffusion with limited data
- Solution: Introduce CLIP supervision to align EEG, text, and image embeddings
- Process: Leverage CLIP's image encoder to refine EEG feature embeddings, enhancing alignment

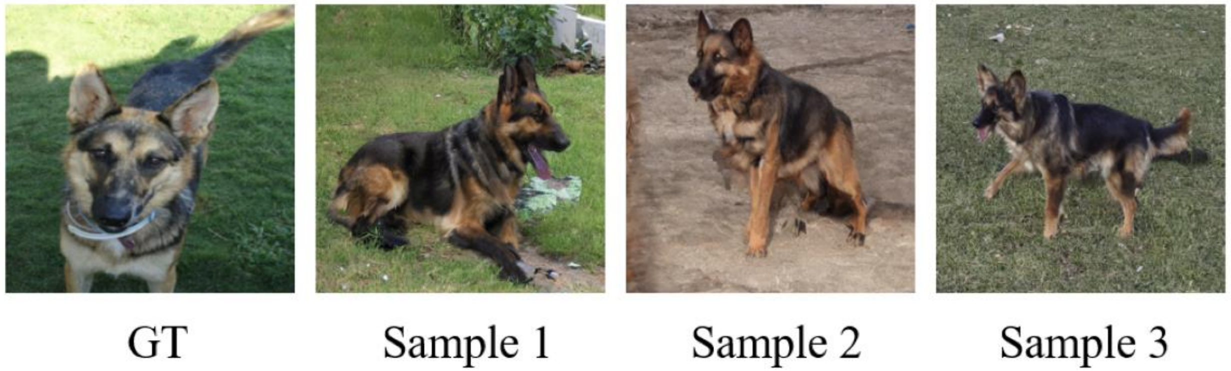


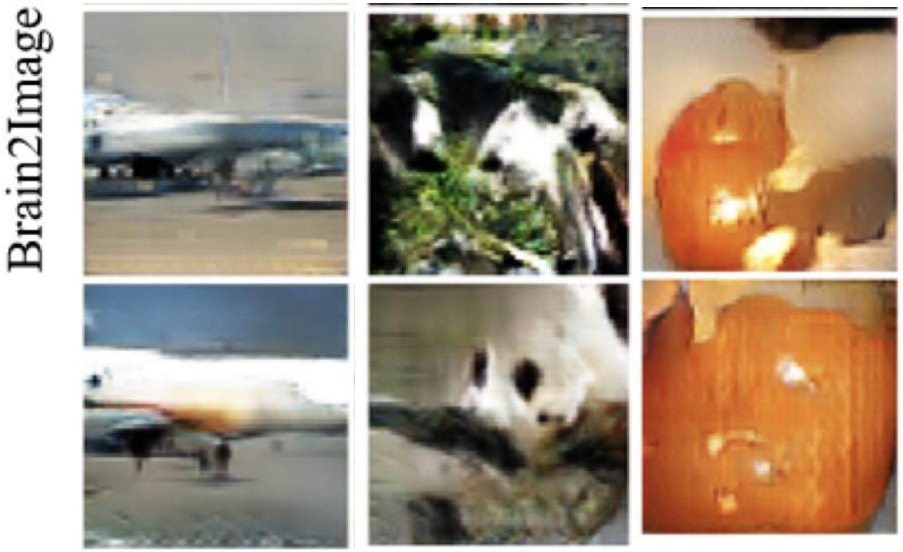
	T_1	T_2	T_3	...	T_N
I_1	$I_1{\cdot}T_1$	$I_1{\cdot}T_2$	$I_1{\cdot}T_3$	...	$I_1{\cdot}T_N$
I_2	$I_2{\cdot}T_1$	$I_2{\cdot}T_2$	$I_2{\cdot}T_3$	...	$I_2{\cdot}T_N$
I_3	$I_3{\cdot}T_1$	$I_3{\cdot}T_2$	$I_3{\cdot}T_3$	...	$I_3{\cdot}T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N{\cdot}T_1$	$I_N{\cdot}T_2$	$I_N{\cdot}T_3$	...	$I_N{\cdot}T_N$

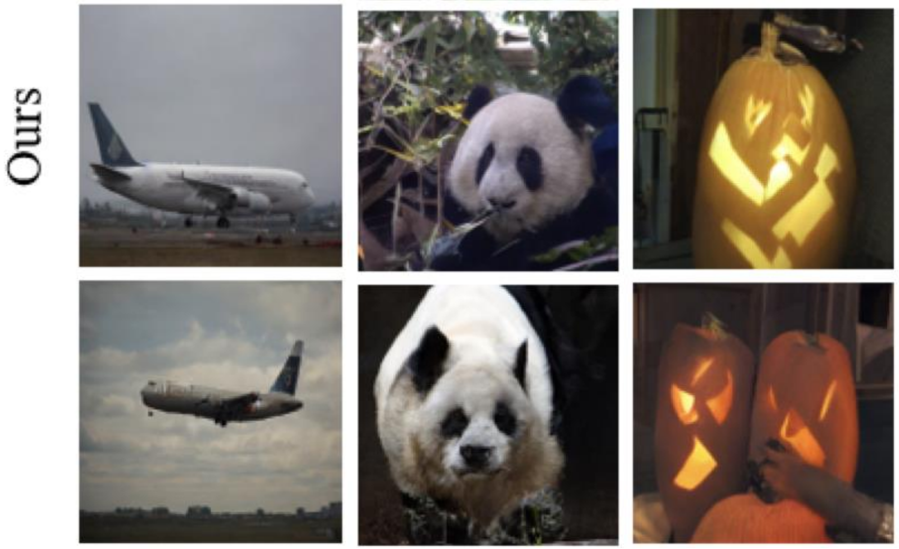
Results

- Quantitative and qualitative results show the method's effectiveness.









Embodied Multimodal Models

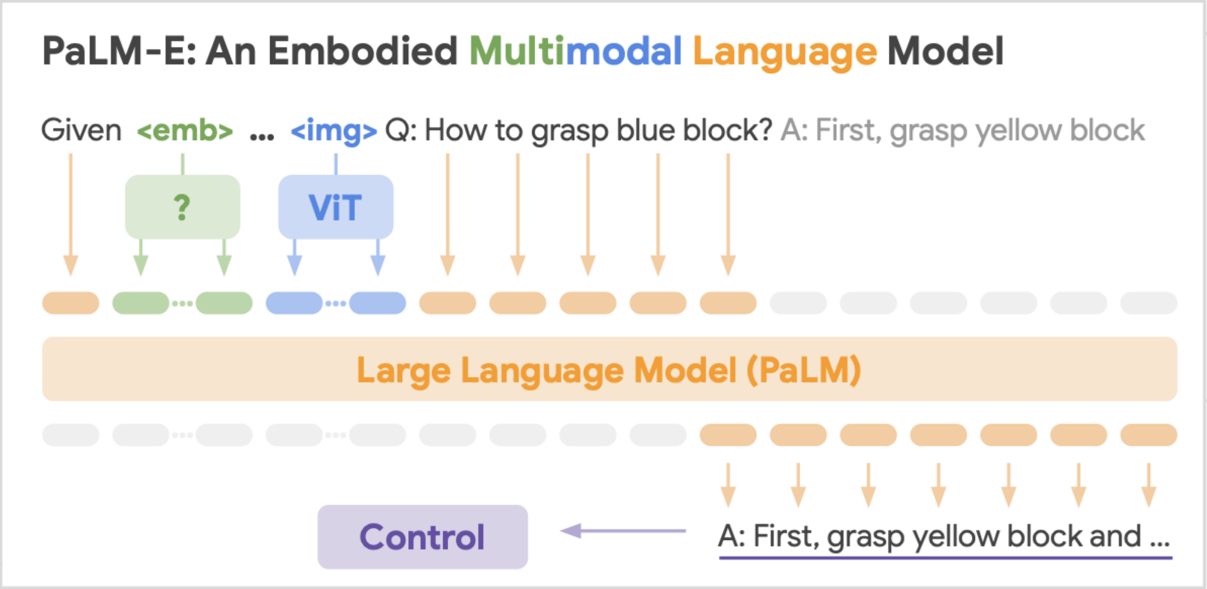
PaLM-E: An Embodied Multimodal Language Model

Motivation:

- Large Language Models (LLMs) excel in various domains.
- Real-world applications, like robotics, demand grounding—connecting words to real-world sensor modalities.
- LLMs trained on massive textual data lack direct connections to real-world visual and physical sensor modalities.
- Existing methods (Ahn et al., 2022) use robotic policies but are limited by providing only textual input.
- Evidence: State-of-the-art visual-language models struggle with direct solutions to robotic reasoning tasks.

Building Embodied Language Models

- Main Idea: Inject continuous, embodied observations into PaLM-E's language embedding space.
- Implementation: Encode multi-modal observations into vectors matching language token dimensions.
- Analogy: Continuous data injected akin to language tokens.


PaLM-E: An Embodied Multimodal Language Model
Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block
?
ViT
Large Language Model (PaLM)
Control
A: First, grasp yellow block and ...

Modalities incorporated into PaLM-E

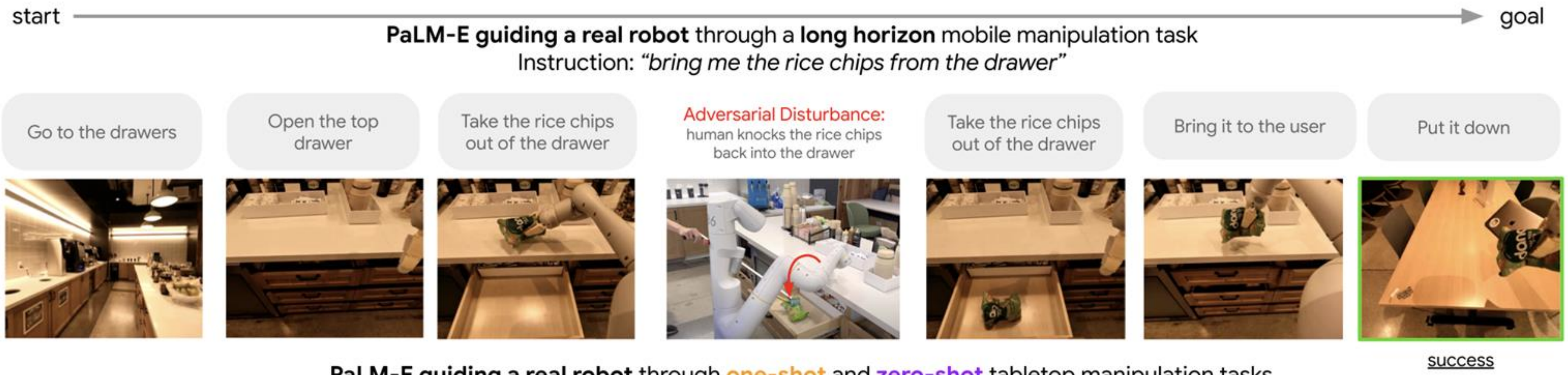
- State Estimation Vectors:
  - Simplest input, representing robot or object states.
- Vision Transformer (ViT):
  - Utilizes ViT for 2D image features.
- Object-Centric Representations:
  - Structured encoders for visual input lacking pre-structured entities.
  - Decomposes ViT's representation into distinct objects using ground-truth masks.
- Object Scene Representation Transformer (OSRT):
  - Learns 3D-centric neural scene representations.
- Entity Referrals:
  - Labels multi-modal tokens for object identification.
  - Enables PaLM-E to reference objects using special tokens

Connecting Output to Embodiment

- PaLM-E is a decoder-only LLM that autoregressively generates textual completions given a prefix or prompt

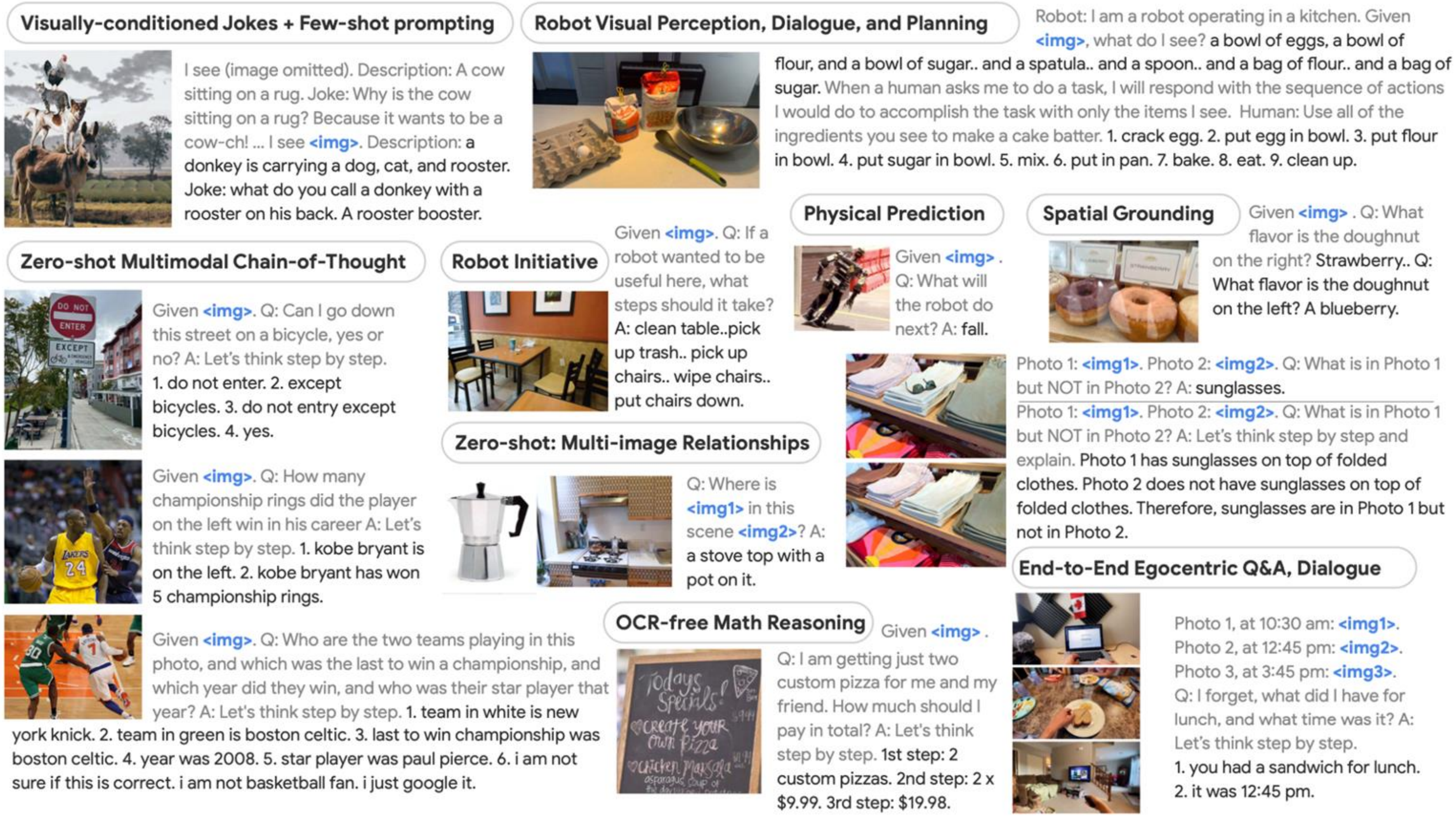
- Hence, we need to distinguish tasks requiring text-only output from those involving embodied planning or control

- Embodied Task Execution assumes access to policies executing low-level skills from a limited vocabulary.


start
goal
PaLM-E guiding a real robot through a long horizon mobile manipulation task
Instruction: "bring me the rice chips from the drawer"
Go to the drawers
Open the top drawer
Take the rice chips out of the drawer
Adversarial Disturbance: human knocks the rice chips back into the drawer
Take the rice chips out of the drawer
Bring it to the user
Put it down
success

PaLM-E guiding a real robot through one-shot and zero-shot tabletop manipulation tasks

Results



Visually-conditioned Jokes + Few-shot prompting

I see (image omitted). Description: A cow sitting on a rug. Joke: Why is the cow sitting on a rug? Because it wants to be a cow-ch! ... I see <img>. Description: a donkey is carrying a dog, cat, and rooster. Joke: what do you call a donkey with a rooster on his back. A rooster booster.

Robot Visual Perception, Dialogue, and Planning

Robot: I am a robot operating in a kitchen. Given <img>, what do I see? a bowl of eggs, a bowl of flour, and a bowl of sugar.. and a spatula.. and a spoon.. and a bag of flour.. and a bag of sugar. When a human asks me to do a task, I will respond with the sequence of actions I would do to accomplish the task with only the items I see. Human: Use all of the ingredients you see to make a cake batter. 1. crack egg. 2. put egg in bowl. 3. put flour in bowl. 4. put sugar in bowl. 5. mix. 6. put in pan. 7. bake. 8. eat. 9. clean up.

Zero-shot Multimodal Chain-of-Thought

Given <img>. Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step. 1. do not enter. 2. except bicycles. 3. do not entry except bicycles. 4. yes.

Given <img>. Q: How many championship rings did the player on the left win in his career A: Let's think step by step. 1. kobe bryant is on the left. 2. kobe bryant has won 5 championship rings.

Given <img>. Q: Who are the two teams playing in this photo, and which was the last to win a championship, and which year did they win, and who was their star player that year? A: Let's think step by step. 1. team in white is new york knick. 2. team in green is boston celtic. 3. last to win championship was boston celtic. 4. year was 2008. 5. star player was paul pierce. 6. i am not sure if this is correct. i am not basketball fan. i just google it.

Robot Initiative

Given <img>. Q: If a robot wanted to be useful here, what steps should it take? A: clean table..pick up trash.. pick up chairs.. wipe chairs.. put chairs down.

Physical Prediction

Given <img> . Q: What will the robot do next? A: fall.

Spatial Grounding

Given <img> . Q: What flavor is the doughnut on the right? Strawberry.. Q: What flavor is the doughnut on the left? A blueberry.

Zero-shot: Multi-image Relationships

Q: Where is <img1> in this scene <img2>? A: a stove top with a pot on it.

Photo 1: <img1>. Photo 2: <img2>. Q: What is in Photo 1 but NOT in Photo 2? A: sunglasses.

Photo 1: <img1>. Photo 2: <img2>. Q: What is in Photo 1 but NOT in Photo 2? A: Let's think step by step and explain. Photo 1 has sunglasses on top of folded clothes. Photo 2 does not have sunglasses on top of folded clothes. Therefore, sunglasses are in Photo 1 but not in Photo 2.

OCR-free Math Reasoning

Given <img> . Q: I am getting just two custom pizza for me and my friend. How much should I pay in total? A: Let's think step by step. 1st step: 2 custom pizzas. 2nd step: 2 x $9.99. 3rd step: $19.98.

End-to-End Egocentric Q&A, Dialogue

Photo 1, at 10:30 am: <img1>.
Photo 2, at 12:45 pm: <img2>.
Photo 3, at 3:45 pm: <img3>.
Q: I forget, what did I have for lunch, and what time was it? A: Let's think step by step.
1. you had a sandwich for lunch.
2. it was 12:45 pm.

"That's all Folks!"

Compositional Reasoning

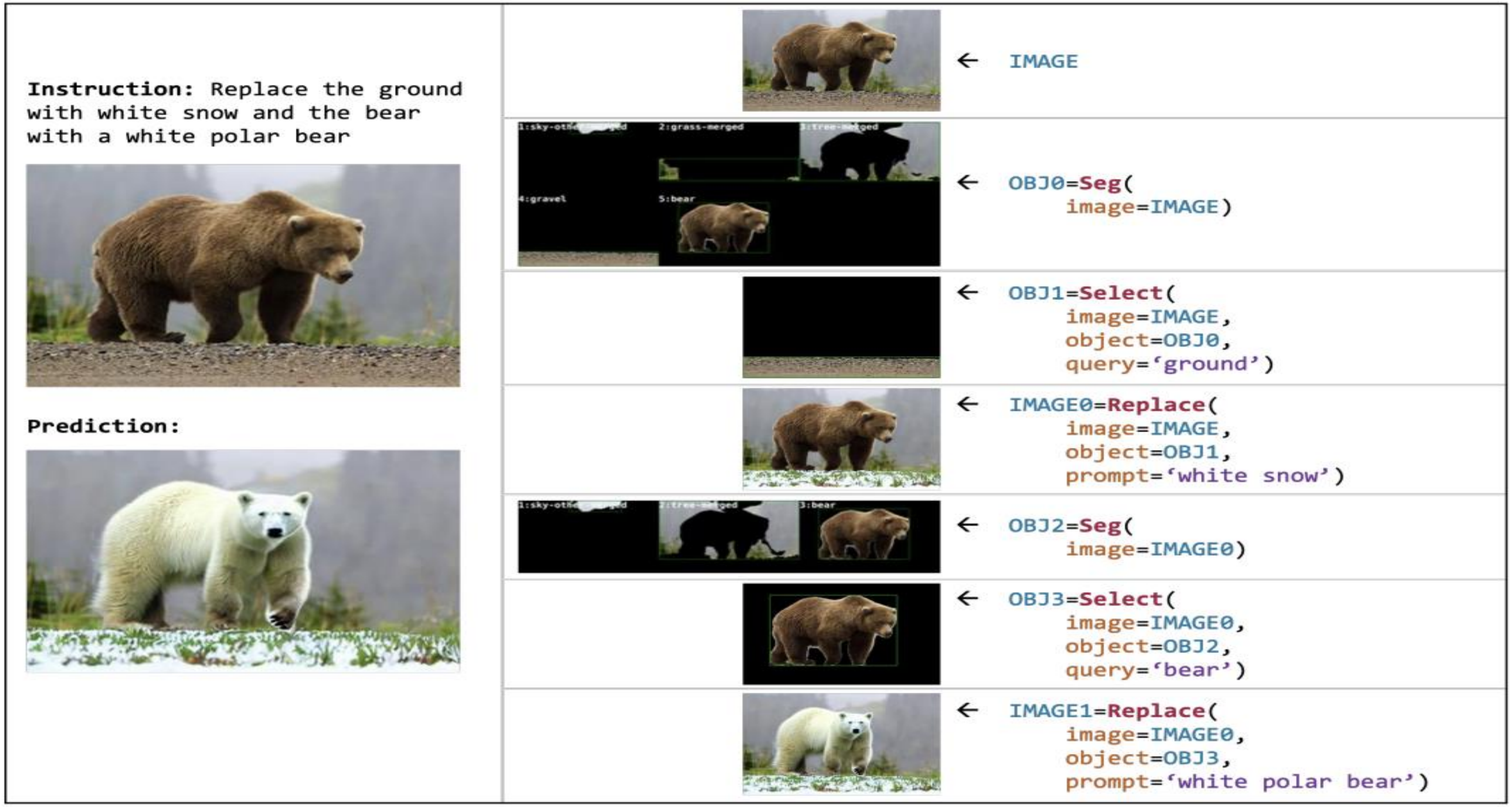
Visual Programming: Compositional visual reasoning without training

Challenge: While existing approaches work well on individual tasks, they struggle with scaling to a wide range of complex tasks.

VISPROG Overview:

- VISPROG inputs visual data and natural language instructions.
- Generates a sequence of steps or a "visual program" to execute tasks.
- Uses modules like computer vision models, language models, and image processing routines.

Visual Programming: Compositional visual reasoning without training



Instruction: Replace the ground with white snow and the bear with a white polar bear

Prediction:

```
← IMAGE

← OBJ0=Seg(
      image=IMAGE)

← OBJ1=Select(
      image=IMAGE,
      object=OBJ0,
      query='ground')

← IMAGE0=Replace(
      image=IMAGE,
      object=OBJ1,
      prompt='white snow')

← OBJ2=Seg(
      image=IMAGE0)

← OBJ3=Select(
      image=IMAGE0,
      object=OBJ2,
      query='bear')

← IMAGE1=Replace(
      image=IMAGE0,
      object=OBJ3,
      prompt='white polar bear')
```

Visual Programming: Compositional visual reasoning without training

Capabilities:

- VISPROG uses GPT-3 for in-context learning and program generation.
- Supports various modules for image understanding, manipulation, and knowledge retrieval.
- Highly interpretable with logical, step-by-step visual rationales for tasks.

Advantages Over Previous Methods:

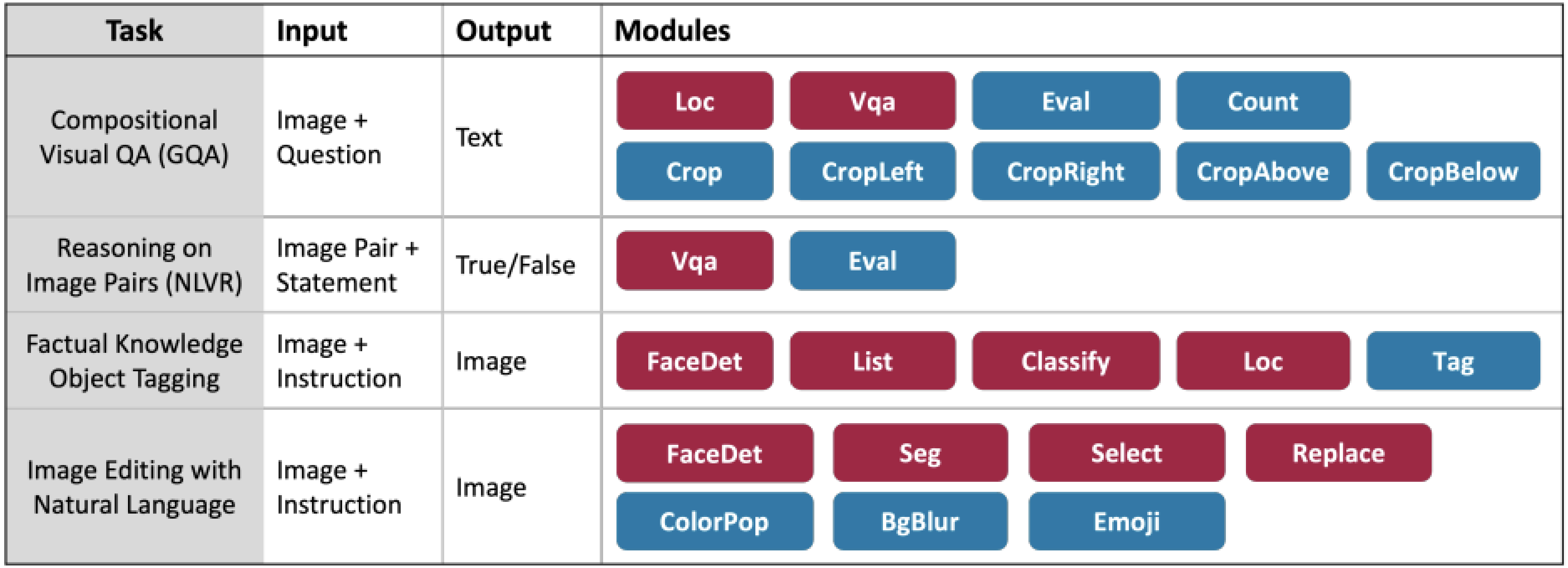
- Higher level of abstraction than Neural Module Networks.
- More flexible and modular, allowing for a wide range of tasks without specific training.

Key Contributions & Use Cases:

- Demonstrates flexibility in tasks like visual question answering, image editing, and object tagging.
- Impressive performance gains in tests (e.g., VQA tasks, zero-shot accuracy in NLVR).

Visual Programming: Compositional visual reasoning without training

Diverse Applications of VISPROG:



Task	Input	Output	Modules
Compositional Visual QA (GQA)	Image + Question	Text	Loc Vqa Eval Count Crop CropLeft CropRight CropAbove CropBelow
Reasoning on Image Pairs (NLVR)	Image Pair + Statement	True/False	Vqa Eval
Factual Knowledge Object Tagging	Image + Instruction	Image	FaceDet List Classify Loc Tag
Image Editing with Natural Language	Image + Instruction	Image	FaceDet Seg Select Replace ColorPop BgBlur Emoji

Visual Programming: Compositional visual reasoning without training

Analysis and Results:

Effect of Prompt Size:

- More in-context examples improve performance in GQA and NLVR tasks.
- Majority voting across different runs enhances accuracy.
- Performance in NLVR saturates with fewer prompts compared to GQA.

Generalization across Tasks:

- Various prompting strategies (random, voting, curated) impact performance differently.
- Curated prompts demonstrate comparable results to voting, with less computational resource.
- VISPROG shows strong zero-shot performance, particularly in single-image VQA for NLVR.

Unified Visio-Linguistic Model

Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks

- Modality Convergence: Integrates language and vision pretraining.
- Multiway Transformers for handling multiple modalities.
- Simplified pre-training with 'mask-then-predict' method.
- Versatile Applications: Effective across object detection, segmentation, classification, and more.

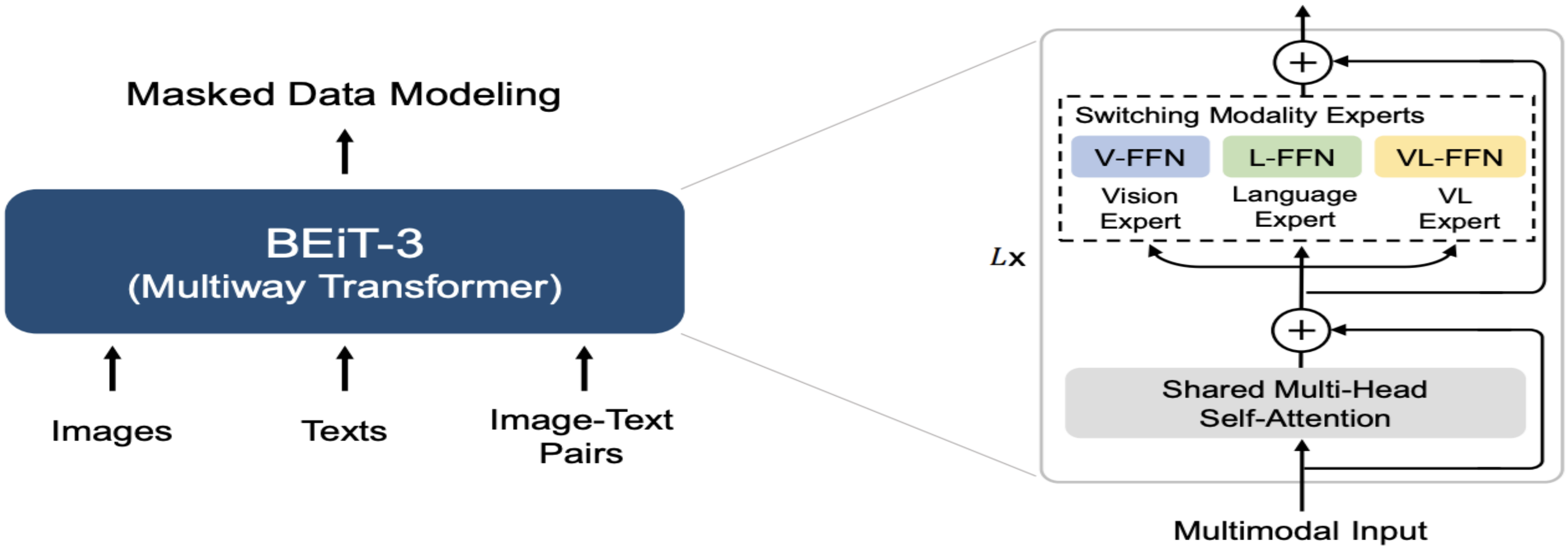


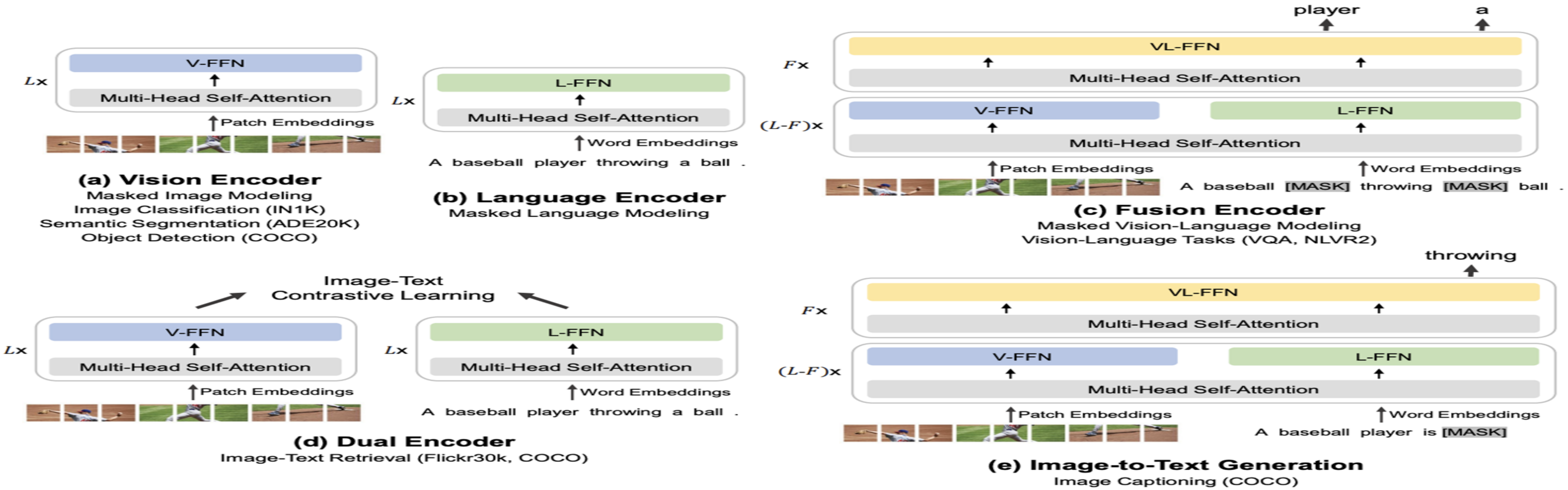
Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks

Multiway Transformer Backbone:

- Shared self-attention module with vision, language, and vision-language experts.
- Enables deep fusion for multimodal tasks.

Masked Data Modeling:

- Unified task for both monomodal and multimodal data.
- Learns representations and alignments between modalities.


Lx
V-FFN
Multi-Head Self-Attention
Patch Embeddings
Lx
L-FFN
Multi-Head Self-Attention
Word Embeddings
A baseball player throwing a ball .
player
a
Fx
VL-FFN
Multi-Head Self-Attention
(L-F)x
V-FFN
L-FFN
Multi-Head Self-Attention
Patch Embeddings
Word Embeddings
A baseball [MASK] throwing [MASK] ball .
Image-Text
Contrastive Learning
Lx
V-FFN
Multi-Head Self-Attention
Patch Embeddings
Lx
L-FFN
Multi-Head Self-Attention
Word Embeddings
A baseball player throwing a ball .
throwing
Fx
VL-FFN
Multi-Head Self-Attention
(L-F)x
V-FFN
L-FFN
Multi-Head Self-Attention
Patch Embeddings
Word Embeddings
A baseball player is [MASK]

(a) Vision Encoder
Masked Image Modeling
Image Classification (IN1K)
Semantic Segmentation (ADE20K)
Object Detection (COCO)

(b) Language Encoder
Masked Language Modeling

(c) Fusion Encoder
Masked Vision-Language Modeling
Vision-Language Tasks (VQA, NLVR2)

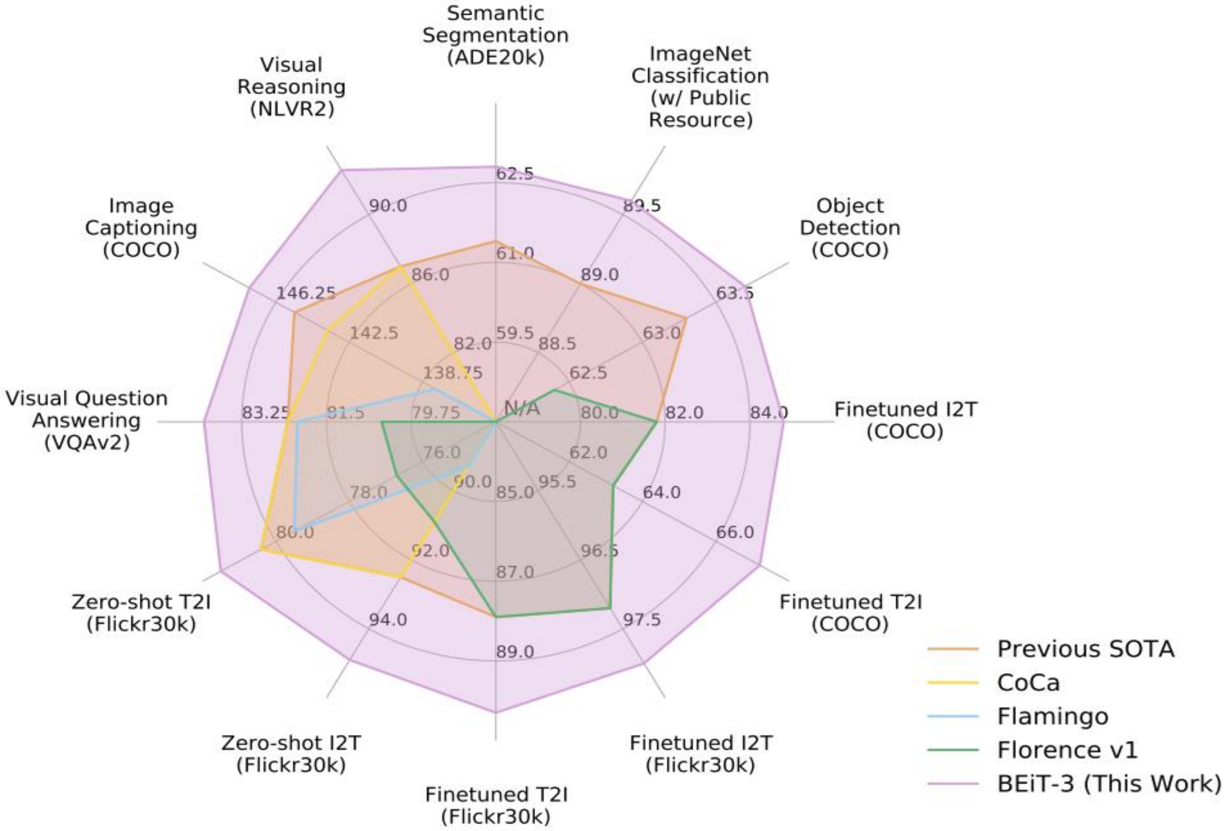
(d) Dual Encoder
Image-Text Retrieval (Flickr30k, COCO)

(e) Image-to-Text Generation
Image Captioning (COCO)

Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks

Some Results:

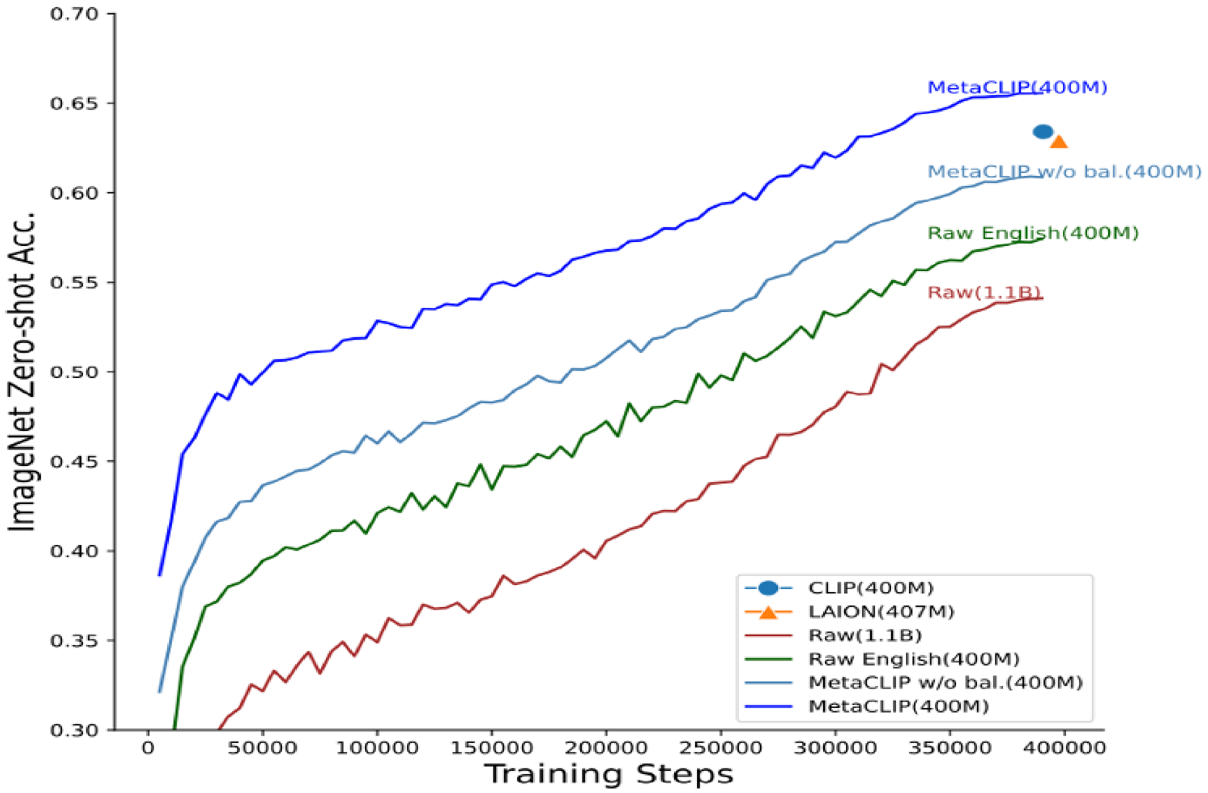
- Fusion encoder model surpasses previous models on VQA v2.0 and achieves 84.03% accuracy
- Sets new record in NLVR2, reaching over 90% accuracy.



Multimodal Alignment

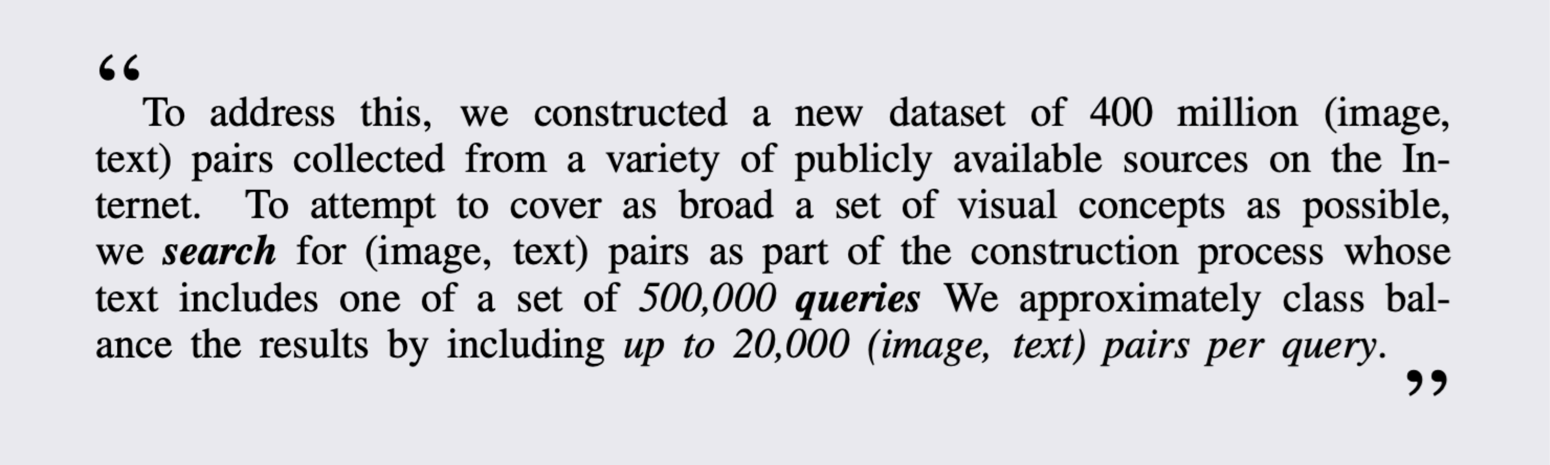
DEMYSTIFYING CLIP DATA

- The increasing availability of pre-trained models for public use contrasts sharply with the lack of transparency regarding their training data. ***What is the significance of good data for a good model?***
- **Demystifying CLIP Data**. The success of CLIP in computer vision is attributed to its data rather than its architecture or pre-training objective.



DEMYSTIFYING CLIP DATA

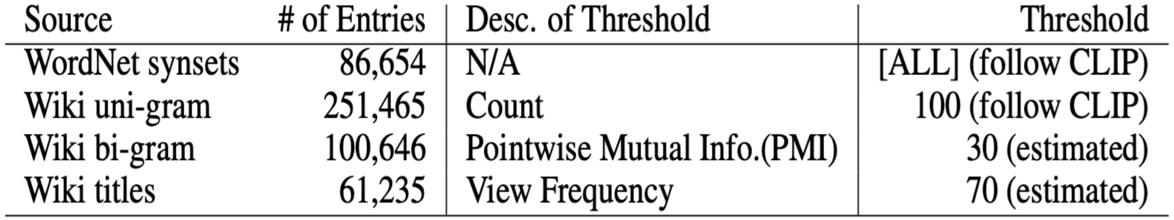
- The limited disclosure of CLIP's data collection process has prompted the need to unveil its curation approach, leading to the creation of ***MetaCLIP***.
- The goal is to uncover CLIP’s data curation process, which involves preserving signal in the data while minimizing noise.
- Details from original CLIP paper:



> “ To address this, we constructed a new dataset of 400 million (image, text) pairs collected from a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we ***search*** for (image, text) pairs as part of the construction process whose text includes one of a set of *500,000* ***queries*** We approximately class balance the results by including *up to 20,000 (image, text) pairs per query*. ”

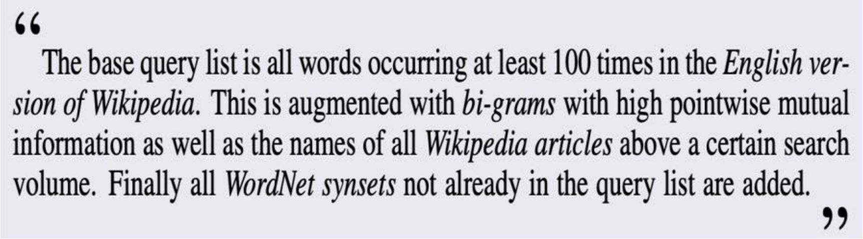
DEMYSTIFYING CLIP DATA

- Data construction process consists of the following steps:
  - Metadata Construction: M = {entry}



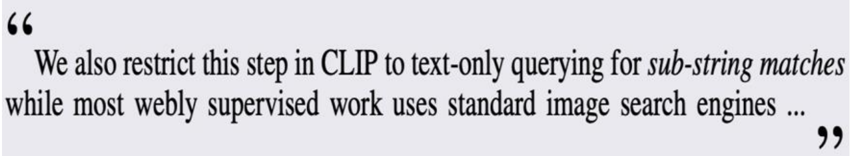
Source	# of Entries	Desc. of Threshold	Threshold
WordNet synsets	86,654	N/A	[ALL] (follow CLIP)
Wiki uni-gram	251,465	Count	100 (follow CLIP)
Wiki bi-gram	100,646	Pointwise Mutual Info.(PMI)	30 (estimated)
Wiki titles	61,235	View Frequency	70 (estimated)

Table 1: Composition of MetaCLIP Metadata.



> “ The base query list is all words occurring at least 100 times in the *English version of Wikipedia*. This is augmented with *bi-grams* with high pointwise mutual information as well as the names of all *Wikipedia articles* above a certain search volume. Finally all *WordNet synsets* not already in the query list are added. ”

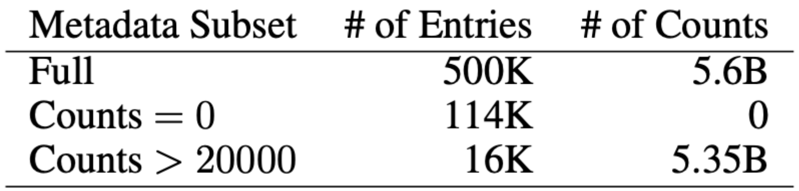
  - Sub-string Matching: text → entry
    - CommonCrawl (CC)^4 as the source (1.6B mage-text pairs)
    - Retains only high-quality matching texts



> “ We also restrict this step in CLIP to text-only querying for *sub-string matches* while most webly supervised work uses standard image search engines ... ”

DEMYSTIFYING CLIP DATA

- Data construction process consists of the following steps:
  - Inverted Indexing: entry → text
    - All texts associated with each metadata entry are aggregated into lists, creating a mapping from each entry to the corresponding texts, *entry* → *text*.
    - Out of the 500k entries, 114k entries have no matches.

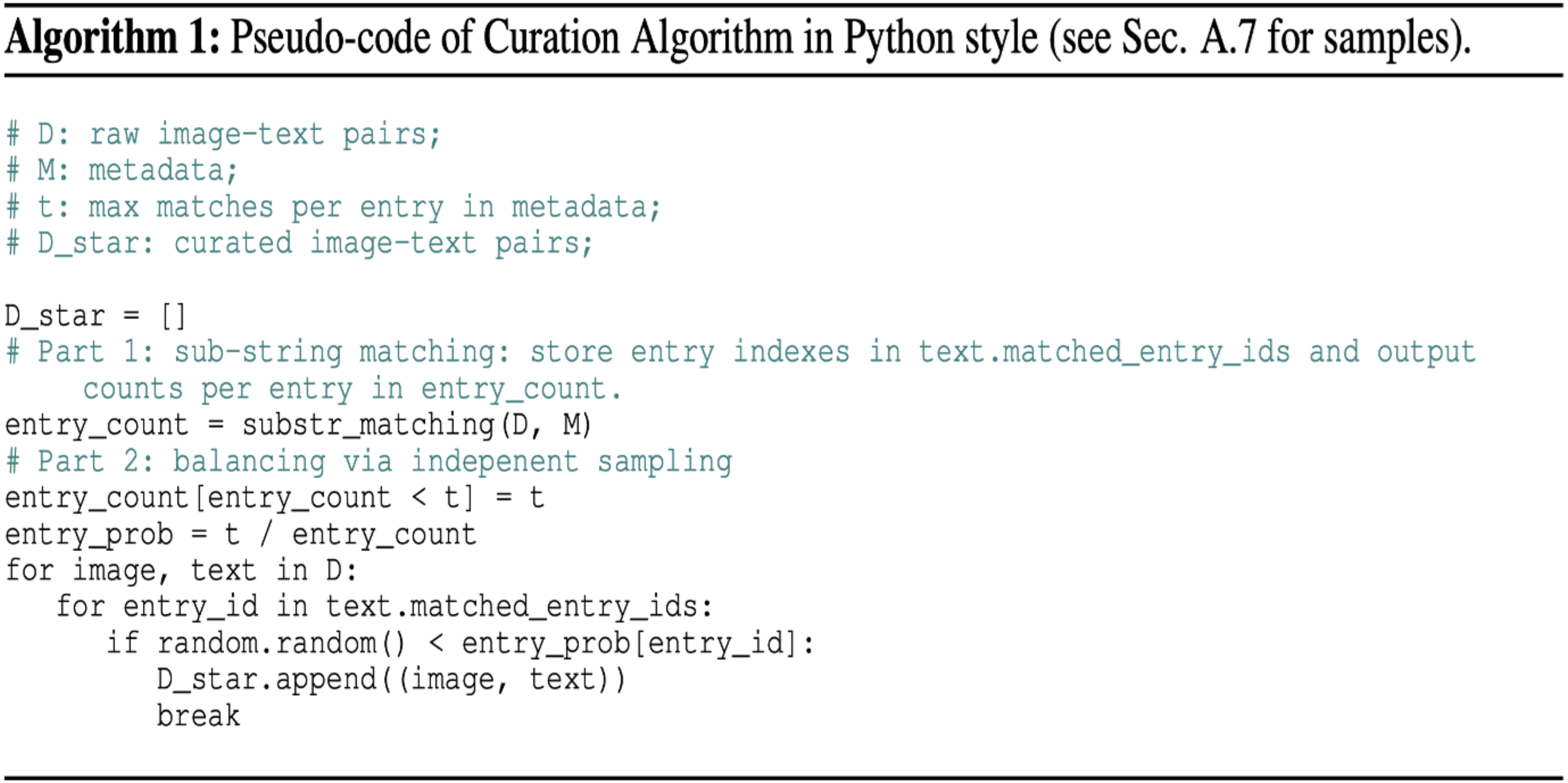


Metadata Subset	# of Entries	# of Counts
Full	500K	5.6B
Counts $= 0$	114K	0
Counts > 20000	16K	5.35B

Table 2: Summary of counts for entries.

  - Query and Balancing with t ≤20K
    - For each metadata entry, the associated list of texts (or image-text pairs) is sub-sampled, ensuring that the resulting data distribution is more balanced.
    - t = 20k is a threshold used to limit the number of texts/pairs for each entry.

DEMYSTIFYING CLIP DATA

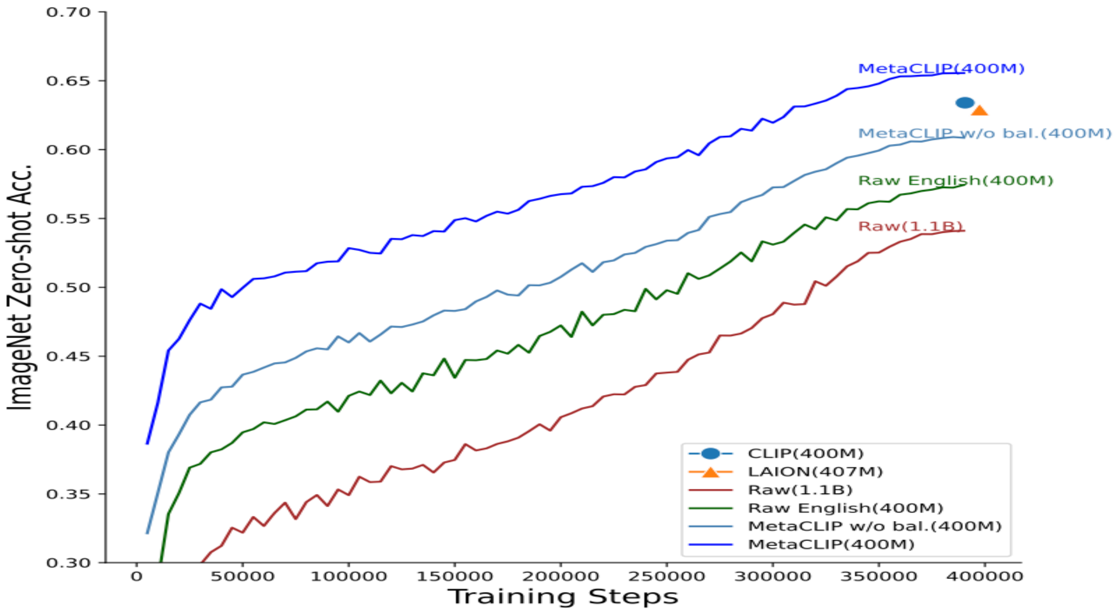


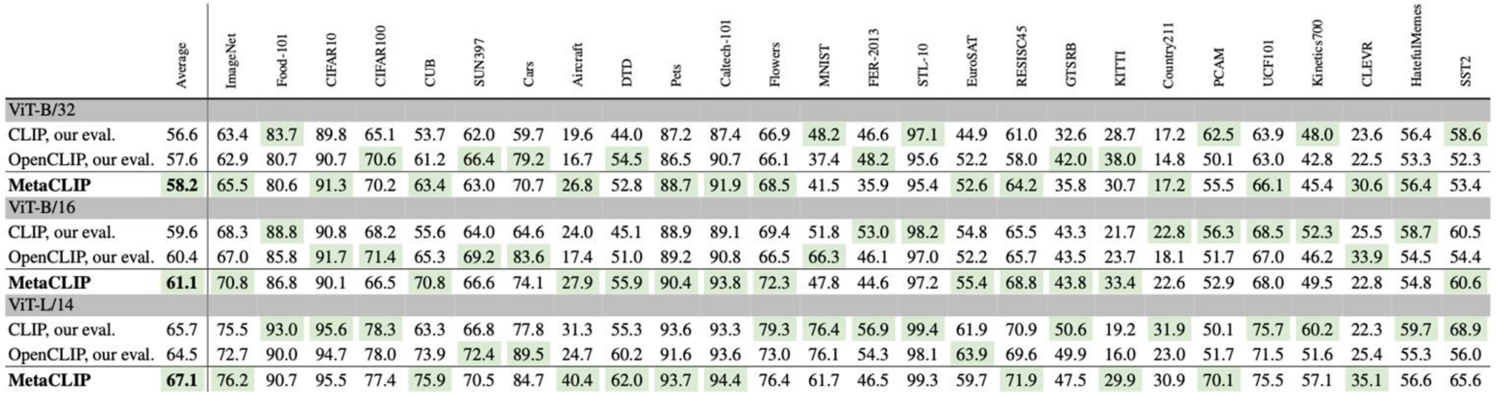
Algorithm 1: Pseudo-code of Curation Algorithm in Python style (see Sec. A.7 for samples).

```
# D: raw image-text pairs;
# M: metadata;
# t: max matches per entry in metadata;
# D_star: curated image-text pairs;

D_star = []
# Part 1: sub-string matching: store entry indexes in text.matched_entry_ids and output
    counts per entry in entry_count.
entry_count = substr_matching(D, M)
# Part 2: balancing via indepenent sampling
entry_count[entry_count < t] = t
entry_prob = t / entry_count
for image, text in D:
   for entry_id in text.matched_entry_ids:
      if random.random() < entry_prob[entry_id]:
         D_star.append((image, text))
         break
```

DEMYSTIFYING CLIP DATA





	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
ViT-B/32																											
CLIP, our eval.	56.6	63.4	83.7	89.8	65.1	53.7	62.0	59.7	19.6	44.0	87.2	87.4	66.9	48.2	46.6	97.1	44.9	61.0	32.6	28.7	17.2	62.5	63.9	48.0	23.6	56.4	58.6
OpenCLIP, our eval.	57.6	62.9	80.7	90.7	70.6	61.2	66.4	79.2	16.7	54.5	86.5	90.7	66.1	37.4	48.2	95.6	52.2	58.0	42.0	38.0	14.8	50.1	63.0	42.8	22.5	53.3	52.3
MetaCLIP	**58.2**	65.5	80.6	91.3	70.2	63.4	63.0	70.7	26.8	52.8	88.7	91.9	68.5	41.5	35.9	95.4	52.6	64.2	35.8	30.7	17.2	55.5	66.1	45.4	30.6	56.4	53.4
ViT-B/16																											
CLIP, our eval.	59.6	68.3	88.8	90.8	68.2	55.6	64.0	64.6	24.0	45.1	88.9	89.1	69.4	51.8	53.0	98.2	54.8	65.5	43.3	21.7	22.8	56.3	68.5	52.3	25.5	58.7	60.5
OpenCLIP, our eval.	60.4	67.0	85.8	91.7	71.4	65.3	69.2	83.6	17.4	51.0	89.2	90.8	66.5	66.3	46.1	97.0	52.2	65.7	43.5	23.7	18.1	51.7	67.0	46.2	33.9	54.5	54.4
MetaCLIP	**61.1**	70.8	86.8	90.1	66.5	70.8	66.6	74.1	27.9	55.9	90.4	93.8	72.3	47.8	44.6	97.2	55.4	68.8	43.8	33.4	22.6	52.9	68.0	49.5	22.8	54.8	60.6
ViT-L/14																											
CLIP, our eval.	65.7	75.5	93.0	95.6	78.3	63.3	66.8	77.8	31.3	55.3	93.6	93.3	79.3	76.4	56.9	99.4	61.9	70.9	50.6	19.2	31.9	50.1	75.7	60.2	22.3	59.7	68.9
OpenCLIP, our eval.	64.5	72.7	90.0	94.7	78.0	73.9	72.4	89.5	24.7	60.2	91.6	93.6	73.0	76.1	54.3	98.1	63.9	69.6	49.9	16.0	23.0	51.7	71.5	51.6	25.4	55.3	56.0
MetaCLIP	**67.1**	76.2	90.7	95.5	77.4	75.9	70.5	84.7	40.4	62.0	93.7	94.4	76.4	61.7	46.5	99.3	59.7	71.9	47.5	29.9	30.9	70.1	75.5	57.1	35.1	56.6	65.6

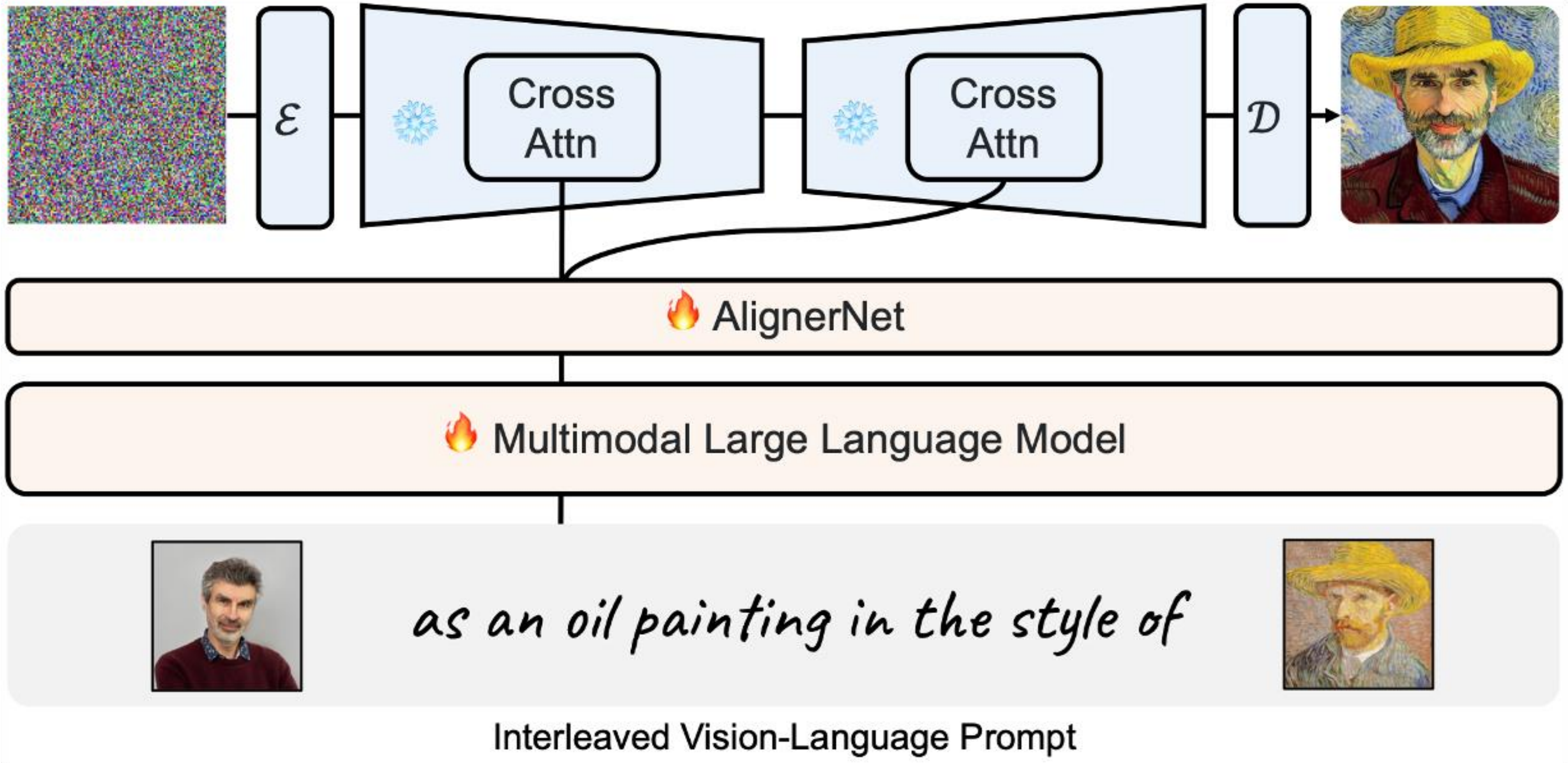
Table 4: MetaCLIP-400M *vs.* CLIP (WIT400M data) and OpenCLIP (LAION-400M data). We use 3 different model scales (ViT-B/32 and -B/16 and -L/14) and an identical training setup as CLIP.

Multimodal Generation

KOSMOS-G: Generating Images in Context with Multimodal Large Language Models

Major advancement in text-to-image (T2I) and vision-language-to-image (VL2I) generation.

- But, how can we generate images from generalized vision-language inputs?



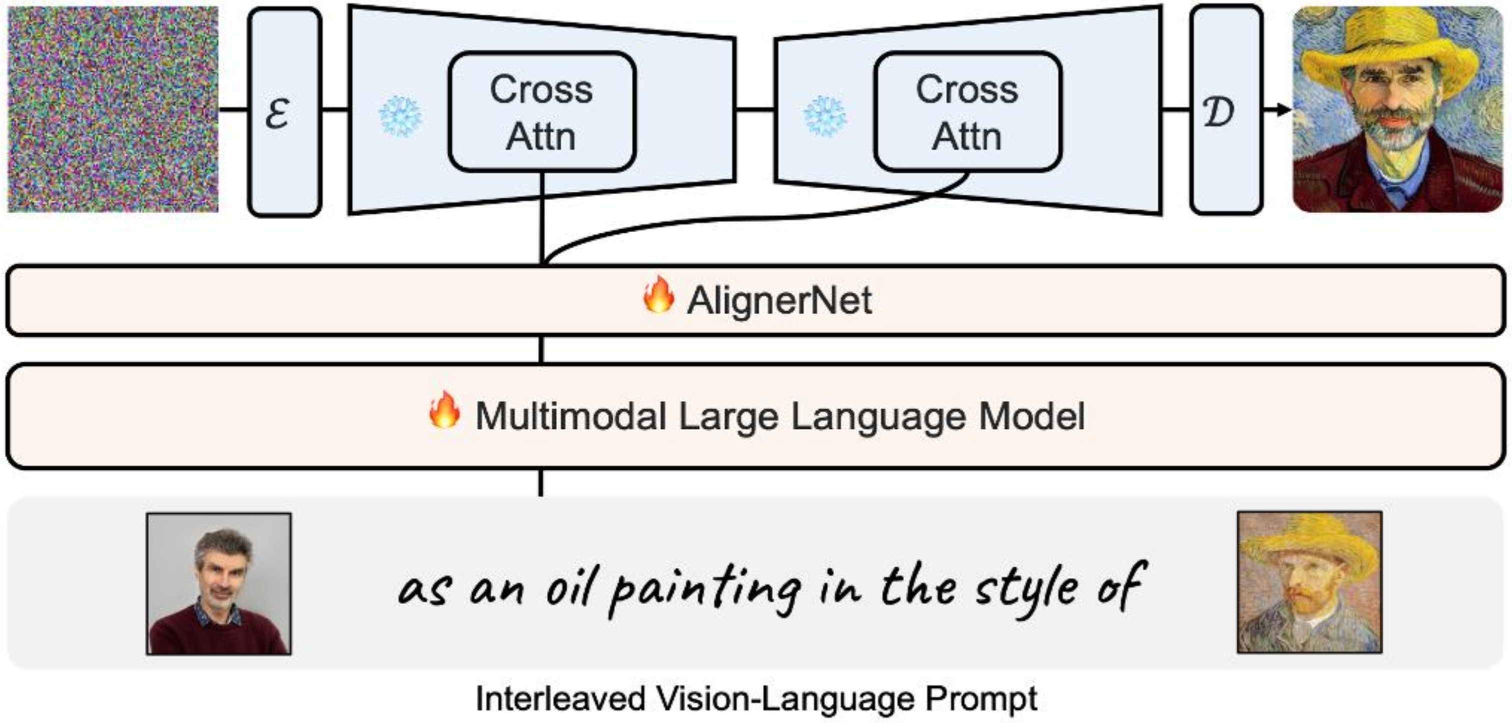
$\mathcal{E}$
Cross Attn
Cross Attn
$\mathcal{D}$
AlignerNet
Multimodal Large Language Model
as an oil painting in the style of
Interleaved Vision-Language Prompt

KOSMOS-G: “Alignment before Instruct”

The backbone of KOSMOS-G MLLM is a Transformer-based causal language model, serving as a general-purpose interface to multimodal input.

Entire pipeline can be divided into 3 stages:

1. Multimodal Language Modeling
2. Image Decoder Aligning
3. Instruction Tuning

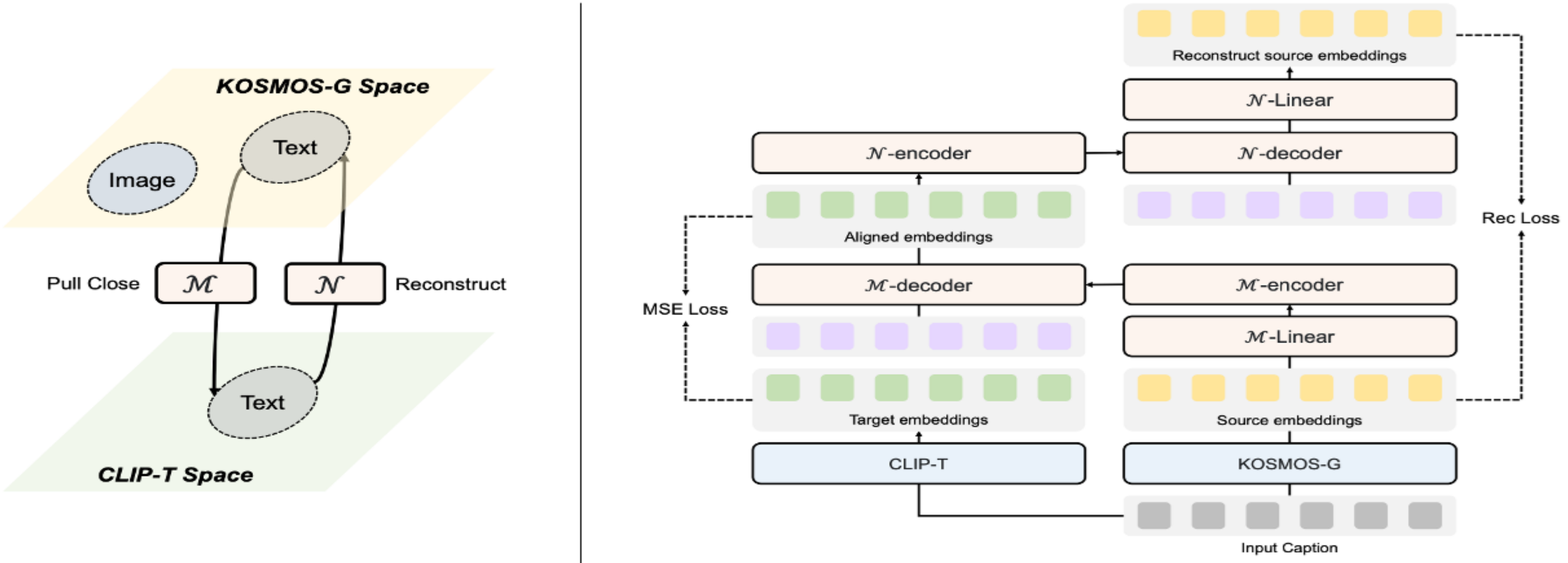


$\mathcal{E}$
Cross Attn
Cross Attn
$\mathcal{D}$
AlignerNet
Multimodal Large Language Model
as an oil painting in the style of
Interleaved Vision-Language Prompt

KOSMOS-G: Image Decoder Aligning

To make KOSMOS-G capable of image generation,

- Diffusion models are incorporated as the image decoder.
- AlignerNet is proposed that consists of an encoder M and a decoder N to learn the alignment between the KOSMOS-G source space and CLIP text encoder target space.



(a) Align process. Text serves as an anchor, image embeddings are naturally aligned throughout the process.

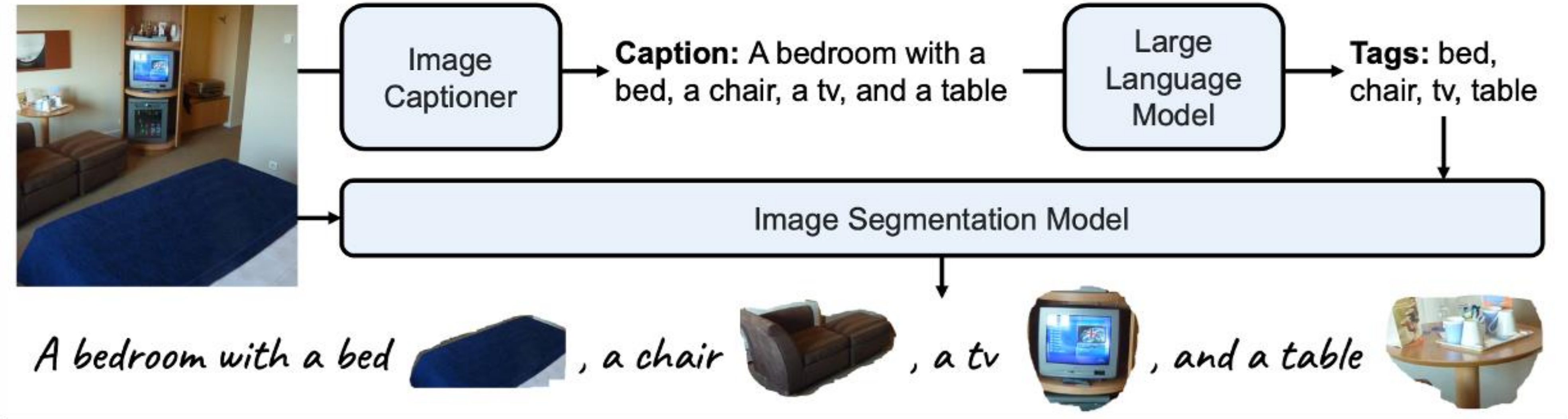
(b) AlignerNet architecture. The Linear layers are used to project the output dimension of MLLM to $d = 768$, the purple elements denote the learned latent queries $\mathbf{Q}_{\mathcal{M}}$ and $\mathbf{Q}_{\mathcal{N}}$.

KOSMOS-G: Instruction Tuning

To pursue the objective of *"image as a foreign language in image generation,"*

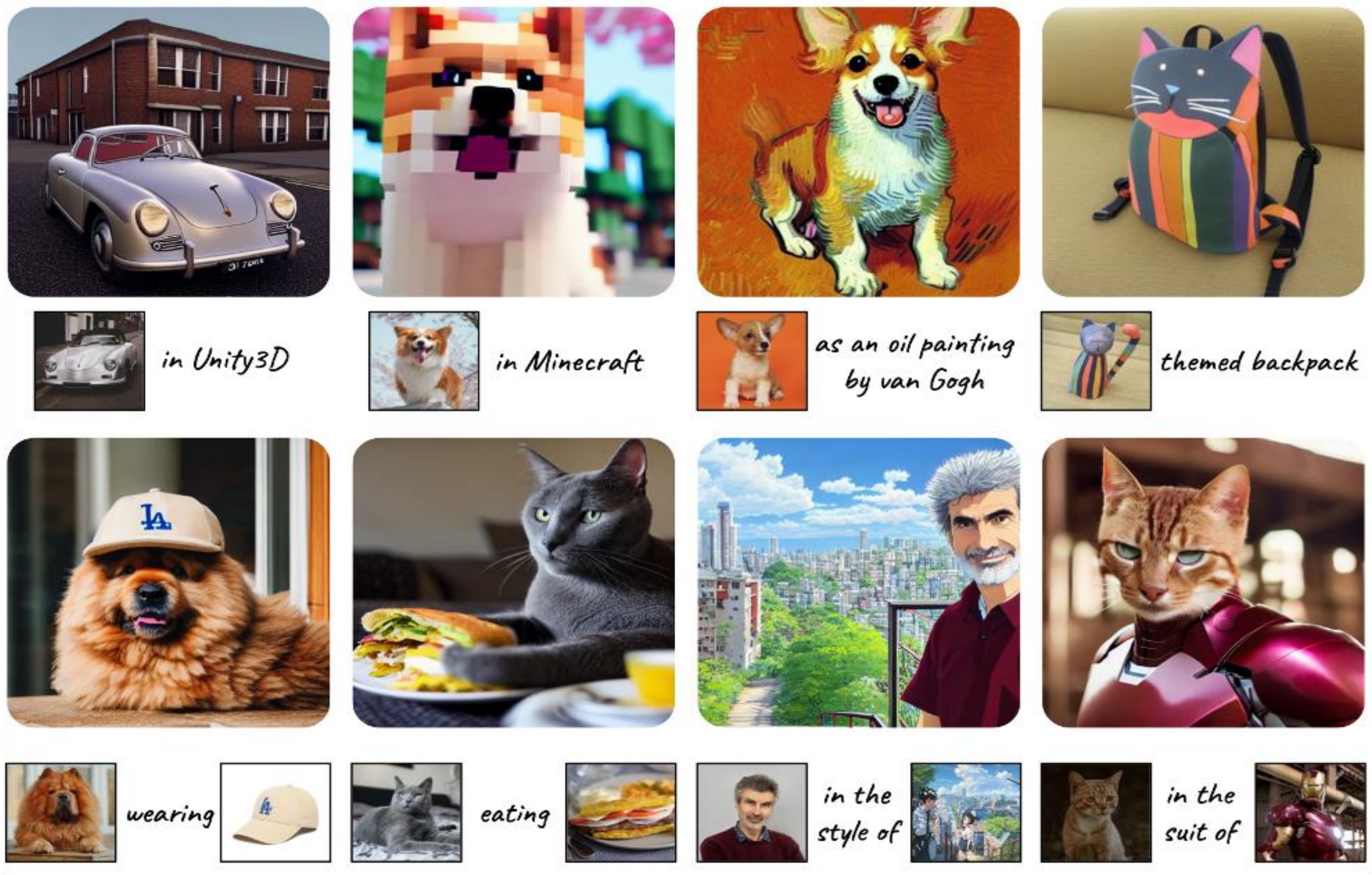
- An interleaved vision-language data has been curated and
- KOSMOS-G is further fine-tuned using the diffusion loss in Equation 3.

$$\mathcal{L}_{diff} = \mathbb{E}_{\mathbf{z}_0, \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\right\|^2\right] \quad (3)$$



KOSMOS-G: Generating Images in Context with Multimodal Large Language Models

KOSMOS-G delivers impressive zero-shot generation results across diverse settings, yielding meaningful and coherent outputs even for highly customized subjects.


in Unity3D
in Minecraft
as an oil painting by van Gogh
themed backpack
wearing
eating
in the style of
in the suit of