Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding

CVPR 2024

提纲

- 作者介绍
- 研究背景
- 研究方法
- 实验效果
- 总结&思考

作者介绍

Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding

Peng Jin[1,2,3] Ryuichi Takanobu Wancai Zhang[4] Xiaochun Cao[5] Li Yuan[1,2,3*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China

[3]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, China

[4]Nari Technology Co.,Ltd., China [5]School of Cyber Science and Tech., Shenzhen Campus of Sun Yat-sen University, Shenzhen, China

jp21@stu.pku.edu.cn yuanli-ece@pku.edu.cn

https://github.com/PKU-YuanGroup/Chat-UniVi

Video-LLaVA: Learning United Visual Representation by Alignment Before Projection

Bin Lin[1] Yang Ye[3] Bin Zhu[1,4] Jiaxi Cui[1,6] Munang Ning[1,2] Peng Jin[1,5] Li Yuan[1,2,5]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China

[3]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China [4]Tencent Data Platform, Shenzhen, China

[5]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, China [6]FarReel Ai Lab

https://github.com/PKU-YuanGroup/Video-LLaVA

LanguageBind, MoE-llava

作者介绍

Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding

Peng Jin[1,2,3] Ryuichi Takanobu Wancai Zhang[4] Xiaochun Cao[5] Li Yuan[1,2,3*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China

[3]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, China

[4]Nari Technology Co.,Ltd., China [5]School of Cyber Science and Tech., Shenzhen Campus of Sun Yat-sen University, Shenzhen, China

jp21@stu.pku.edu.cn yuanli-ece@pku.edu.cn

https://github.com/PKU-YuanGroup/Chat-UniVi

金鹏

Phd year3

Text-Video Retrieval, Cross-Modal Representation Learning; Multimodal Large Language Models

袁粒

北大深圳研究生院助理教授、国家优青（海外）

Computer Vision; Multi-Modal Machine Learning; AI for Science

提纲

- 作者介绍
- 研究背景
- 研究方法
- 实验效果
- 总结&思考

研究背景

- Large-scale LLM
  - LLM作为调度器
    - HuggingGPT, Visual ChatGPT, ViperGPT,MM-REACT

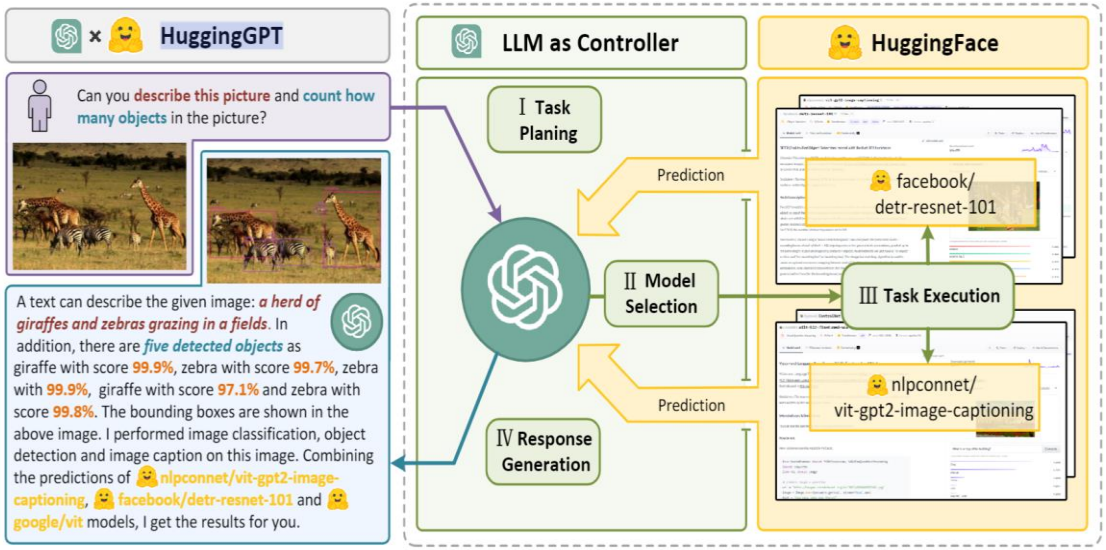


Figure 1: *Language serves as an interface for LLMs (e.g., ChatGPT) to connect numerous AI models (e.g., those in Hugging Face) for solving complicated AI tasks.* In this concept, an LLM acts as a controller, managing and organizing the cooperation of expert models. The LLM first plans a list of tasks based on the user request and then assigns expert models to each task. After the experts execute the tasks, the LLM collects the results and responds to the user.

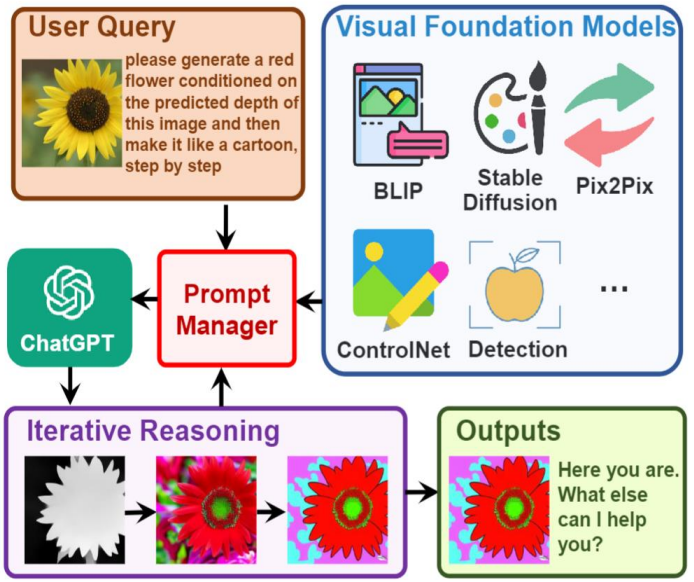
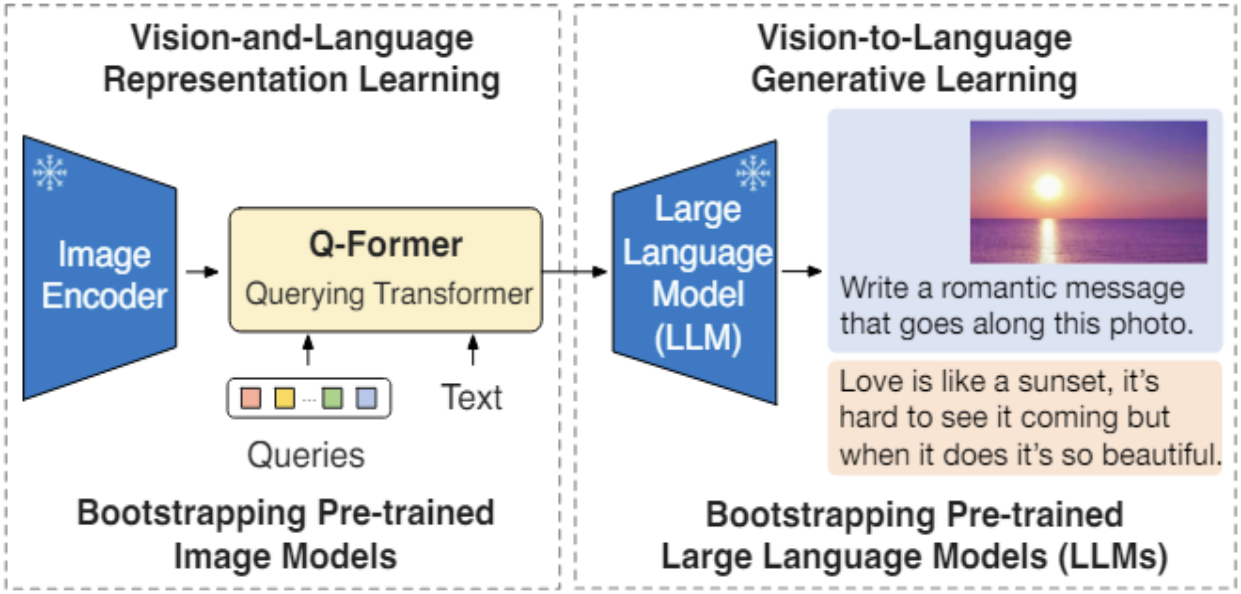


Figure 1. Architecture of Visual ChatGPT.

研究背景

- Large-scale LLM
  - LLM作为调度器
    - HuggingGPT, Visual ChatGPT, ViperGPT,MM-REACT
  - 端到端训练的LLM
    - BLIP2, GPT4,…
    - 存在问题：大多只能专对图像或者视频输入
      - 图像输入：visualtoken获得更精细的空间理解



研究背景

- Large-scale LLM
  - LLM作为调度器
    - HuggingGPT, Visual ChatGPT, ViperGPT,MM-REACT
  - 端到端训练的LLM
    - BLIP2, GPT4,…
    - 存在问题：大多只能专对图像或者视频输入
      - 图像输入：更多visualtoken获得更精细的空间理解
      - 视频输入：更多建模时间信息损失空间信息

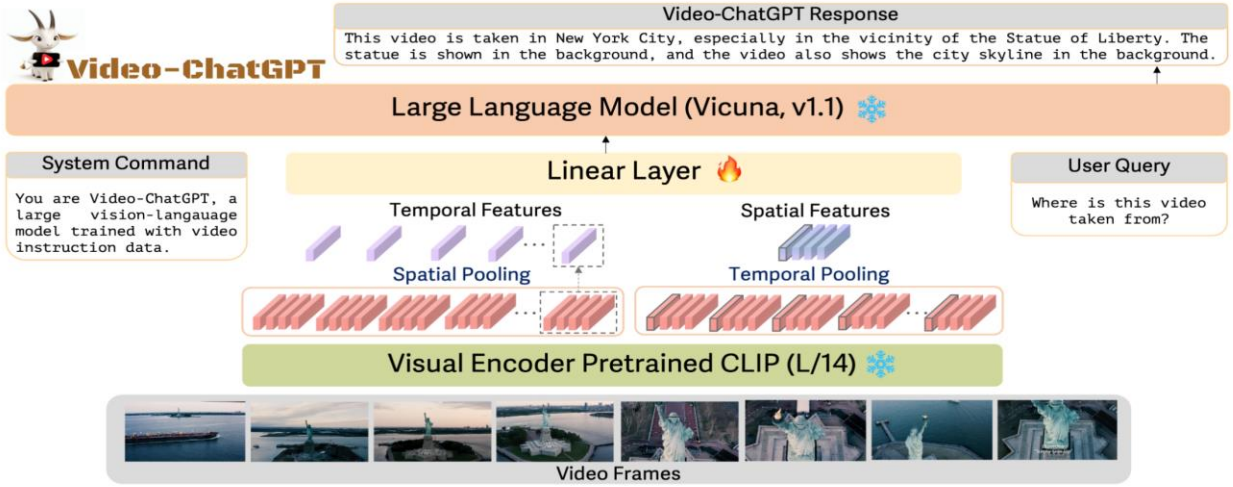


Figure 1: **Architecture of Video-ChatGPT.** Video-ChatGPT leverages the CLIP-L/14 visual encoder

研究背景

- Large-scale LLM
  - LLM作为调度器
    - HuggingGPT，Visual ChatGPT，ViperGPT,MM-REACT
  - 端到端训练的LLM
    - BLIP2，GPT4,…
    - 存在问题：大多只能专对图像或者视频输入
      - 图像输入：使用更多visualtoken获得更精细的空间理解
      - 视频输入：更多建模时间信息损失空间信息
      - 混合输入：Flamingo主要重点是图像理解，缺乏有效建

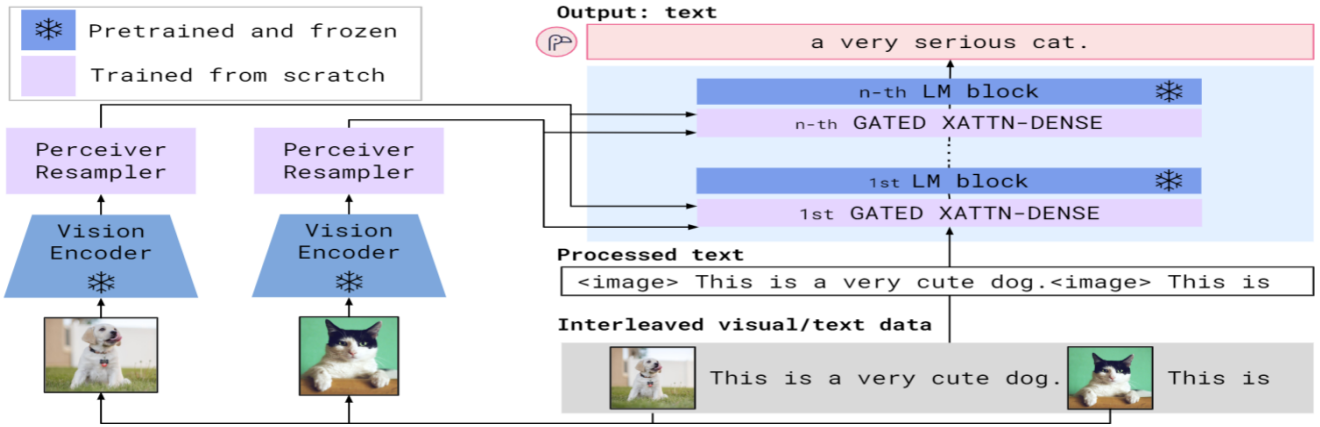
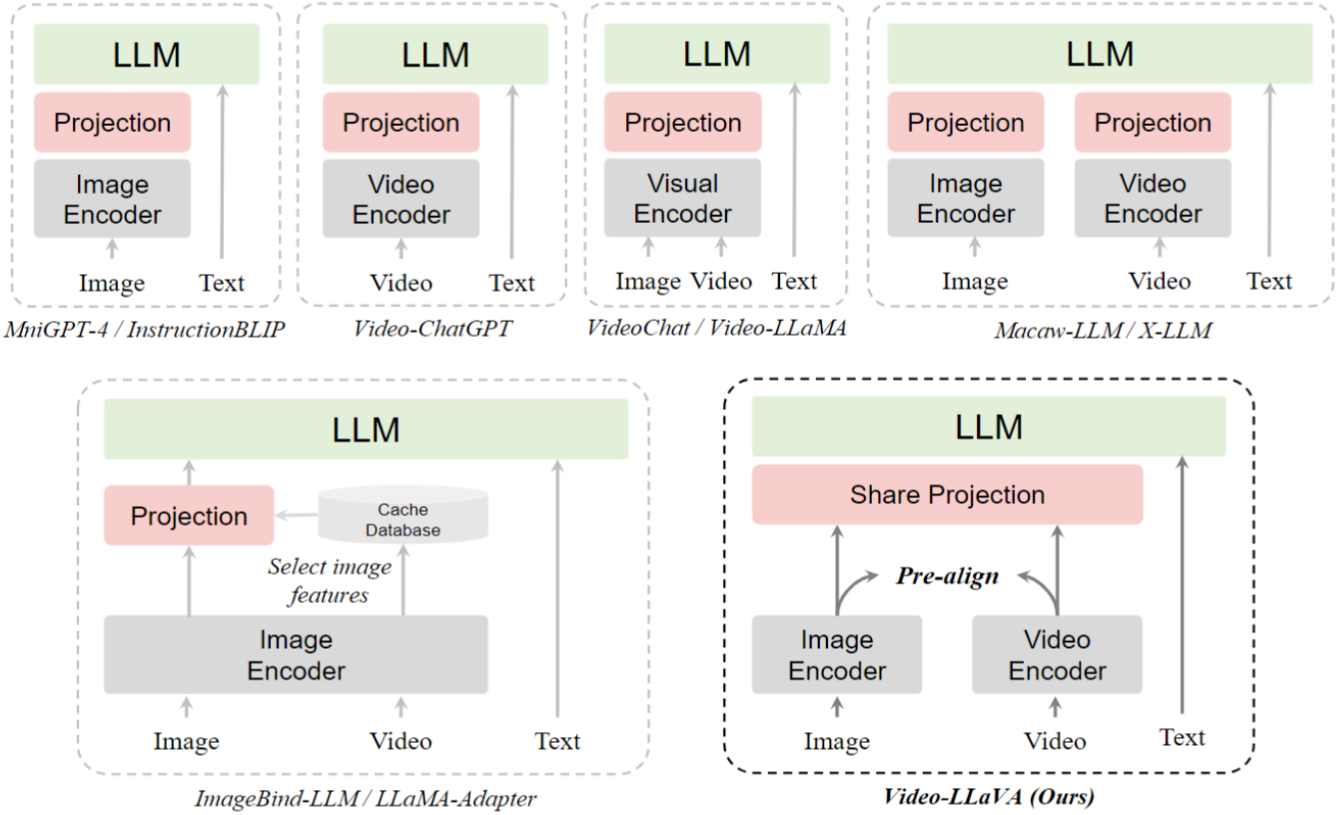


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

研究背景

- Large-scale LLM
  - LLM作为调度器
    - HuggingGPT, Visual ChatGPT, ViperGPT,MM-REACT
  - 端到端训练的LLM



理解

建

研究背景

- 动态token聚合
  - 只在图像上
    - EVIT,Dynamic ViT, ToME,…
  - Parameter-cost

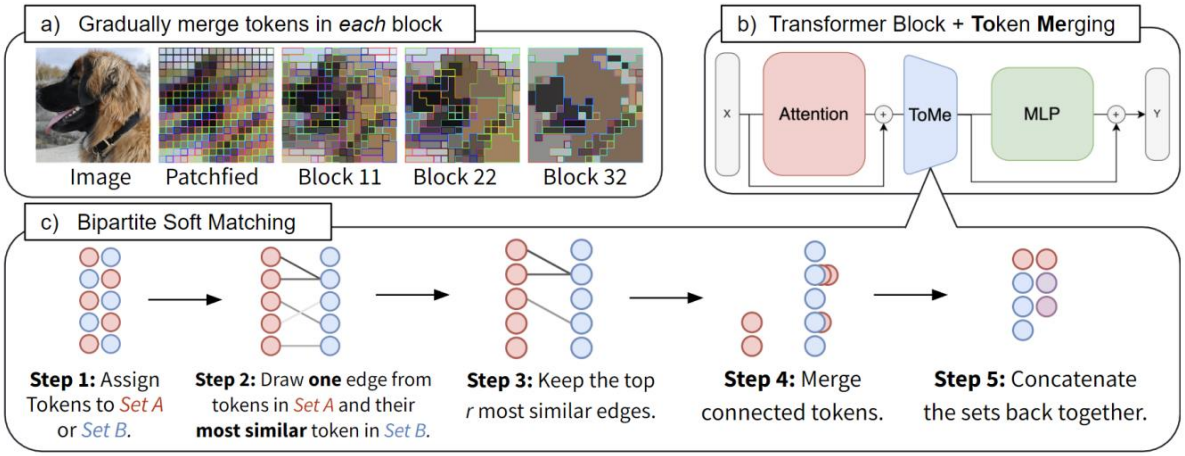


Figure 1: **Token Merging.** (a) With ToMe, similar patches are merged in each transformer block: for example, the dog's fur is merged into a single token. (b) ToMe is simple and can be inserted inside the standard transformer block. (c) Our fast merging algorithm, see Appendix D for implementation.

提纲

- 作者介绍
- 研究背景
- 研究方法
- 实验效果
- 总结&思考

研究方法

核心思想： 信息重要程度不同 → 动态token聚合 （ DPC-KNN ）

- **空间上：** 重要物体，如前景保留，融合背景冗余信息
- 时序上：将视频分成几个关键事件，同一事件中的帧信息融合。

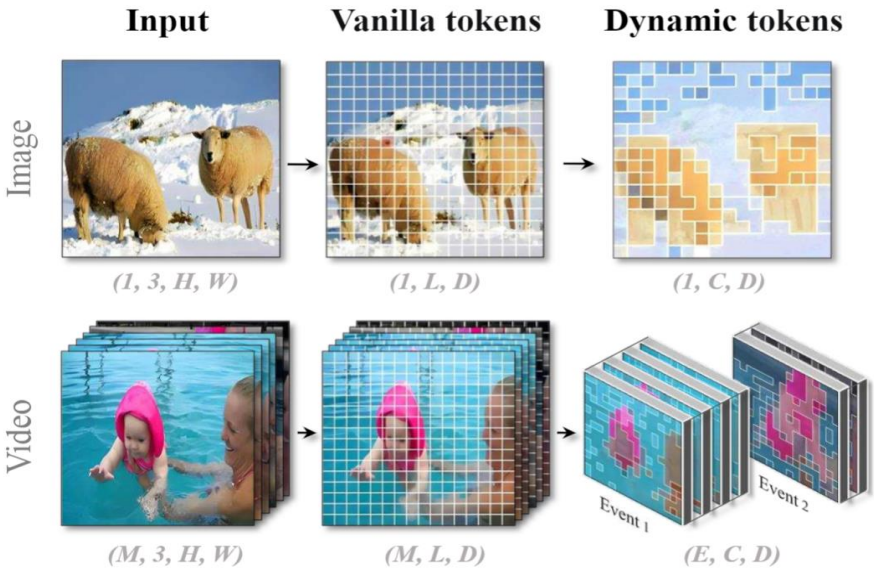


Figure 1. **The unified representation framework for images and videos utilizing a collection of dynamic visual tokens.** "H" and "W" represent the height and width of the input, respectively. "L", "D", "M", "C", and "E" denote the number of vanilla visual tokens, the feature dimension, the frame length, the number of dynamic visual tokens, and the number of events, respectively.

local density:

$$\rho_i = \exp\Big(-\frac{1}{K}\sum_{z_k \in \mathrm{KNN}(z_i, \boldsymbol{Z})} \|z_k - z_i\|^2\Big),$$

relative density：

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} \|z_j - z_i\|^2, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i. \\ \max\limits_{j} \|z_j - z_i\|^2, & \text{otherwise.} \end{cases}$$

cluster center: $\rho_i \times \delta_i$

相同cluster token取平均

研究方法

核心思想： 信息重要程度不同 → 动态token聚合 （ DPC-KNN ）

- 空间上：重要物体，如前景保留，融合背景冗余信息
- **时序上：**将视频分成几个关键事件，同一事件中的帧信息融合。

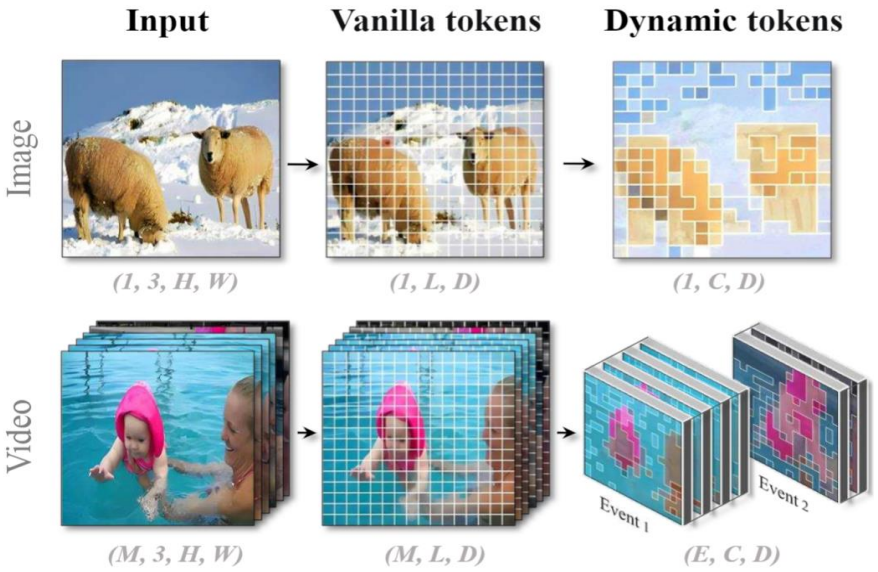


Figure 1. **The unified representation framework for images and videos utilizing a collection of dynamic visual tokens.** "H" and "W" represent the height and width of the input, respectively. "L", "D", "M", "C", and "E" denote the number of vanilla visual tokens, the feature dimension, the frame length, the number of dynamic visual tokens, and the number of events, respectively.

$$\boldsymbol{Z}^m = \{z_i^m\}_{i=1}^L \quad \rightarrow \quad f^m$$

$$\tilde{\boldsymbol{Z}}_n = \{z_i^m | m \in \boldsymbol{F}_n,\ i \in \{1, 2, ..., L\}\}.$$

local density:

$$\tilde{\rho}_i = \exp\Big(-\frac{1}{K} \sum_{z_k \in \mathrm{KNN}(z_i, \tilde{\boldsymbol{Z}})} \|z_k - z_i\|^2\Big),$$

relative density：

$$\tilde{\delta}_i = \begin{cases} \min\limits_{j:\tilde{\rho}_j > \tilde{\rho}_i} \|z_j - z_i\|^2, & \text{if } \exists j \text{ s.t. } \tilde{\rho}_j > \tilde{\rho}_i. \\ \max\limits_{j} \|z_j - z_i\|^2, & \text{otherwise.} \end{cases}$$

相同cluster token取平均

研究方法

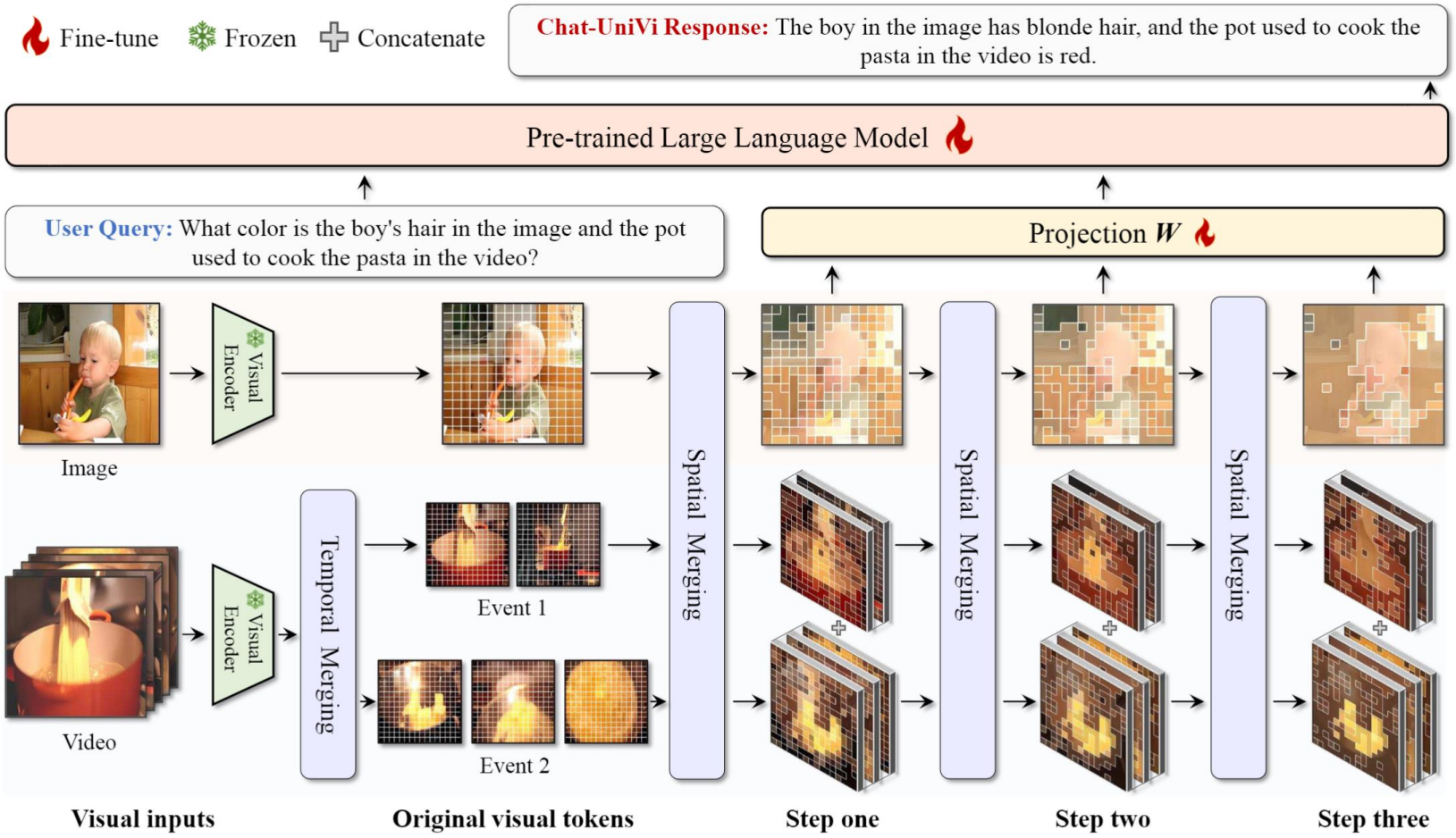


Figure 3. **The overview of the proposed Chat-UniVi for conversations containing both images and videos.** Chat-UniVi uniformly represents images and videos using a collection of dynamic visual tokens and provides a multi-scale representation that equips large language models to perceive both high-level semantic concepts and low-level visual details.

研究方法

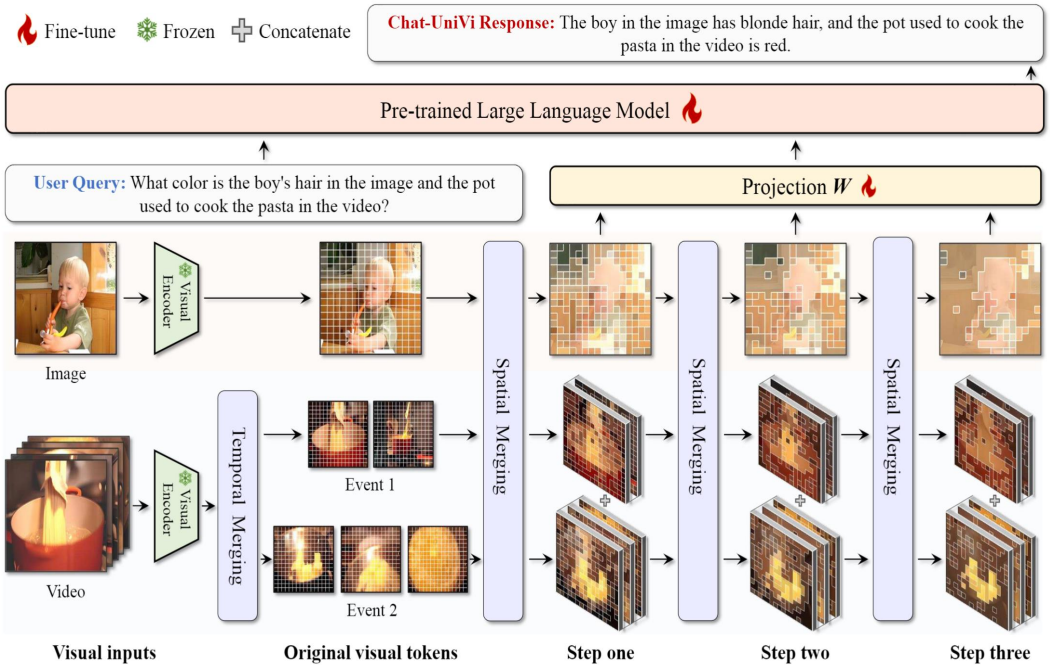


Figure 3. **The overview of the proposed Chat-UniVi for conversations containing both images and videos.** Chat-UniVi uniformly represents images and videos using a collection of dynamic visual tokens and provides a multi-scale representation that equips large language models to perceive both high-level semantic concepts and low-level visual details.

两阶段训练

1. Projection W

COCO and CC3M-595K

1. Projection W + LLM（可以lora）

数据集：混合数据集（单/多轮图像/视频对话数据集）

(i) multimodal in-context instruction datasets, such as MIMIC-IT

(ii) visual instruction datasets, such as LLaVA

(iii) video instruction data from VideoChatGPT

Vision encoder: CLIP (ViT-L/14)

LLM: Vicuna-v1.5

提纲

- 作者介绍
- 研究背景
- 研究方法
- 实验效果
- 总结&思考

实验效果

Experiments

Methods	LLM Size	Visual Tokens	Conversation	Detail	Reason	All
LLaVA [40]	13B	256	83.1	75.3	96.5	85.1
LLaVA [40]	7B	256	70.3	56.6	83.3	70.1
LLaVA [40]†	7B	256	78.8	70.2	91.8	80.4
Chat-UniVi	7B	**112**	**84.1**	**74.2**	**93.7**	**84.2**

Table 1. **GPT-based evaluation for image understanding.** "†" denotes our own re-implementation of LLaVA under our training settings (same foundation model, same image data, and same training scheme) for a fair comparison.

Methods	LLM Size	Correct	Detail	Context	Temporal	Consistency
Video-LLaMA [76]	7B	39.2	43.6	43.2	36.4	35.8
LLaMA-Adapter [77]	7B	40.6	46.4	46.0	39.6	43.0
VideoChat [33]	7B	44.6	50.0	50.6	38.8	44.8
Video-ChatGPT [45]	7B	48.0	50.4	52.4	39.6	47.4
Chat-UniVi	7B	**57.8**	**58.2**	**69.2**	**47.9**	**56.2**

Table 2. **GPT-based evaluation for video understanding.** The results reported in Maaz et al. [45] span a range from 0 to 5. To standardize the metrics, we normalize all scores to a scale of 0 to 100.

实验效果

Experiments

Methods	LLM Size	Subject			Context Modality			Grade		Average
		NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Random Choice [42]	-	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Human [42]	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
Zero-shot Question Answering Accuracy (%)										
GPT-4 [40]	1T+	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
GPT-3 [42]	175B	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87	74.04
LLaVA [40]†	7B	47.78	41.96	53.64	47.90	44.03	51.92	49.63	45.29	48.08
Chat-UniVi	7B	**58.61**	**61.08**	**61.82**	**57.33**	**58.25**	**61.39**	**62.04**	**56.23**	**59.96**
Fine-tuning Question Answering Accuracy (%)										
LLaVA [40]	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA [40]†	7B	79.71	91.68	82.82	80.94	83.24	81.46	83.74	81.74	83.02
LLaMA-Adapter [77]	7B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaMA-SciTune [20]	7B	84.50	**94.15**	82.91	88.35	83.64	**88.74**	85.05	85.60	86.11
Chat-UniVi	7B	**88.50**	93.03	**85.91**	**88.51**	**85.97**	88.15	**88.88**	**88.60**	**88.78**

Table 3. **Zero-shot and fine-tuning question answering accuracy on the ScienceQA test set.** Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. "†" denotes our own re-implementation of LLaVA under our training settings for a fair comparison.

实验效果

Experiments

Methods	LLM Size	MSRVTT-QA		MSVD-QA		TGIF-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM [71]	1B	16.8	-	32.2	-	41.0	-	24.7	-
Video-LLaMA [76]	7B	29.6	1.8	51.6	2.5	-	-	12.4	1.1
LLaMA-Adapter [77]	7B	43.8	2.7	54.9	3.1	-	-	34.2	2.7
VideoChat [33]	7B	45.0	2.5	56.3	2.8	34.4	2.3	26.5	2.2
Video-ChatGPT [45]	7B	49.3	2.8	64.9	3.3	51.4	3.0	35.2	2.7
Chat-UniVi	7B	**55.0**	**3.1**	**69.3**	**3.7**	**69.0**	**3.8**	**46.1**	**3.3**

Table 4. **Zero-shot video question answering accuracy.** We follow the evaluation protocol in Maaz et al. [45], *i.e.*, employing GPT-assisted evaluation to assess the capabilities of models. "Score" denotes the confidence score from 0 to 5 assigned by the GPT model.

Methods	LLM Size	Random (POPE-R)			Popular (POPE-P)			Adversarial (POPE-A)		
		Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes
LLaVA [40]	13B	64.12	73.38	83.26	63.90	72.63	81.93	58.91	69.95	86.76
MiniGPT-4 [84]	13B	79.67	80.17	52.53	69.73	73.02	62.20	65.17	70.42	67.77
InstructBLIP [14]	13B	88.57	89.27	56.57	82.77	84.66	62.37	72.10	77.32	73.03
MultiModal-GPT [19]	7B	50.10	66.71	99.90	50.00	66.67	100.00	50.00	66.67	100.00
mPLUG-Owl [73]	7B	53.97	68.39	95.63	50.90	66.94	98.57	50.67	66.82	98.67
LLaVA [40]†	7B	72.16	78.22	76.29	61.37	71.52	85.63	58.67	70.12	88.33
Chat-UniVi w/o multi-scale	7B	73.88	79.30	74.63	56.36	69.01	90.83	55.63	68.67	91.63
Chat-UniVi w/ multi-scale	7B	**85.19**	**86.05**	**54.67**	**69.50**	**74.39**	**69.10**	**64.97**	**71.54**	**73.10**

Table 5. **Zero-shot object hallucination evaluation on the COCO validation set.** We report the results of the polling-based object probing evaluation (POPE). "Yes" represents the proportion of positive answers that the model outputs. "†" denotes our own re-implementation of LLaVA under our training settings (same foundation model, same image data, and same training scheme) for a fair comparison.

多尺度表示提高了抵抗幻觉的能力

实验效果

Experiments

Methods	Image Understanding				Video Understanding				
	Conversation	Detail	Reason	All	Correct	Detail	Context	Temporal	Consistency
Only Image	84.0	69.3	89.3	81.5	43.4	48.6	56.8	36.6	46.2
Only Video	72.7	55.8	71.5	66.8	57.4	**58.8**	69.0	47.0	56.0
Image + Video	45.5	31.3	76.1	50.9	51.2	55.6	64.8	40.3	50.4
Video + Image	79.0	69.2	88.5	79.1	45.6	49.8	58.2	38.8	47.8
Image & Video	**84.1**	**74.2**	**93.7**	**84.2**	**57.8**	58.2	**69.2**	**47.9**	**56.2**

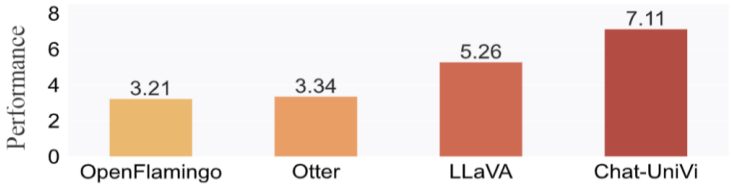
Table 6. **Ablation study about instruction tuning scheme.** "Only Image" indicates training solely on image data. "Image + Video" means training on image data followed by fine-tuning on video data. "Image & Video" denotes training on a mixed dataset.

C_1	C_2	C_3	Visual Tokens	Conversation	Detail	Reason	All
16	8	4	28	78.6	69.0	**95.1**	81.1
32	16	8	56	82.7	67.2	94.5	81.6
64	32	16	112	**84.1**	**74.2**	93.7	**84.2**
128	64	32	224	79.8	68.7	83.8	79.8

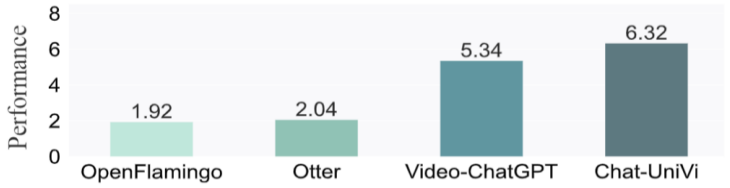
Table 7. **Ablation study about the number of spatial visual clusters.** "C_1", "C_2", and "C_3" denote the number of clusters at the first step, the second step, and the last step, respectively.

Clustering Ratio	Correct	Detail	Context	Temporal	Consistency
$1/M$	51.2	41.8	47.6	28.0	42.2
$1/32$	57.2	58.0	**69.6**	45.8	54.2
$1/16$	**57.8**	**58.2**	69.2	**47.9**	56.2
$1/8$	56.8	**58.2**	68.0	46.2	**57.8**

Table 8. **Ablation study about the number of temporal visual clusters.** "M" is the frame length. "$1/M$" denotes that the model directly consolidates all frames into a single event.



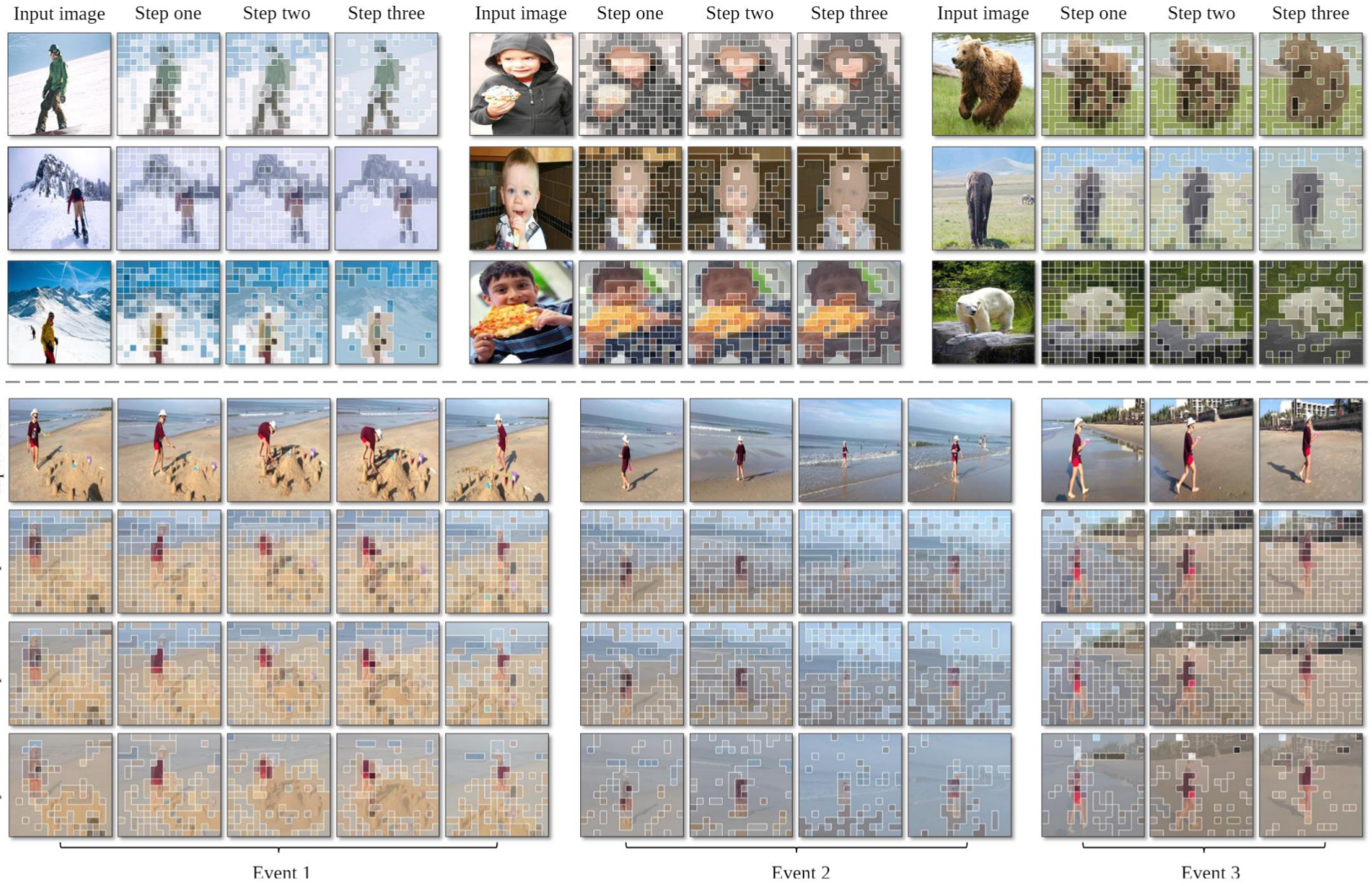
(a) Comparison of image-based conversations



(b) Comparison of video-based conversations

Figure 4. **Human evaluations.** In 30 image conversation scenarios and 30 video conversation scenarios, the evaluators rate the model on a scale of 0 to 10 based on its multimodal conversation performance. Finally, we use the average score as the final model score.

实验效果


Input image
Step one
Step two
Step three
Input image
Step one
Step two
Step three
Input image
Step one
Step two
Step three
Input video
Step one
Step two
Step three
Event 1
Event 2
Event 3

提纲

- 作者介绍
- 研究背景
- 研究方法
- 实验效果
- 
  总结&思考

总结

- 引入token聚合，保留更多视频信息
- 多尺度融合
- 视频+图像数据混合训练，模型拥有同时处理视频和图像的能力

- 用一种parameter-free的融合方法，带参数训练容易崩

- 专注一个setting，融合多个简单的想法，work

Thank you !