香港中文大學
The Chinese University of Hong Kong

LLaMA-VID:

An Image is Worth 2 Tokens in Large Language Models

Yanwei Li, Chengyao Wang, Jiaya Jia

CUHK

Introduction

How to integrate Vision into Language Model?

3 key parts in current Vision Language Model (VLM):

Representation

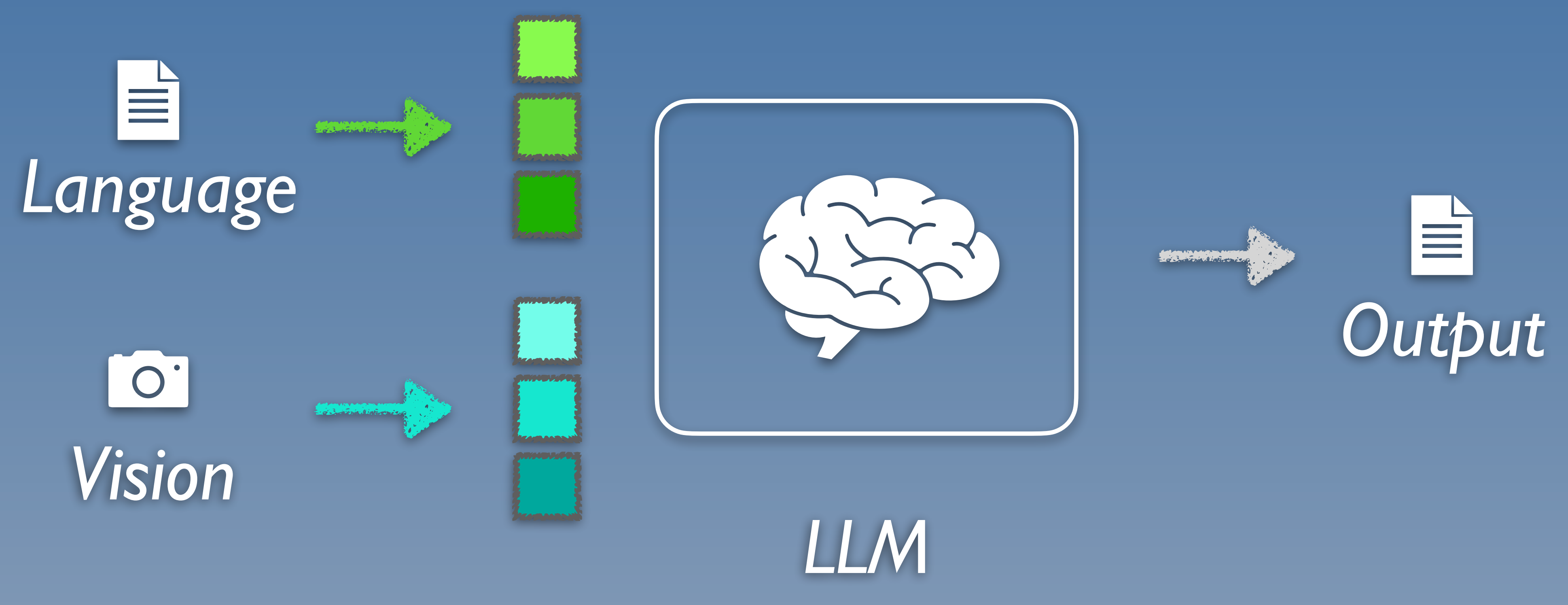
- *Language —> Tokenizer —> Text Token*
- *Vision —> Transformer —> Image Token*

Processing

- *Process tokens from different modalities in LLM*

Prediction

- *Predict text or images from the generated token*



General pipeline of current VLM

Introduction

How to integrate Vision into Language Model?

3 key parts in current Vision Language Model (VLM):

Representation

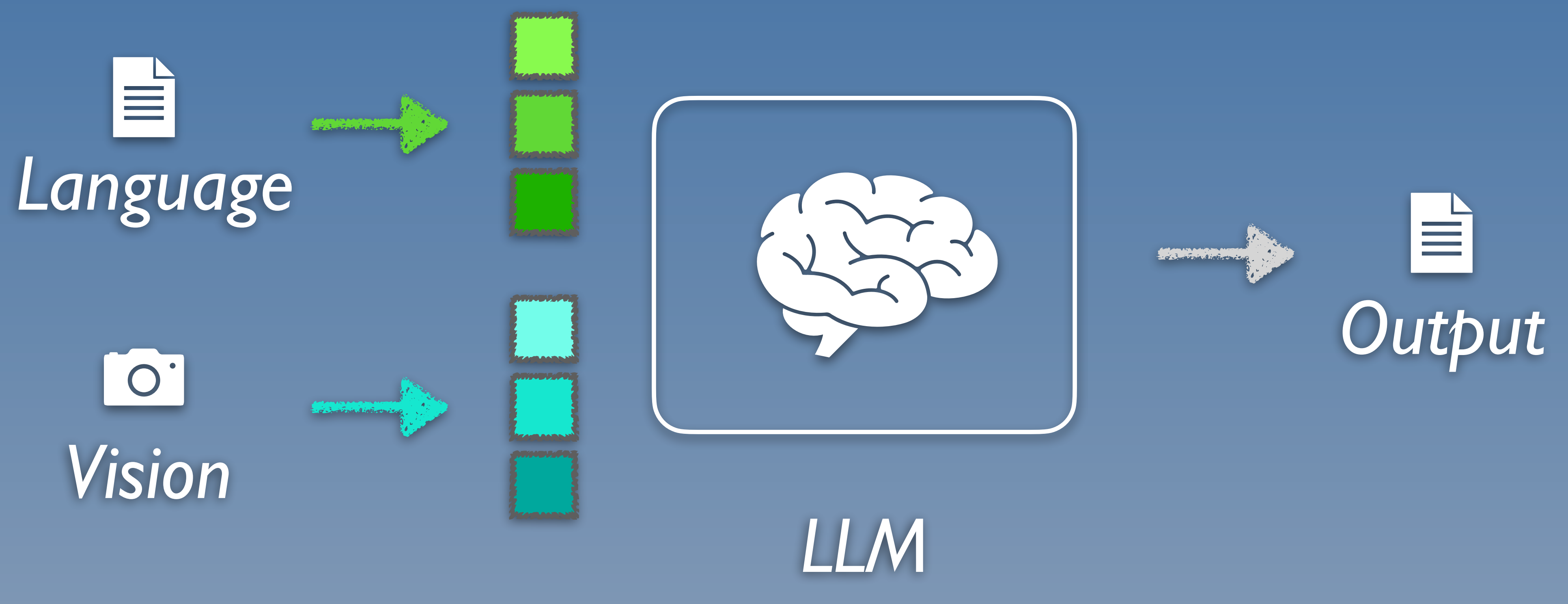
- *Language —> Tokenizer —> Text Token*
- *Vision —> Transformer —> Image Token*

Processing

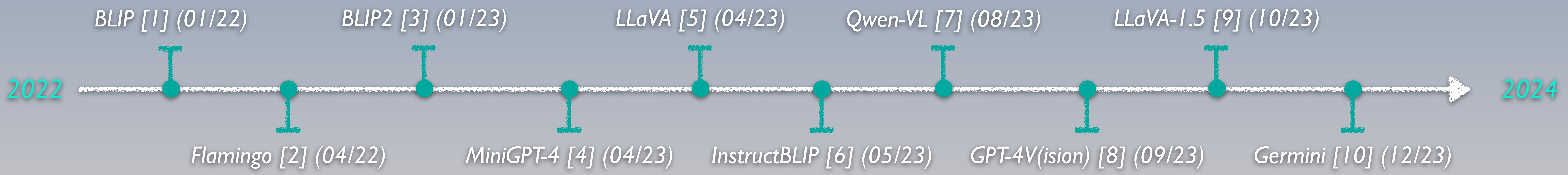
- *Process tokens from different modalities in LLM*

Prediction

- *Predict text or images from the generated token*

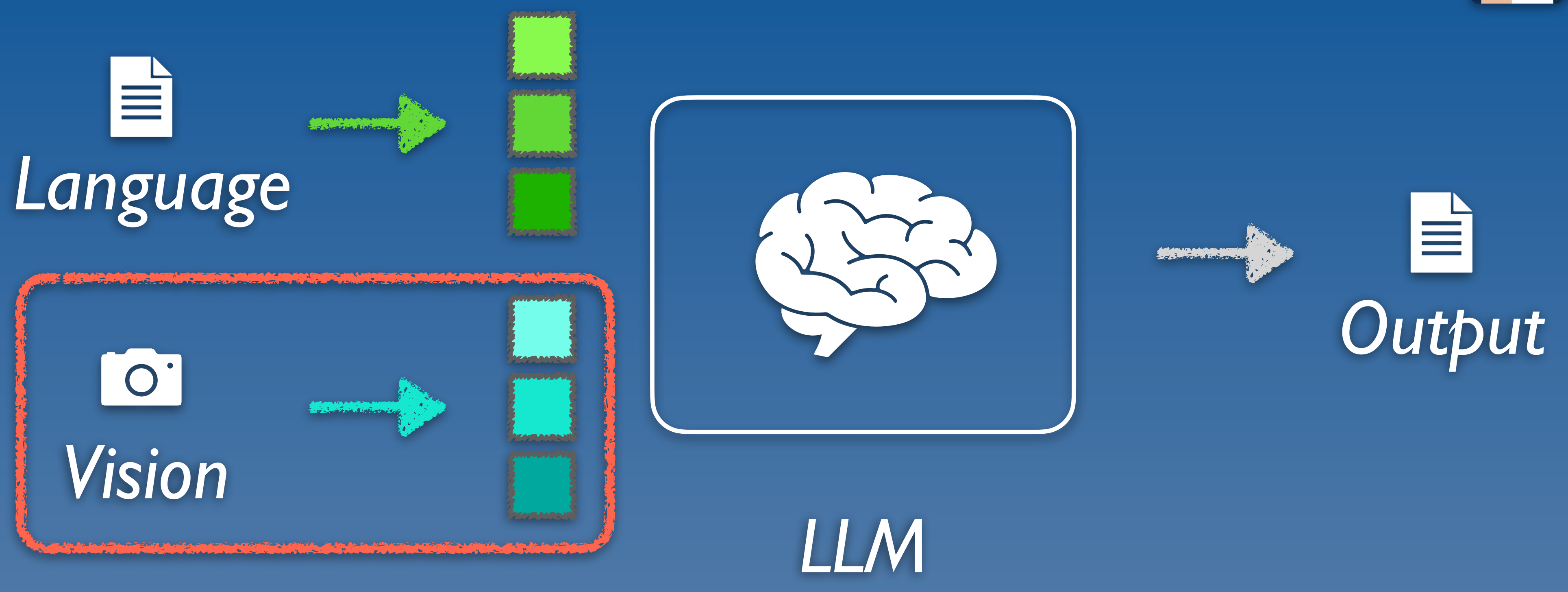


General pipeline of current VLM



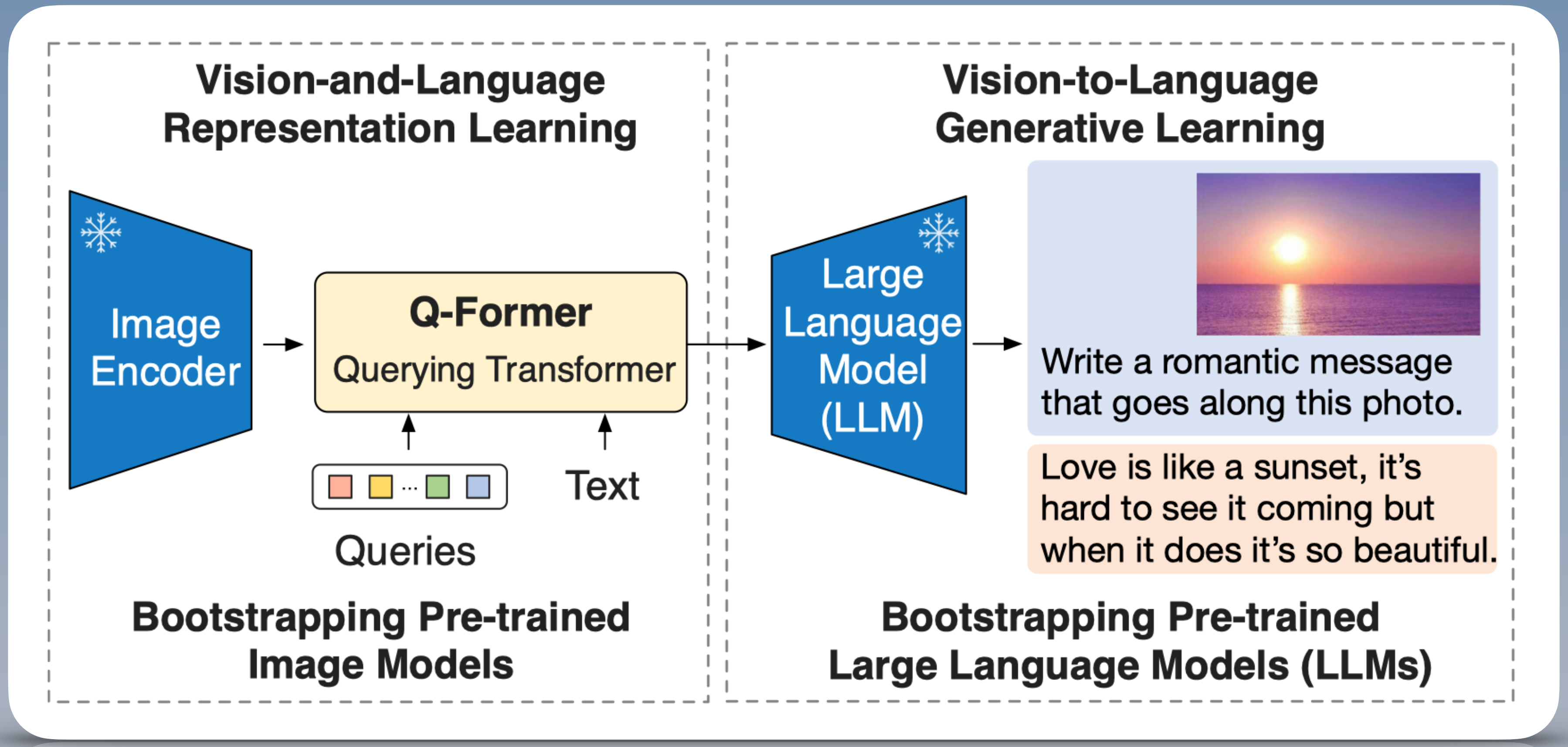
Introduction

How to integrate Vision into Language Model?



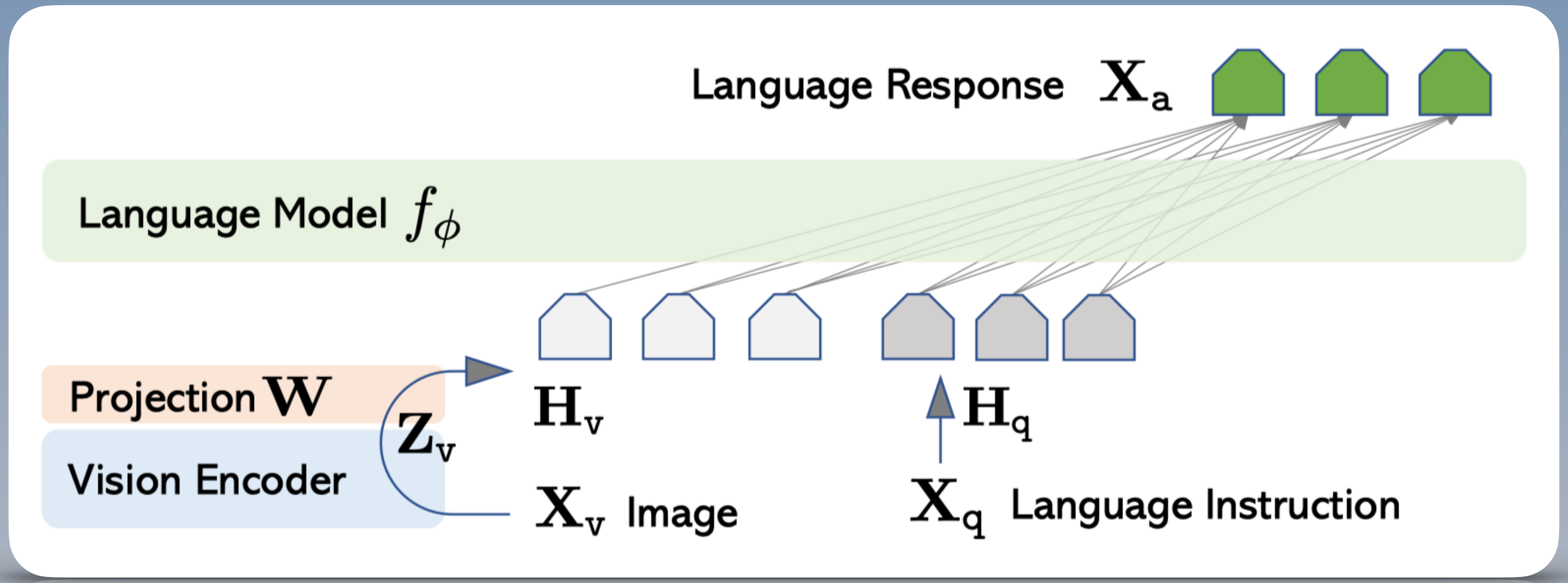
Token Generation

- *Query-based: Use N pre-trained queries to represent each image with N tokens, like Flamingo [2] (N=64), BLIP2 [3] (N=32).*
- *Projector-based: Directly project patch-wise features from ViT to vision tokens increasing with image size, like LLaVA [5].*



Query-based Vision token generation in BLIP2 [3].

32 queries for each image.

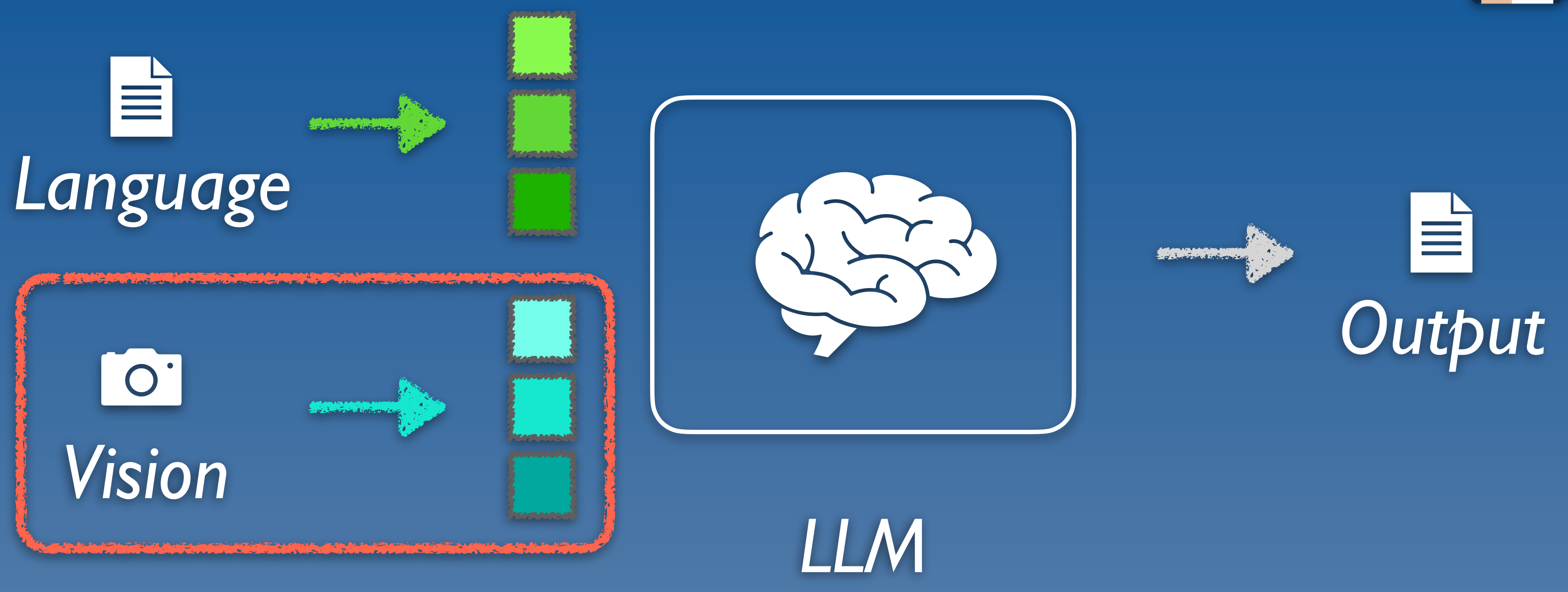


Projector-based Vision token generation in LLaVA [5].

256 queries for each image with 224 size.

Introduction

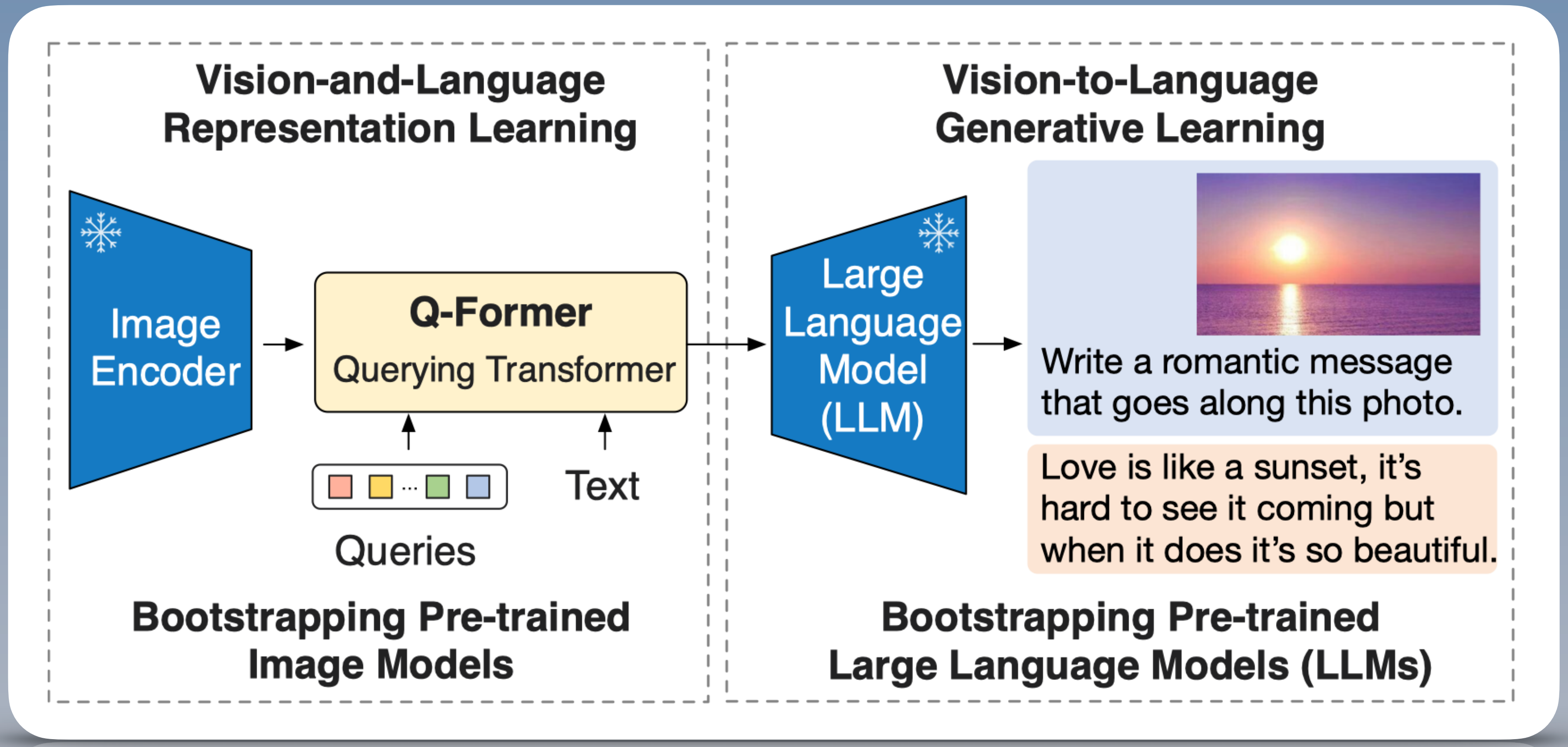
What if we want to process a 3-hour video?



Token Cost

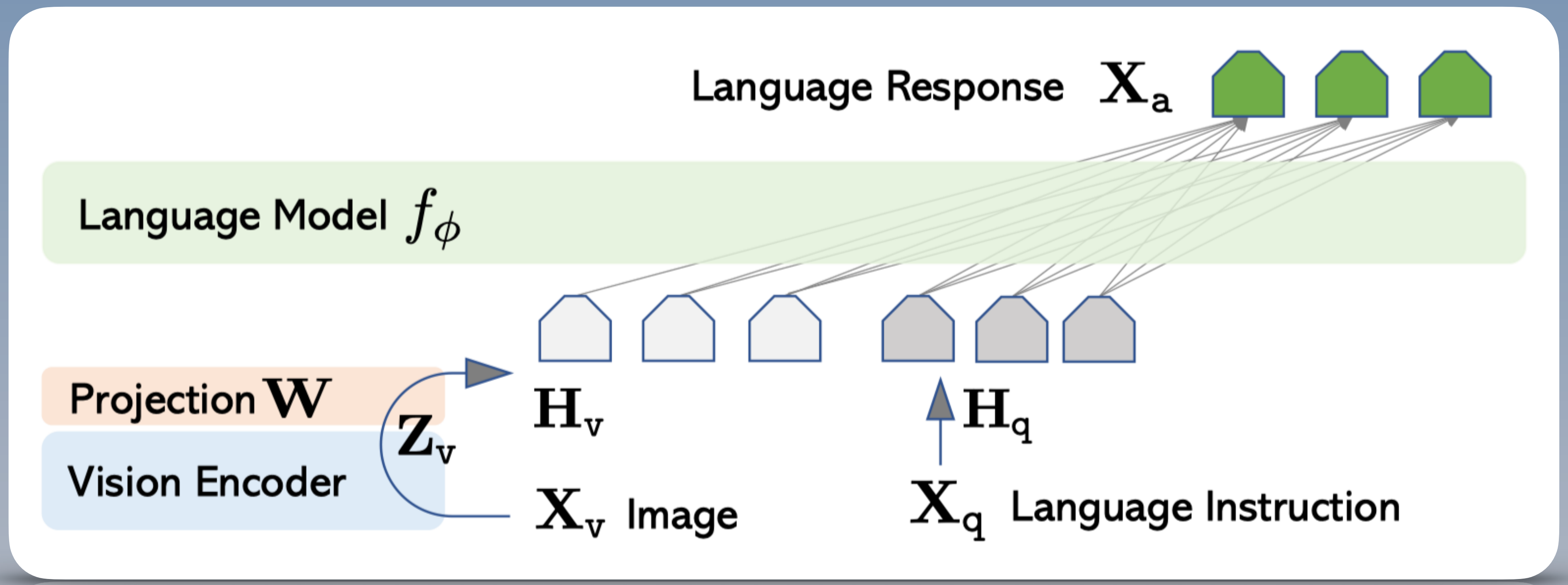
- *Query-based: if sampled with 1FPS, Flamingo (640K), BLIP2 (320K).*
- *Projector-based: if sampled with 1FPS, LLaVA (2.5M).*

We cannot afford such high token cost in current LLMs!



Query-based Vision token generation in BLIP2 [3].

32 queries for each image.



Projector-based Vision token generation in LLaVA [5].

256 queries for each image with 224 size.

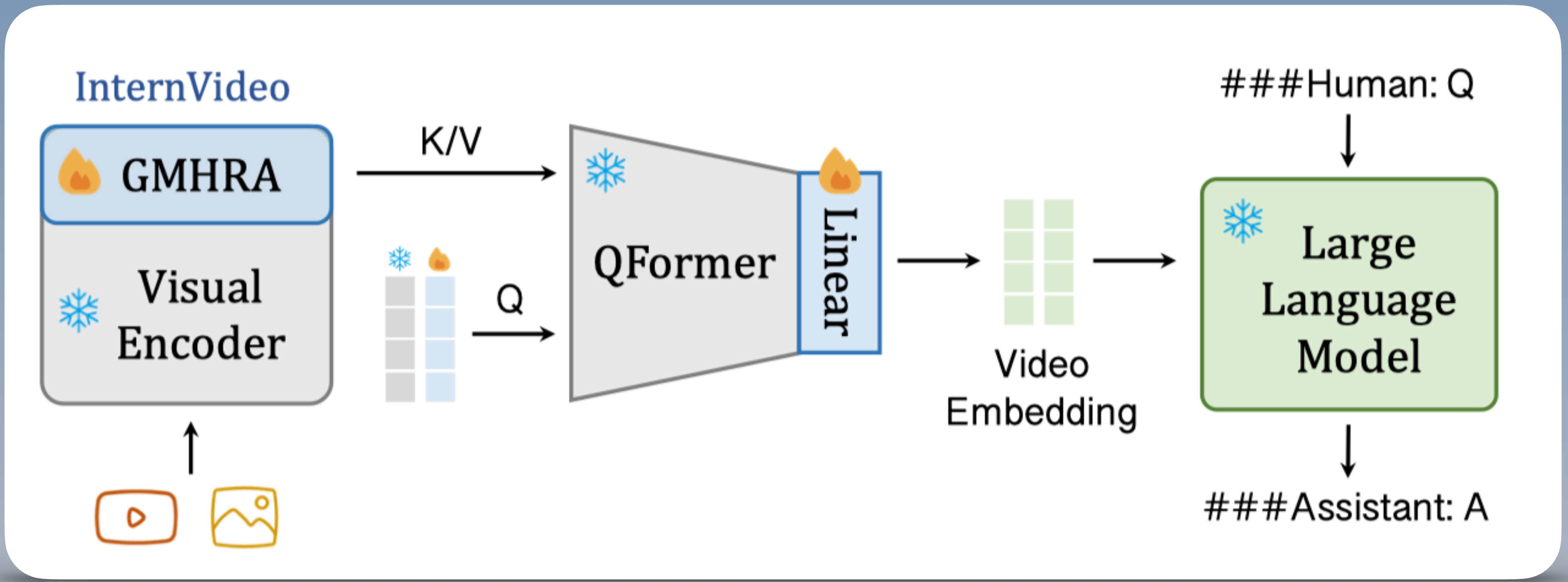
Introduction

How current VLM deal with video?

Current VLM for video

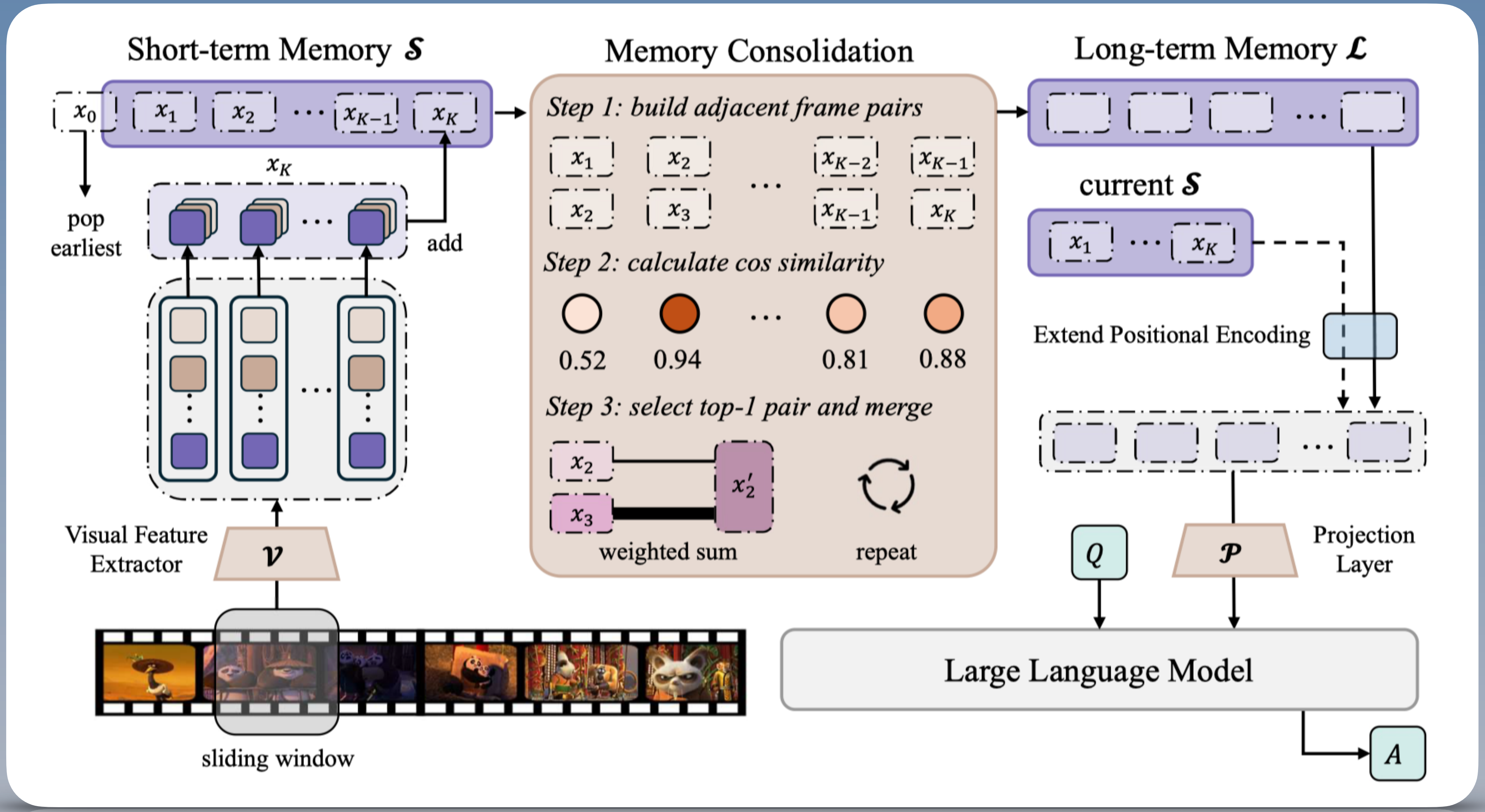
- *Concatenation:* directly concat query-based tokens from all frames, like VideoChat [11], Video-LLaMA [12].
- *Memory-based:* consolidate short-term tokens or use temporal pooling to reduce consumption, like MovieChat [13], Video-ChatGPT [14].

Current VLMs cannot well handle long videos!



Query-based token concatenation in VideoChat [11].

Cannot handle long video with 32 tokens/frame



Memory-based token generation in MovieChat [13].

Cannot well handle subtitles or audios in each frame.

Introduction

How GPT-4V deal with video?

You

GTA6.mp4
File

Describe this video in detail

ChatGPT

To provide a detailed description of the video you've uploaded, I'll need to watch it first. Please give me a moment to review the content of the video.

Finished analyzing

The video you've uploaded is approximately 90 seconds long, with a frame rate of 30 frames per second, and a resolution of 640×360 pixels. I will now proceed to watch the video and provide a detailed description. This may take a few moments. [>-]

I have extracted five frames from different points in the video to get a representative overview of its content. I will describe each frame in detail to give you an understanding of the video's content. Let's start with the first frame. [>-]

Frame 1

Frame 1 Description

- **Setting**: Appears to be an urban environment, likely a city street. Buildings can be seen in the background.
- **Characters**: There are no visible characters in this frame.
- **Vehicles**: There's a car, possibly a sedan, in the foreground. Its make and model are not clearly discernible.
- **Time of Day**: It looks like daytime, judging by the lighting and shadows.
- **Other Details**: The overall color tone is realistic, and there are street elements like road markings and possibly traffic lights or signs.

Frame 4 Description

- **Setting**: This frame continues the urban setting theme. It shows a street corner, with buildings and a pedestrian crossing.
- **Characters**: There are several characters, mostly pedestrians. They appear to be engaged in various activities like walking or standing, possibly waiting to cross the street.
- **Vehicles**: The street is filled with cars, suggesting a busy urban area. The variety of vehicles, including sedans and larger vehicles, adds to the bustling city atmosphere.
- **Time of Day**: It's still daytime, with clear visibility and shadows indicating sunlight.
- **Other Details**: The presence of traffic lights, road markings, and signage is evident. There's also a sense of movement and daily life in a city.

Finally, I will describe the fifth and last frame. [>-]

Frame 5

Frame 5 Description

- **Setting**: Similar to the previous frames, this one is set in an urban environment. The focus is on a section of the street with buildings on either side.
- **Characters**: This frame features several characters, likely pedestrians, who are distributed across the sidewalk and near the street. Their activities are not clearly discernible, but they seem to be going about their daily routines.
- **Vehicles**: There are cars on the street, indicating ongoing traffic. The variety of cars suggests a typical city scene.
- **Time of Day**: As in the earlier frames, it appears to be daytime, with sunlight casting shadows on the street.
- **Other Details**: Urban elements like street lamps, traffic signs, and possibly shopfronts or business entrances are visible. The frame captures the essence of a bustling city environment.

GPT-4V process 5 sampled frames from the video (1.5 min) and generate response frame-by-frame.
It cannot handle video as a whole!

You

Try to describe the whole video without frame-by-frame analysis.

ChatGPT

To provide a description of the entire video without a frame-by-frame analysis, I will watch it in its entirety and then summarize the overall content and themes. This process will take a few moments.

Error analyzing

Unfortunately, I encountered a technical limitation that prevents me from watching the entire video in one go. However, I can still provide a comprehensive description by analyzing the video in segments. I'll proceed with this approach and share my observations shortly. [>-]

Error analyzing

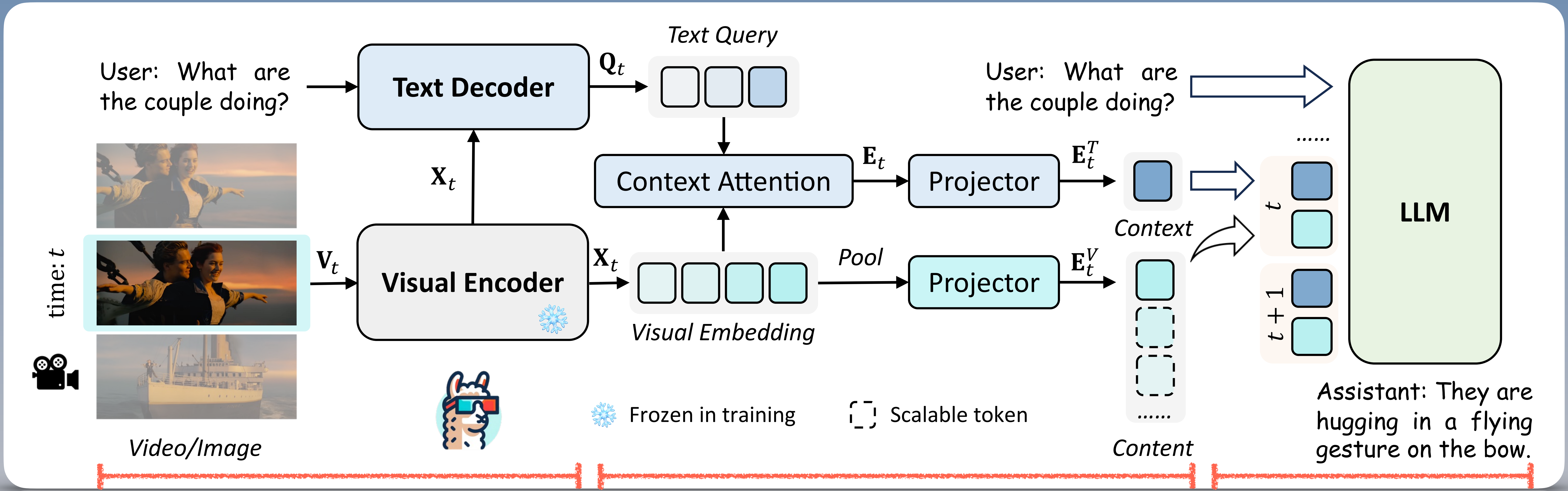
It seems I encountered another technical issue while attempting to analyze the video in segments. Given these limitations, I'm unable to watch the video directly or analyze it in larger segments.

However, if you can provide a brief summary or key points about the video, I can offer insights or further information based on that description. Alternatively, if there are specific aspects of the video you are curious about or want more information on, please let me know, and I'll do my best to assist you. [>-]

LLaMA-VID

Overview

- *Encoder and Decoder:* encode video frames/image and produce text queries from the text decoder.
- *Token Generation:* generate context token and content token with the context attention and projectors.
- *Training Strategy:* optimize LLM and the framework with designed data and training strategy.

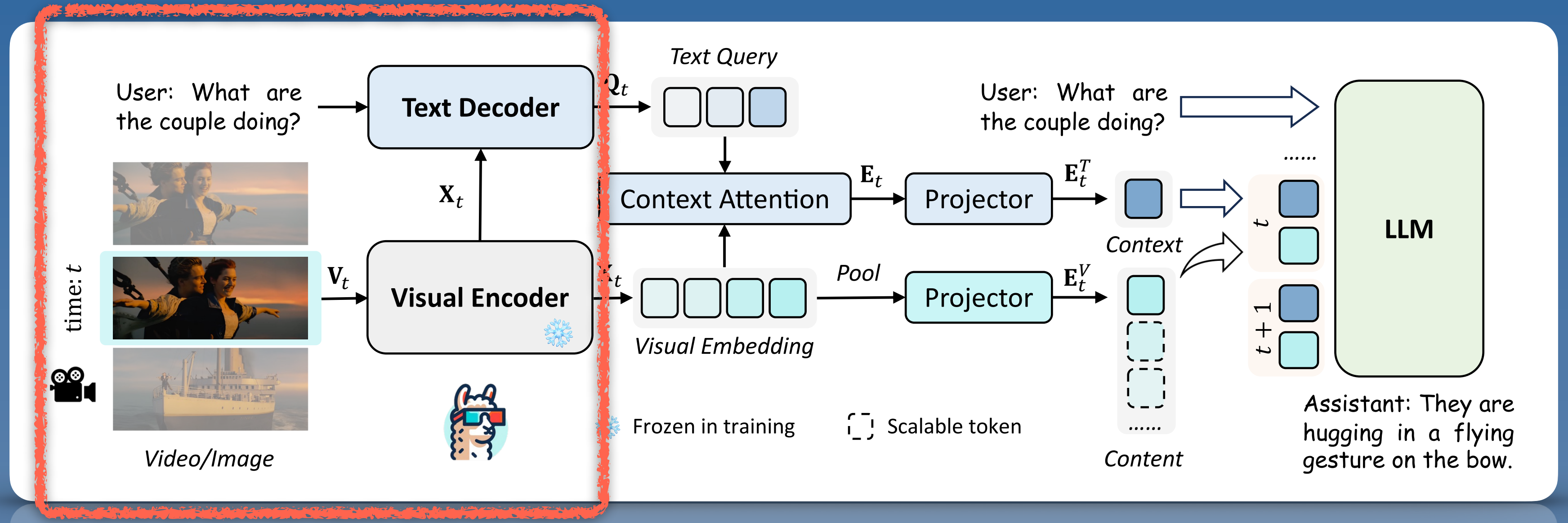


The framework of LLaMA-VID [15].

LLaMA-VID

Encoder and Decoder

- *Visual Encoder:* ViT-based visual encoder for visual embedding.
- *Text Decoder:* BERT-based text decoder to generate instruction-guided text queries, which can be instantiated with QFormer.



LLaMA-VID

Encoder and Decoder

- Visual Encoder: ViT-based visual encoder for visual embedding.
- Text Decoder: BERT-based text decoder to generate instruction-guided text queries, which can be instantiated with QFormer.

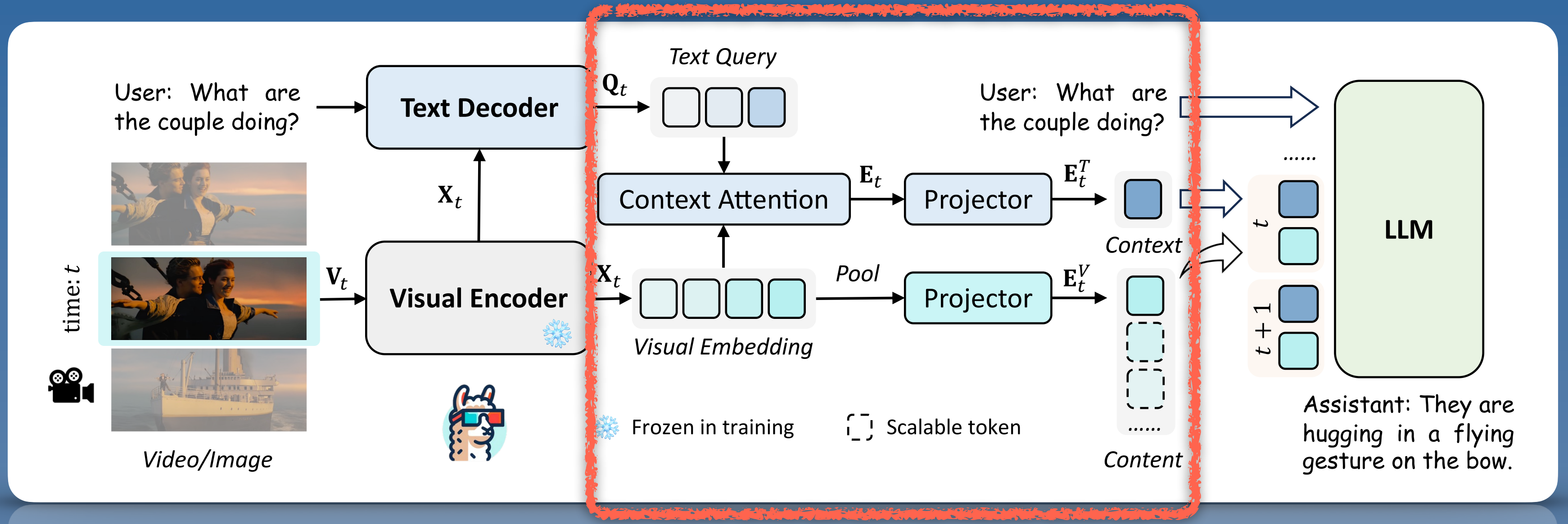
Token Generation

- Context Token: generate context-related embedding with the text queries and visual embedding in Context Attention:

$$\mathbf{E}_t = \mathrm{Mean}(\mathrm{Softmax}(\mathbf{Q}_t \times \mathbf{X}_t^{\mathrm{T}}) \times \mathbf{X}_t)\,.$$

- Content Token: adaptive pooling strategy for the visual embedding according to computational constraints.

- 2 Token/Frame: concat context token and content token to represent each frame in video.



Algorithm 1 Pseudo Code for Token Generation.

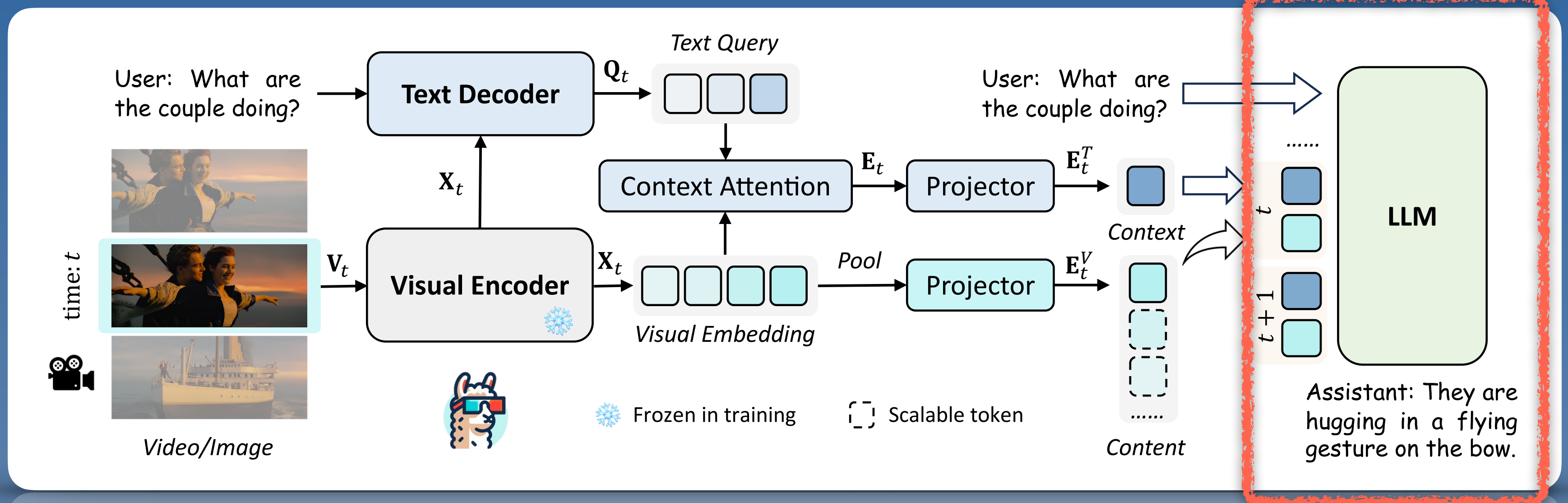
```
# B: batch size; C: channel size; n: content shape
# M: query length; N: shape of flatten image pacthes;
# text_q: text query in shape (B, M, C)
# vis_embed: visual embedding in shape (B, N, C)

# Key part 1: calculate context-related embedding
ctx_embed = text_q @ vis_embed.transpose(-1,-2)
ctx_embed = ctx_embed / (vis_embed.shape[-1]**0.5)
ctx_embed = (ctx_embed.softmax(-1)@vis_embed).mean(1)
ctx_embed = self.ctxproj(ctx_embed[:,None])

# Key part 2: calculate visual embedding
cur_shape = int(vis_embed.shape[1]**0.5)
vis_embed = vis_embed.reshape(B, cur_shape, -1, C)
vis_embed = F.avg_pool2d(vis_embed.permute(0,3,1,2),
    kernel_size=cur_shape//n, stride=cur_shape//n)
vis_embed = vis_embed.permute(0,2,3,1).flatten(1,2)
vis_embed = self.visproj(vis_embed)

# concat token in shape (B, n+1, C), n in [1,N]
final_token = torch.cat([ctx_embed, vis_embed], dim=1)
```

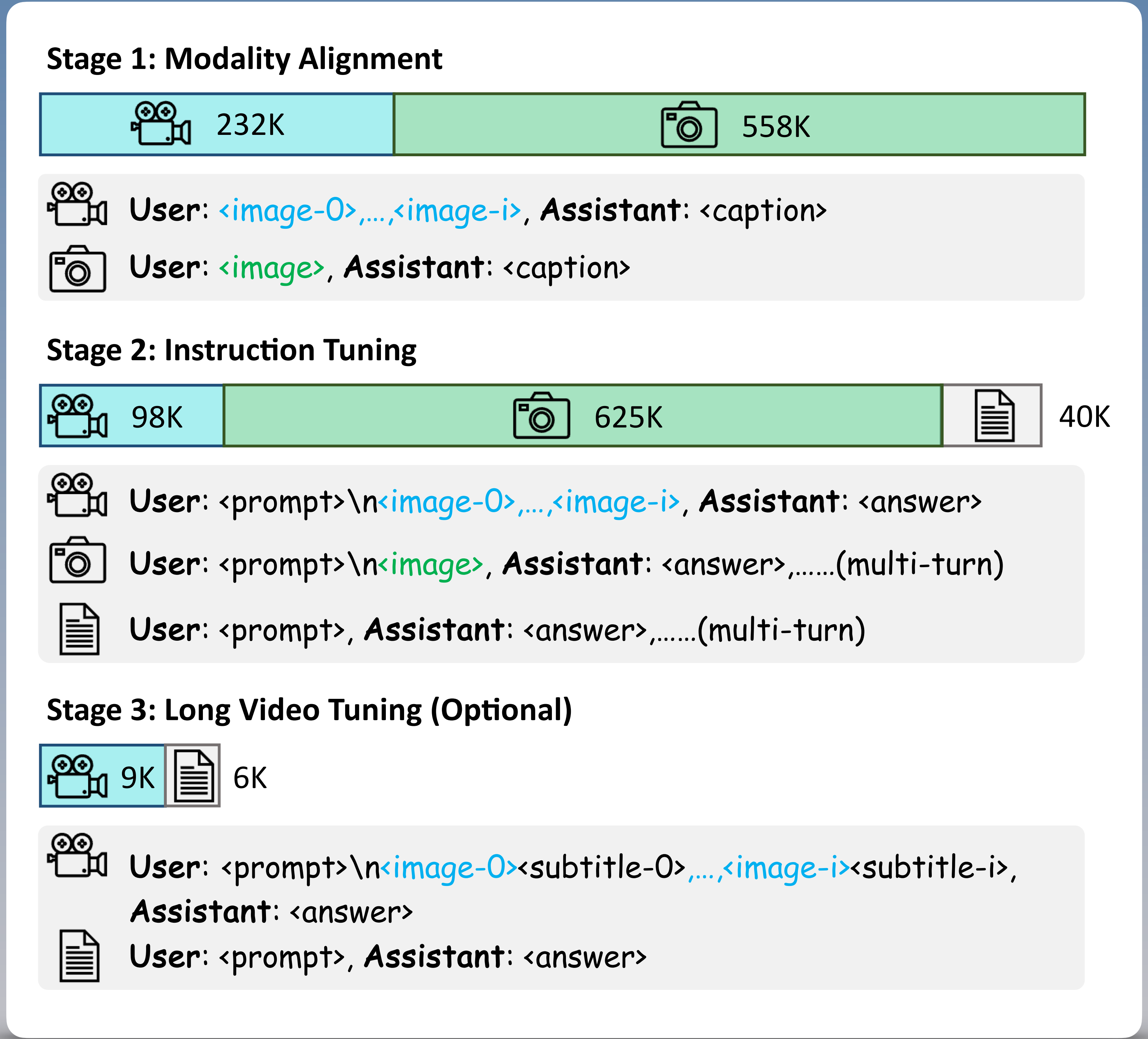
F: torch.nn.functional; `ctxproj`, `visproj`: predefined linear projectors.

LLaMA-VID

Training Strategy

- *Modality Alignment:* *optimize projectors with 232K video caption pairs and 558K image caption pairs.*
- *Instruction Tuning:* *optimize Text Decoder, projectors, and LLMs with 98K video pairs, 625K image pairs, and 40K text pairs.*





Settings	Stage 1	Stage 2	Stage 3
Batch size	256	128	8
Learning rate	1e-3	2e-5	2e-5
Learning schedule		Cosine decay	
Warmup ratio		0.03	
Weight decay		0	
Epoch		1	
Optimizer		AdamW	
DeepSpeed stage		2	
Vision encoder		Freeze	
Text decoder	Freeze	Open	Freeze
Max token	2048	2048	65536

LLaMA-VID

Training Strategy

- *Modality Alignment:* optimize projectors with 232K video caption pairs and 558K image caption pairs.
- *Instruction Tuning:* optimize Text Decoder, projectors, and LLMs with 98K video pairs, 625K image pairs, and 40K text pairs.
- *Long Video Tuning:* optimize LLM to support hour-long videos with collected 9K long video pairs and 6K long text pairs.
- *Long VideoQA dataset:* generate 6K question-answer pairs using GPT-4 and Claude-2 using movie synopsis and scripts.

Settings	Stage 1	Stage 2	Stage 3
Batch size	256	128	8
Learning rate	1e-3	2e-5	2e-5
Learning schedule		Cosine decay	
Warmup ratio		0.03	
Weight decay		0	
Epoch		1	
Optimizer		AdamW	
DeepSpeed stage		2	
Vision encoder		Freeze	
Text decoder	Freeze	Open	Freeze
Max token	2048	2048	65536

Video Frame: Only for illustration here, not used to produce instruction data.

Synopsis for the whole movie

Synopsis: Young Rose, angry and distraught that her mother has apparently a rranged the marriage, considers committing suicide by jumping from the ster n; Jack manages to pull her back over the rail after she loses her footing……

Instruction pairs for movie summary

User: <prompt> Create 2 plot summary of this movie. The first one should be a brief summary written in one paragraph. The second one should be a det ail summary written in multiple paragraphs……

GPT-4: Brief Summary:……, Detail Summary:……

Instruction pairs for movie plot and characters

User: <prompt> Create 5 questions about the movie plot, including plot und erstanding, plot description, plot analysis, etc. Create 5 questions about char acters, including relationship, personality, behavior……

GPT-4: Question: What ultimately happens to the Heart of the Ocean ne cklace?. Answer:……

Script for the whole movie

Script: Rose runs along the B deck promenade. She is dishevelled, her hair fl ying. She is crying, her cheeks streaked with tears. But also angry, furious! S haking with emotions she doesn't understand… hatred, self-hatred……

Instruction pairs for movie reasoning and details

User: <prompt> Create 5 complex questions about plot reasoning rather th an simply describe the plot. Create 5 complex questions about detail scene an d activity description……

Claude-2: Question:Why doesn't Rose get in the lifeboat with her mother when she has a chance. Answer: ……

Results & Analysis

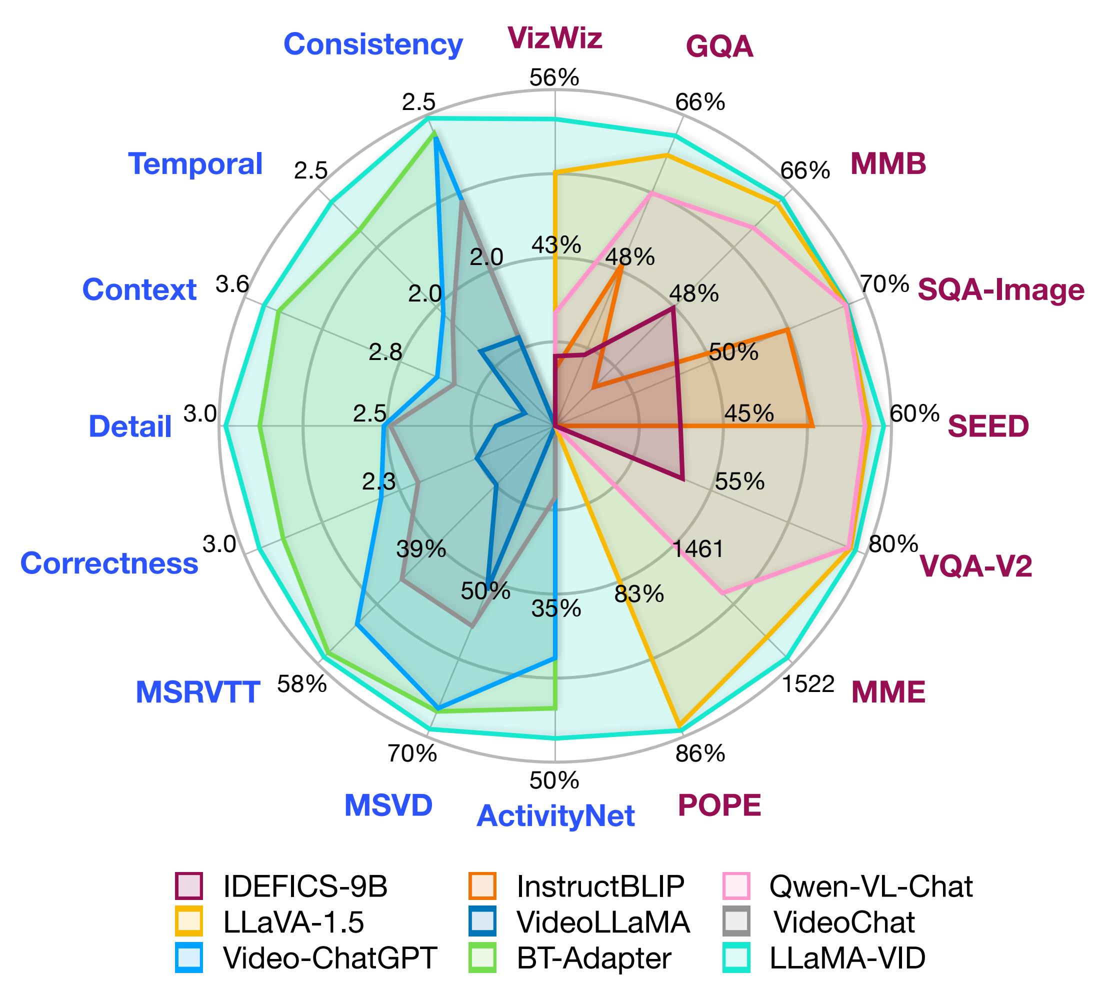
Results

- *Video-based benchmarks: achieve top rank on 4 video-based QA benchmarks with 2 tokens for each frame.*

Comparisons with different VLMs on video-based benchmarks.

Method	LLM	Res.	MSVD-QA Acc	MSVD-QA Score	MSRVTT-QA Acc	MSRVTT-QA Score	ActivityNet-QA Acc	ActivityNet-QA Score
FrozenBiLM [59]	DeBERTa-V2	224	32.2	–	16.8	–	24.7	–
VideoLLaMA [62]	Vicuna-7B	224	51.6	2.5	29.6	1.8	12.4	1.1
LLaMA-Adapter [63]	LLaMA-7B	224	54.9	3.1	43.8	2.7	34.2	2.7
VideoChat [30]	Vicuna-7B	224	56.3	2.8	45.0	2.5	26.5	2.2
Video-ChatGPT [39]	Vicuna-7B	224	64.9	3.3	49.3	2.8	35.2	2.7
BT-Adapter [34]	Vicuna-7B	–	67.5	**3.7**	57.0	3.2	45.7	3.2
LLaMA-VID	Vicuna-7B	224	69.7	**3.7**	57.7	3.2	47.4	**3.3**
LLaMA-VID	Vicuna-13B	224	**70.0**	**3.7**	**58.9**	**3.3**	**47.5**	**3.3**

Method	LLM	Res.	Correctness	Detail	Context	Temporal	Consistency
VideoLLaMA [62]	Vicuna-7B	224	1.96	2.18	2.16	1.82	1.79
LLaMA-Adapter [63]	LLaMA-7B	224	2.03	2.32	2.30	1.98	2.15
VideoChat [30]	Vicuna-7B	224	2.23	2.50	2.53	1.94	2.24
Video-ChatGPT [39]	Vicuna-7B	224	2.40	2.52	2.62	1.98	2.37
BT-Adapter [34]	Vicuna-7B	–	2.68	2.69	3.27	2.34	2.46
LLaMA-VID	Vicuna-7B	224	2.96	3.00	3.53	2.46	2.51
LLaMA-VID	Vicuna-13B	224	**3.07**	**3.05**	**3.60**	**2.58**	**2.63**



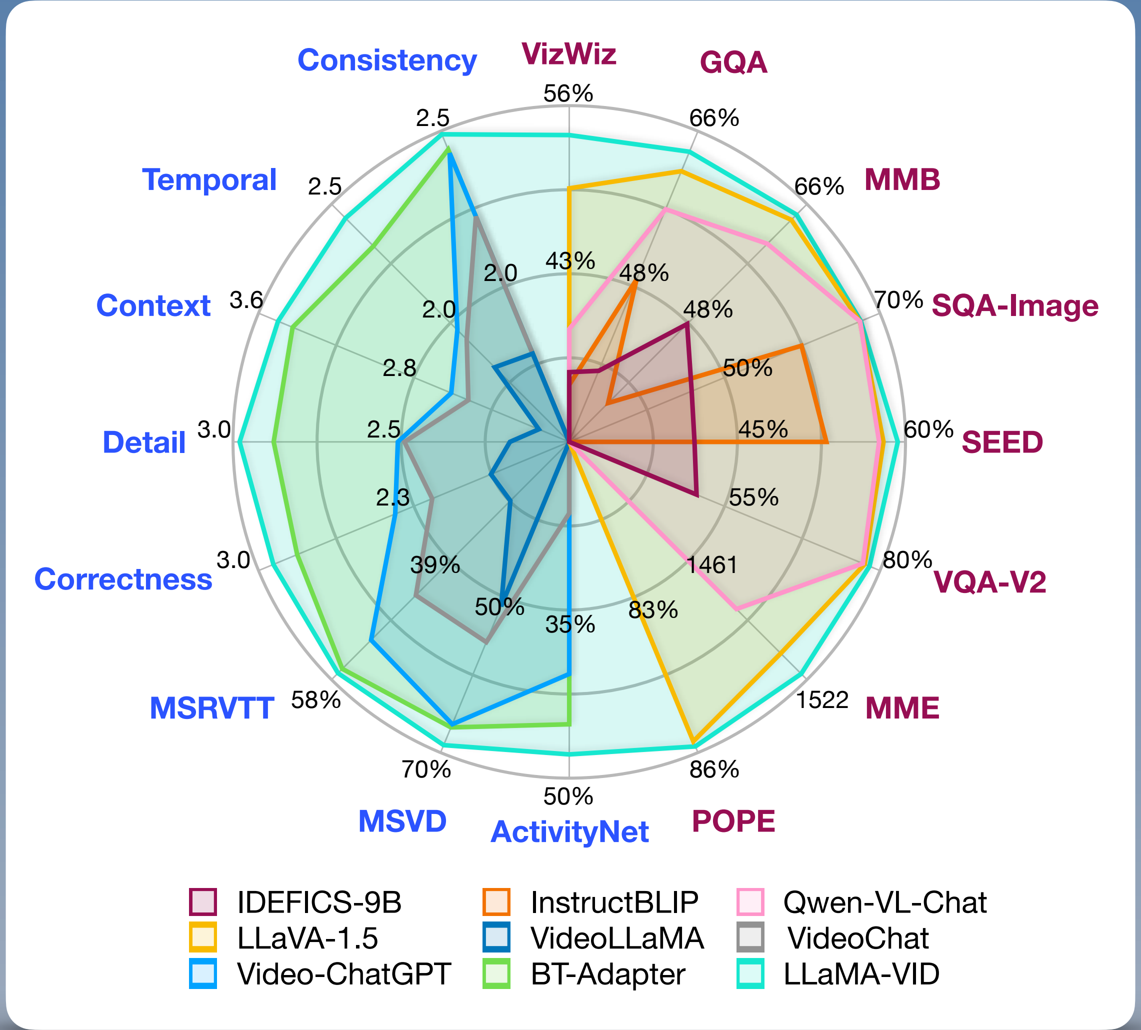
Results & Analysis

Results

- *Video-based benchmarks: achieve top rank on 4 video-based QA benchmarks with 2 tokens for each frame.*
- *Image-based benchmarks: achieve top rank on 8 image-based QA benchmarks with +1 context token in LLaVA.*

Comparisons with different VLMs on image-based benchmarks.

Method	LLM	Res.	GQA	MMB	MME	POPE	SEED	SQA^{I}	VizWiz	VQA^{v2}
InstructBLIP [14]	Vicuna-7B	224	49.2	36.0	–	–	53.4	60.5	34.5	–
IDEFICS-9B [23]	LLaMA-7B	224	38.4	48.2	–	–	–	–	35.5	50.9
Qwen-VL† [4]	Qwen-7B	448	59.3*	38.2	–	–	56.3	67.1	35.2	78.8*
Qwen-VL-Chat† [4]	Qwen-7B	448	57.5*	60.6	1487.5	–	58.2	68.2	38.9	78.2*
LLaVA-1.5 [32]	Vicuna-7B	336	62.0*	64.3	1510.7	85.9	58.6	66.8	50.0	78.5*
LLaMA-VID	Vicuna-7B	336	**64.3***	**65.1**	**1521.4**	**86.0**	**59.9**	**68.3**	**54.2**	**79.3***
BLIP-2 [29]	Vicuna-13B	224	41.0	–	1293.8	85.3	46.4	61.0	19.6	41.0
InstructBLIP [14]	Vicuna-13B	224	49.5	–	1212.8	78.9	–	63.1	33.4	–
Shikra [9]	Vicuna-13B	224	–	58.8	–	–	–	–	–	77.4*
IDEFICS-80B [23]	LLaMA-65B	224	45.2	54.5	–	–	–	–	36.0	60.0
LLaVA-1.5 [32]	Vicuna-13B	336	63.3*	**67.7**	1531.3	85.9	61.6	**71.6**	53.6	**80.0***
LLaMA-VID	Vicuna-13B	336	**65.0***	66.6	**1542.3**	**86.0**	**62.3**	70.0	**54.3**	**80.0***



Results & Analysis

Analysis

- *Token Type:* *both context token and content token contribute to the performance for image-based QA.*
- *Token Number:* *generally, more content tokens bring better performance for image-based QA.*
- *Text Decoder:* *both raw BERT or pre-trained QFormer bring better results, while QFormer is better.*

Ablation study on token number.

context	*content*	**GQA**	**POPE**	**SQA^I**	**VQA^T**
0	256	61.9	85.5	67.5	53.0
1	256	**63.0**	**86.6**	67.7	**53.8**
1	64	60.8	85.1	68.7	52.3
1	16	58.2	83.1	67.4	50.8
1	4	56.2	83.5	68.7	49.1
1	1	55.5	83.1	**68.8**	49.0

Ablation study on token type.

context	*content*	**GQA**	**POPE**	**SQA^I**	**VQA^T**
✗	✓	53.3	80.9	66.1	46.5
✓	✗	54.3	82.4	67.7	48.3
✓	✓	**55.5**	**83.1**	**68.8**	**49.0**

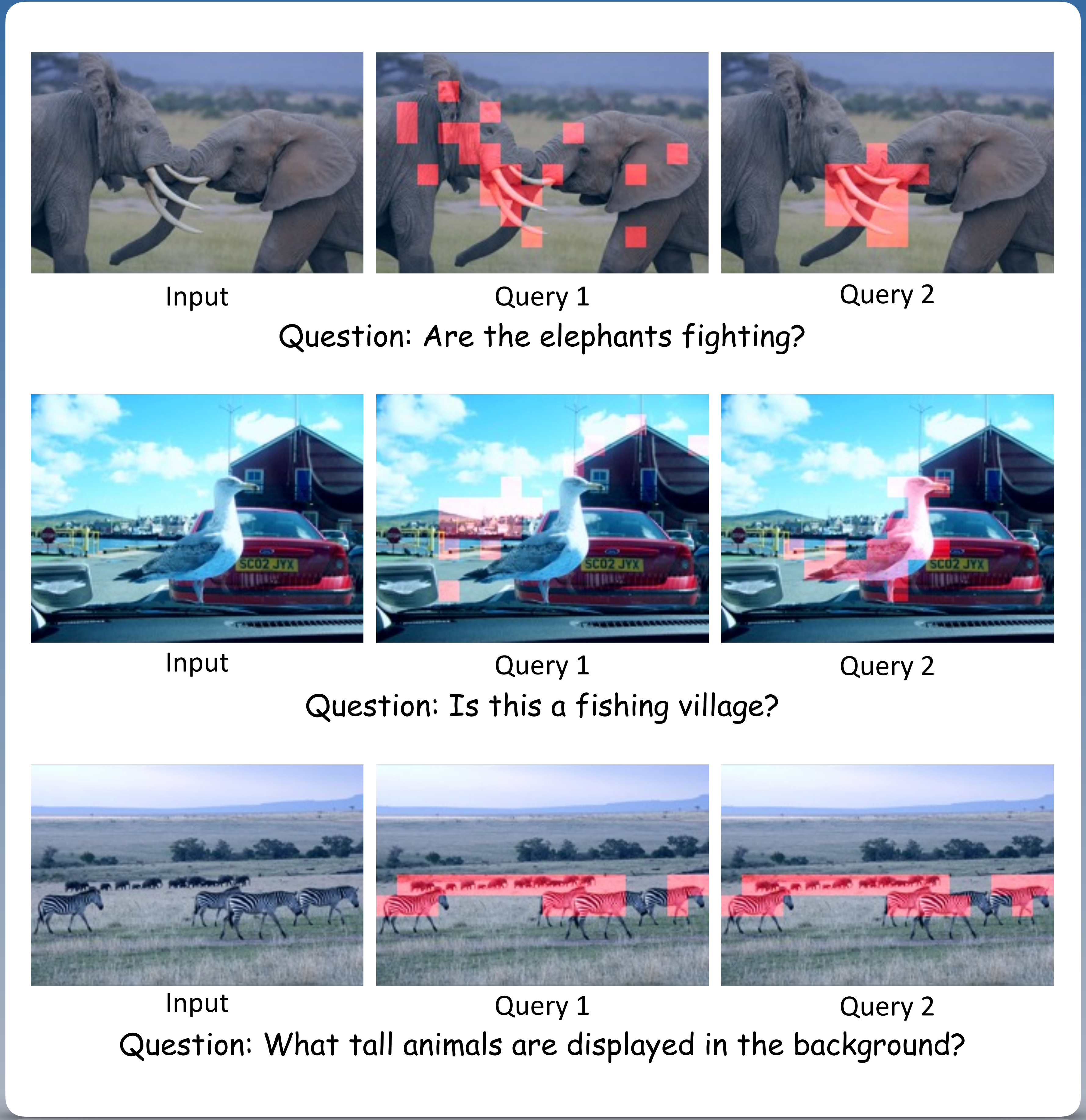
Ablation study on text decoder.

text	**GQA**	**POPE**	**SQA^I**	**VQA^T**
–	53.3	80.9	66.1	46.5
BERT	54.1	80.8	67.9	48.1
QFormer	**55.5**	**83.1**	**68.8**	**49.0**

Results & Analysis

Analysis

- Token Type: both context token and content token contribute to the performance for image-based QA.
- Token Number: generally, more content tokens bring better performance for image-based QA.
- Text Decoder: both raw BERT or pre-trained QFormer bring better results, while QFormer is better.
- Response in Context Attention: the text-guided query effectively focuses on important areas relevant to the input questions.



High response areas with top scores to input questions.

Results & Analysis

Input with 3-hour movie Interstellar

LLaMA-VID

User: How did Cooper transmit data from the black hole to Murphy?

Assistant: Cooper transmitted data from the black hole to Murphy by using a watch that had been given to Murphy by Cooper.

User: Why does time pass slower on Miller's planet than on Earth? How slowly does it go?

Assistant: Time passes slower on Miller's planet because it is in orbit around a black hole, which has a strong gravitational pull. The stronger the gravitational pull, the slower time moves. On Miller's planet, time moves at about 7 years per hour, compared to Earth's 1 hour per hour. This is because the black hole's gravity is so strong that it distorts space-time, causing time to run more slowly.

LLaMA 2

User: How did Cooper transmit data from the black hole to Murphy?

Assistant: There is no character named Murphy in the movie Interstellar, and there is no scene where a character named Cooper transmits data from a black hole. The movie Interstellar is a work of science fiction and does not depict any real events or technology. Therefore, it is not possible for Cooper to transmit data from a black hole to Murphy, as there is no Murphy character in the movie and no such technology exists.

User: Why does time pass slower on Miller's planet than on Earth? How slowly does it go?

Assistant: In the movie Interstellar, time runs slower on Miller planet than on Earth due to the planet's strong gravitational pull. The stronger the gravitational force, the slower time passes, according to Einstein's theory of general relativity. On Miller planet, the gravitational force is much stronger than on Earth, causing time to slow down significantly. The exact slowing down of time on Miller planet is not explicitly stated in the movie, but it is implied to be around 1.5 to 2 times slower than on Earth. This means that for. This time dilation effect become every hour that passes on Earth, only around 45 minutes would pass on Miller planets more pronounced as you approach the planet's surface, making time appear to slow down even more.

LongLoRA

User: How did Cooper transmit data from the black hole to Murphy?

Assistant: Cooper realizes that TARS, the robot, is the only one standing between him and being able to transmit data from the black hole to Murphy. TARS has the ability to transmit data through radio waves, so Cooper comes up with a plan to have TARS transmit the quantum data he has collected while inside the black hole. Cooper programs TARS to transmit the data in Morse code, which TARS is able to understand and decode. Through this method, Cooper is able to successfully transmit the data from the black hole to Murphy, who is able to receive and understand it.

User: Why does time pass slower on Miller's planet than on Earth? How slowly does it go?

Assistant: Miller planet orbits a black hole, which has a huge gravitational pull. This causes time to run slower on Miller planet compared to Earth. Specifically, time runs 7 years slower on Miller planet for every 1 year spent on Earth. The effect is more pronounced the closer one gets to the black hole. So time runs significantly slower on Miller planet, especially in the vicinity of the black hole.

Comparisons with LLaMA2 [16] and LongLoRA [17] given the same questions.

Results & Analysis

Single Image

User: Please describe this image in detail.

Assistant: The image features a young man in a blue outfit, likely from the game series "The Legend of Zelda," standing on top of a large rock formation.
He appears to be looking out over the landscape, possibly contemplating his next move or enjoying the view. The rock formation is surrounded by a beautiful blue sky, with clouds scattered throughout the scene......

User: Try to write a short poem on it.

Assistant: A man in blue, On a cliff high above, With a sword in his hand, He looks out to the land.

The sky is blue and clear, The clouds are white and fluffy, The man stands tall and strong, A warrior in his own way......

User: What's the name of this young man and game, and where he comes from?

Assistant: The young man in the image is named Link, and he comes from the land of Hyrule, which is the setting for the popular video game series "The Legend of Zelda."

3-min Video
Zootopia Trailer

User: Please describe this video in detail.

Assistant: The video starts with a cartoon character standing in front of a castle, which is surrounded by a beautiful landscape. The character is holding a bucket and appears to be observing the surroundings. The scene then transitions to a train, where the character is seen sitting inside the train. The train is surrounded by lush greenery and appears to be moving through a scenic route......

User: What's the name of the movie described in this video?

Assistant: The movie described in this video is "Zootopia."

User: Who produce the movie?

Assistant: The movie is produced by Walt Disney Pictures.

User: What's the type of this movie?

Assistant: The movie is a Disney animated film.

3-hour Video
Avatar

User: Summarize the plot of this movie.

Assistant: Avatar follows Jake Sully, a disabled former Marine who is recruited by the RDA to work on Pandora, a moon rich in a valuable mineral called unobtanium. Jake's twin brother, Tommy, died in a helmet-cam accident, and Jake is offered his place in the Avatar Program, which allows humans to link with the native Na'vi, a humanoid species. Jake is assigned to the avatar of his deceased brother......

User: How human work on Pandora?

Assistant: Humans work on Pandora by using avatars, which are grown from human DNA mixed with the DNA of the native Na'vi, to interact with the environment and the Na'vi.

User: Why Jake Sully decide to fight against humanity with the Na'vi?

Assistant: Jake Sully decides to fight against humanity with the Na'vi after he falls in love with the Na'vi and their way of life, and after he realizes that the human military is destroying their home and culture.

Examples of LLaMA-VID with single image, short video, and hour-long video.

Results & Analysis

Input with 1.5-min video GTA6 Trailer

User: Describe this video in detail.

Assistant: This video begins with a woman sitting at a table, typing on a computer. She then gets up and walks away. The next scene shows people walking around a beach, enjoying the sun and the sand. The camera then cuts to a city where people are seen dancing in the streets. The woman from the beach scene is shown walking down the street, and she is followed by a man. The two then enter a building and start dancing together. The video then transitions to a car chase where people are seen riding around in cars and performing stunts. Finally, the video ends with a game of Grand Theft Auto being played. Overall, this video seems to be a collection of different scenes from various settings. The beach scene shows people enjoying the sun and the sand, while the city scene shows people dancing and having fun. The car chase scene seems to be a stunt sequence, and the Grand Theft Auto game is the final scene. The video seems to be a compilation of different moments from various settings, showcasing people enjoying themselves and having a good time.

User: Which game do you think this video prompt?

Assistant: The video prompts the game Grand Theft Auto.

User: What will attract you most in this game?

Assistant: The game's graphics and the diverse range of activities it offers are likely to attract a wide range of players. The game's graphics are likely to be visually stunning, with a range of different environments and settings to explore. Additionally, the game offers a range of different activities, from racing and stunt driving to shooting and exploration, which are likely to attract players who enjoy a mix of different gameplay styles.

User: If I told you this game is set in Florida, which city do you prefer? and why?

Assistant: If this game is set in Florida, I would prefer Miami. The video shows a woman sitting at a table and a man walking away, which suggests that the game may be set in a city with a vibrant nightlife. Additionally, the video shows a beach scene and a city street with people dancing, which further supports the idea that Miami would be a great setting for this game.

User: Who produce this game?

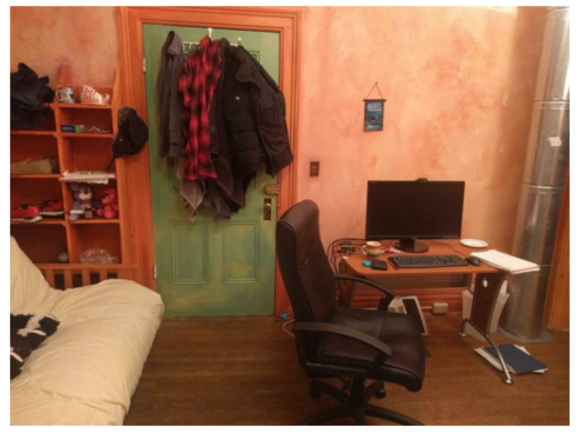
Assistant: The game is produced by Rockstar Games.

Short Video

User: Why this video looks funny? Explain in detail.

Assistant: The video seems to be a comedy skit or a parody of a serious situation. The man is seen speaking to the camera while making exaggerated expressions. He appears to be acting funny and making the audience laugh. As the scene progresses, he starts to laugh louder and harder, further emphasizing the humor of the situation. The other people in the scene are also laughing, which indicates that they are enjoying the humor as well. Overall, the video seems to be a light-hearted and entertaining comedy skit that aims to make the audience laugh.

Single Image



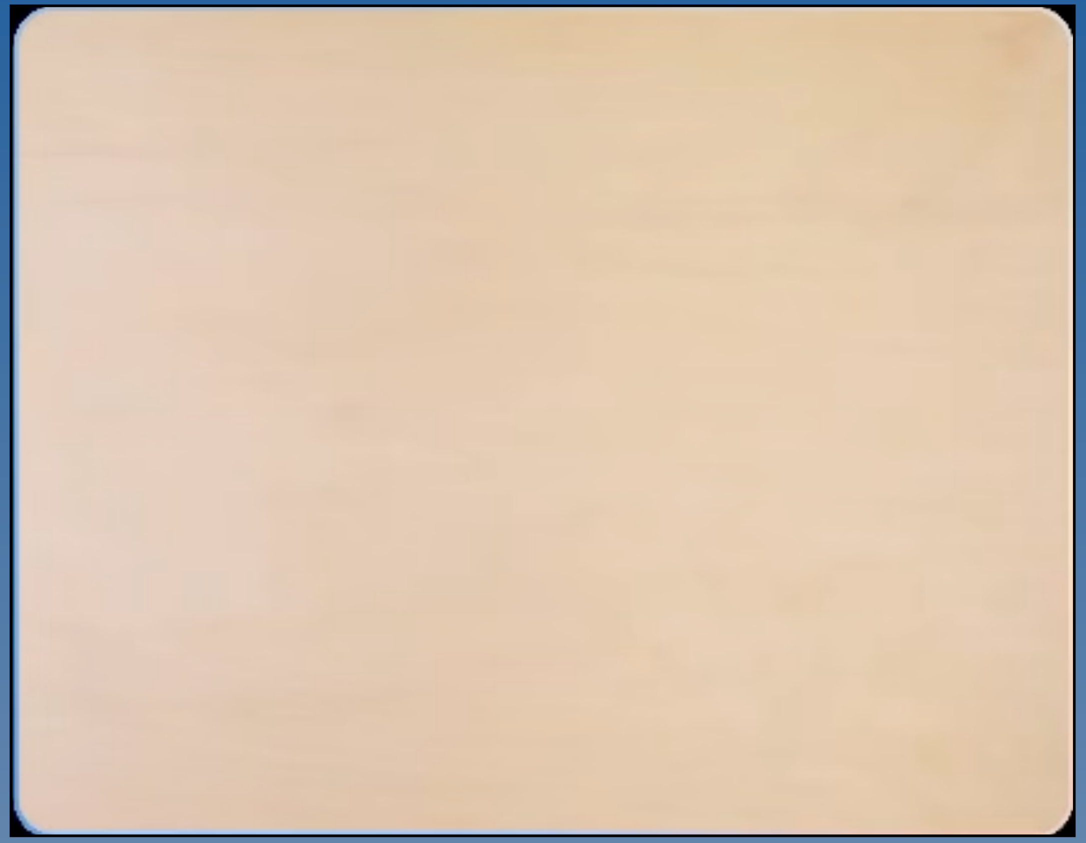
User: Suppose you are a detective, what can you infer from the visual clues in the image?

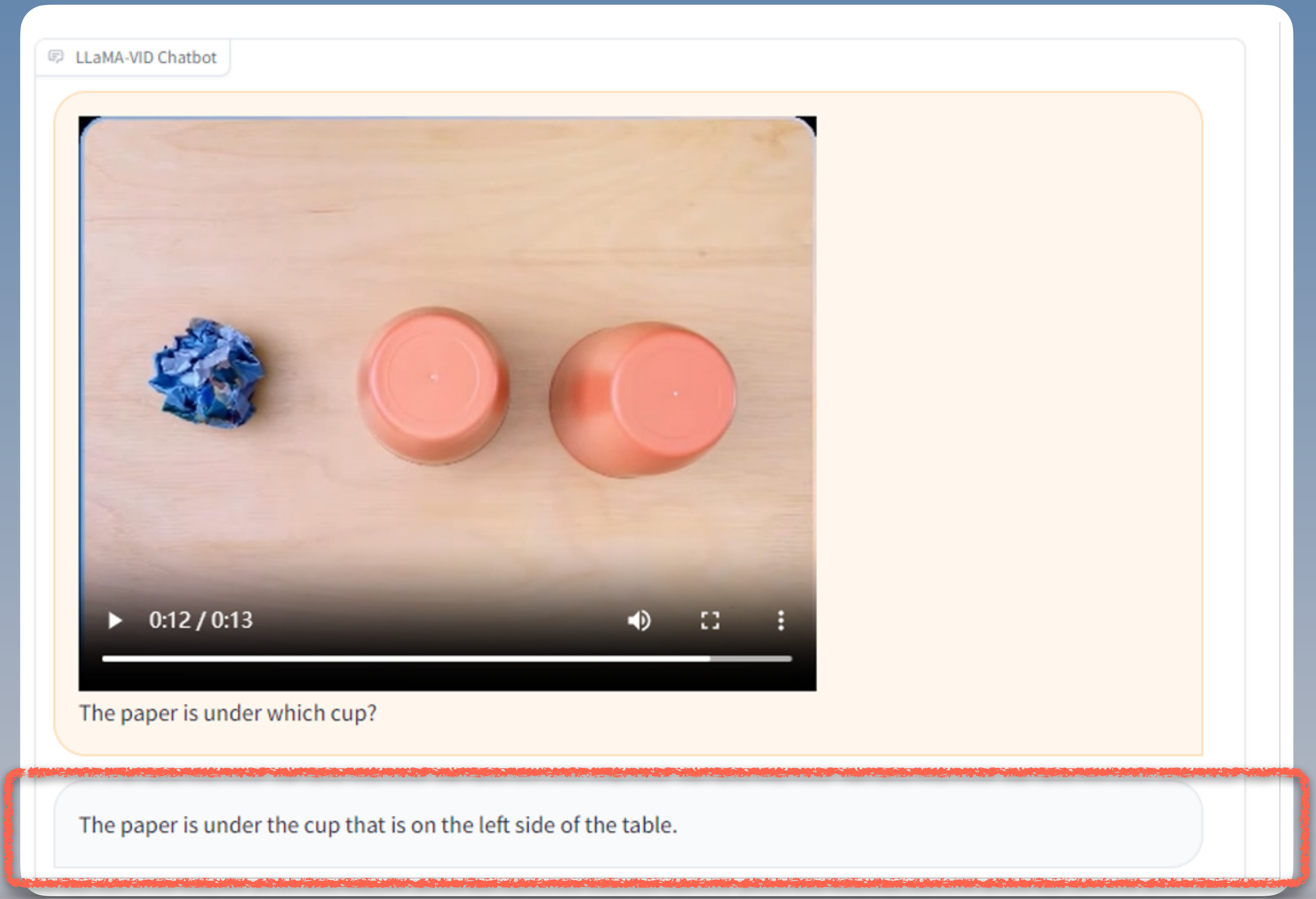
Assistant: As a detective, I can infer that the person living in this space has a preference for a cluttered environment. The room features a bed, a desk with a computer, a chair, and a door with coats hanging on it. The presence of multiple coats hanging on the door suggests that the person might have a busy lifestyle or frequently goes out. The cluttered nature of the room might indicate that the person has limited time to organize or tidy up the space. Additionally, the presence of a computer and a keyboard on the desk suggests that the person might be involved in work, study, or other computer-related activities.

Future Plan

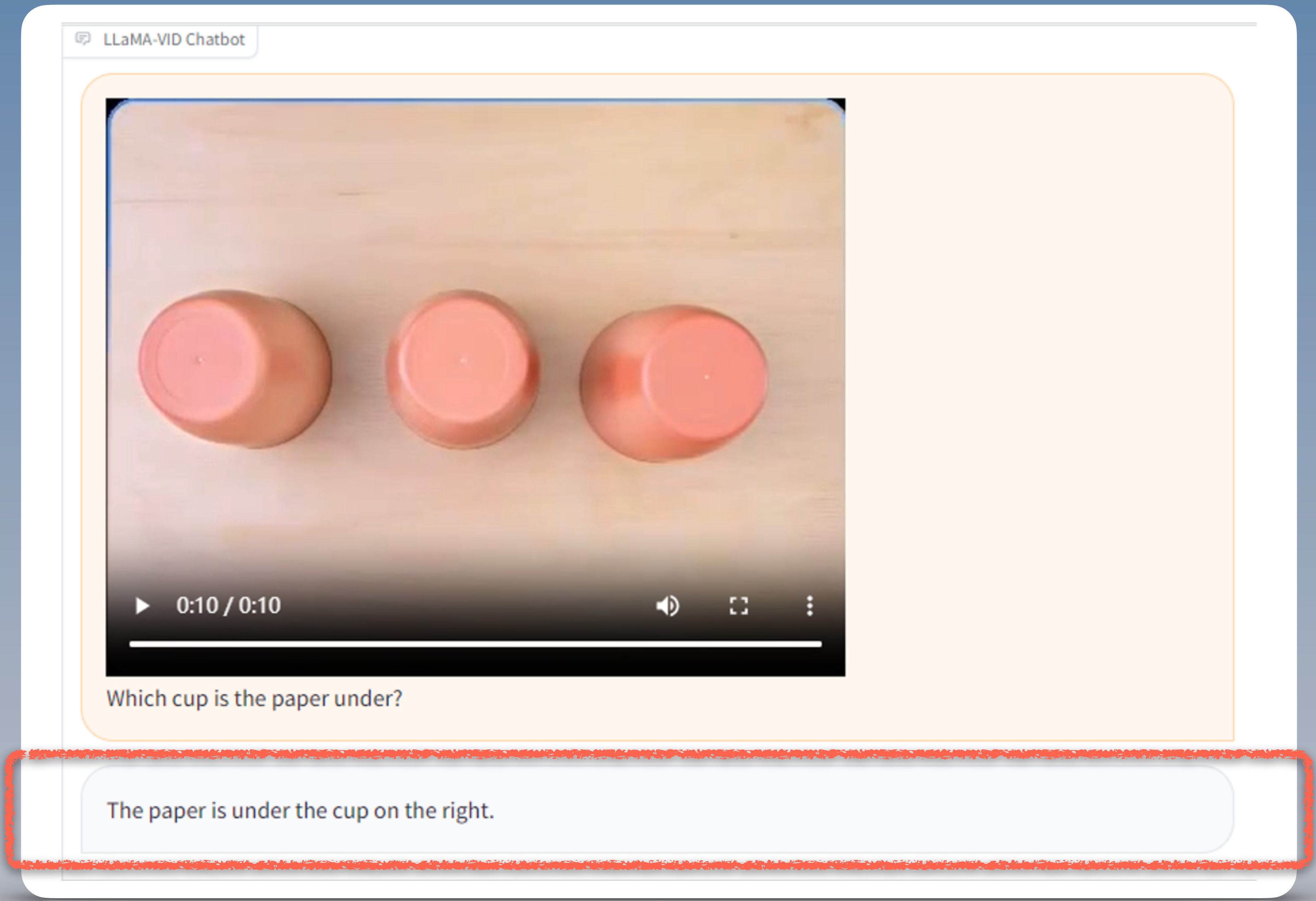
What's the Next?

- *Motion:* how to model motion across frames?
- *Detail:* how to preserve frame details in some key frames?





Inference with the last frame.



Inference without the last frame.

Reference

[1] Junnan Li, et.al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, In ICML, 2022.
[2] Jean-Baptiste Alayrac, et.al., Flamingo: a Visual Language Model for Few-Shot Learning, In NeurIPS, 2022.
[3] Junnan Li, et.al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, arXiv, 2023.
[4] Deyao Zhu, et.al., MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, arXiv, 2023.
[5] Haotian Liu, et.al.,Visual Instruction Tuning, In NeurIPS, 2023.
[6] Wenliang Dai, et.al., InstructBLIP:Towards General-purpose Vision-Language Models with Instruction Tuning, arXiv, 2023.
[7] Jinze Bai, et.al., Qwen-VL:A Versatile Vision-Language Model for Understanding, Localization,Text Reading, and Beyond, arXiv, 2023.
[8] OpenAI, GPT-4V(ision) System Card, https://openai.com/research/gpt-4v-system-card, 2023
[9] Haotian Liu, et.al., Improved Baselines with Visual Instruction Tuning, arXiv, 2023.
[10] Google, Gemini:A Family of Highly Capable Multimodal Models, https://deepmind.google/technologies/gemini/, 2023.
[11] KunChang Li, et.al.,VideoChat: Chat-Centric Video Understanding, arXiv, 2023.
[12] Hang Zhang, et.al.,Video-LLaMA:An Instruction-tuned Audio-Visual Language Model for Video Understanding, arXiv, 2023.
[13] Enxin Song, et.al., MovieChat: From Dense Token to Sparse Memory for Long Video Understanding, arXiv, 2023.
[14] Muhammad Maaz, et.al.,Video-ChatGPT:Towards Detailed Video Understanding via Large Vision and Language Models, arXiv, 2023.
[15] Yanwei Li, et.al., LLaMA-VID:An Image is Worth 2 Tokens in Large Language Models, arXiv, 2023.
[16] Hugo Touvron, et.al., Llama 2: Open foundation and fine-tuned chat models, arXiv, 2023.
[17] Yukang Chen, et.al., Longlora: Efficient fine-tuning of long-context large language models, arXiv, 2023.

Thanks!

Q&A

[Project] [Code] [Demo] [Paper] [Model]