VISTA-LLAMA: Reducing Hallucination in Video Language Models via Equal Distance to Visual Tokens

Fan Ma[1], Xiaojie Jin[2*], Heng Wang[2], Yuchen Xian[1], Jiashi Feng[2], Yi Yang[1*]
[1]Zhejiang University [2]ByteDance Inc.

关键词： MLLM，幻觉，位置编码，视频理解

曹耘宁
2024/07/09
智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

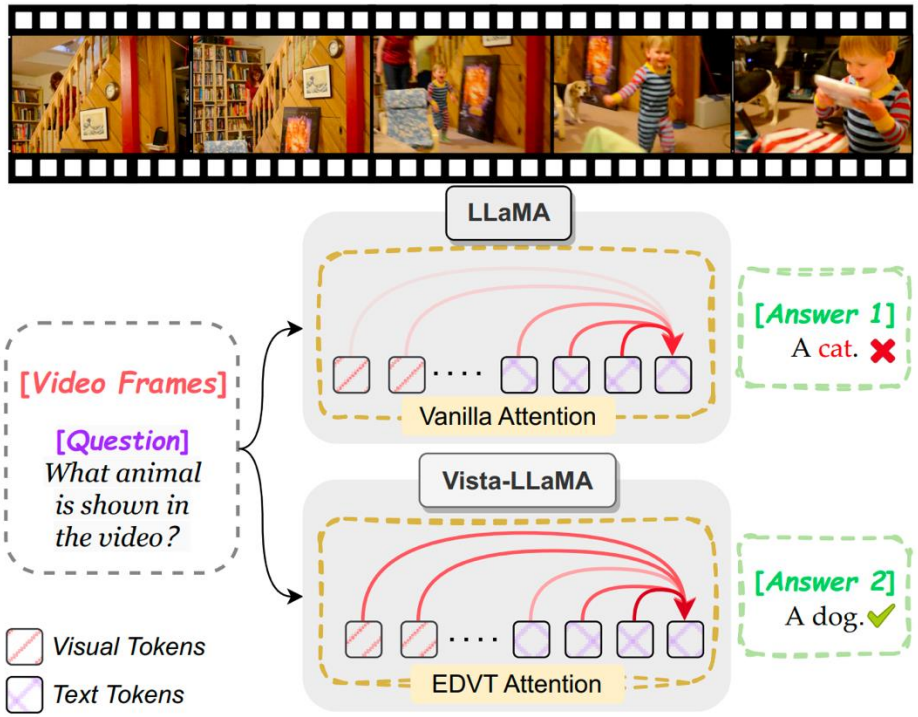
- 研究背景
- 研究方法
- 实验效果
- 总结

MLLM存在幻觉问题

- MLLM存在幻觉问题，答非所问。而且越长的文本越容易出现。

研究动机

- RoPE存在距离衰减问题，text token对visual token关注衰减



- 研究背景
- 研究方法
- 实验效果
- 总结

总体框架

- 1. EDVT attention, 2. Sequential visual projector

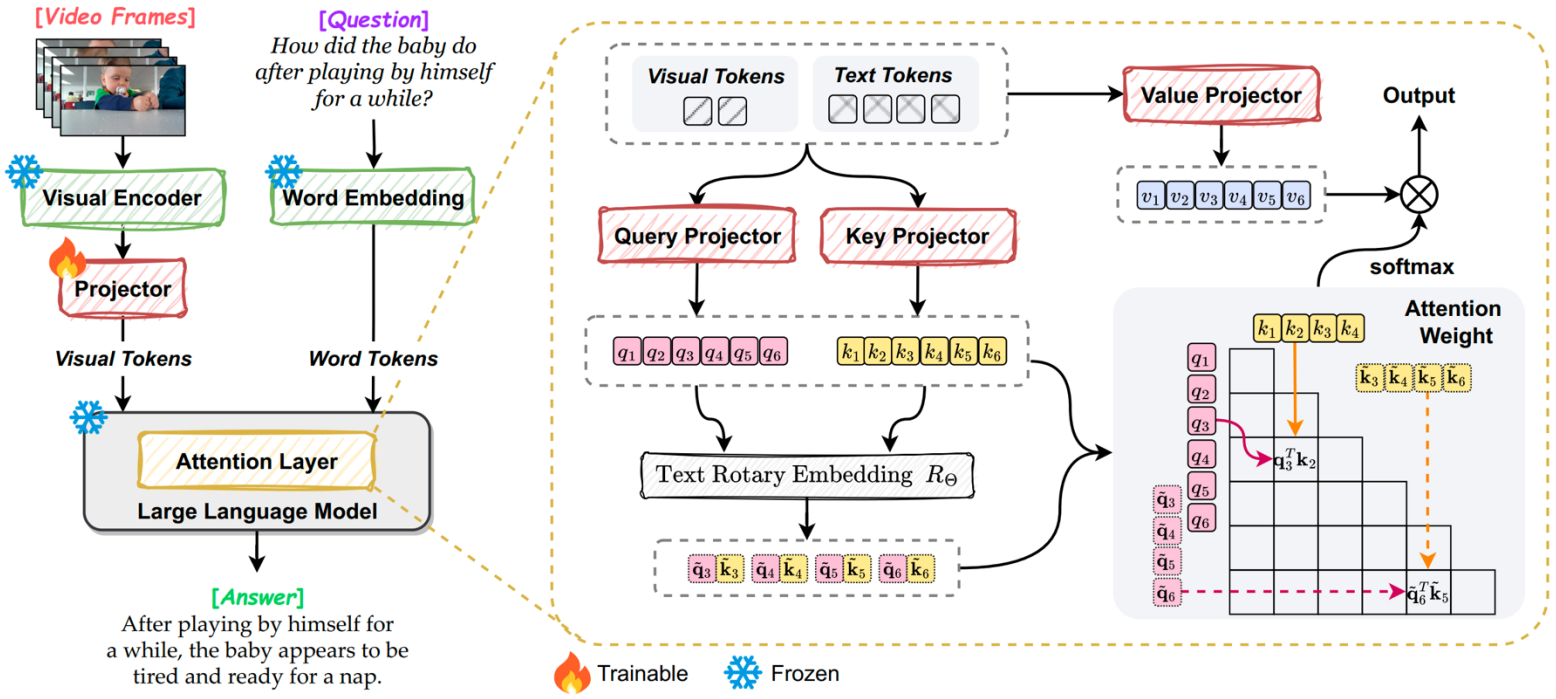


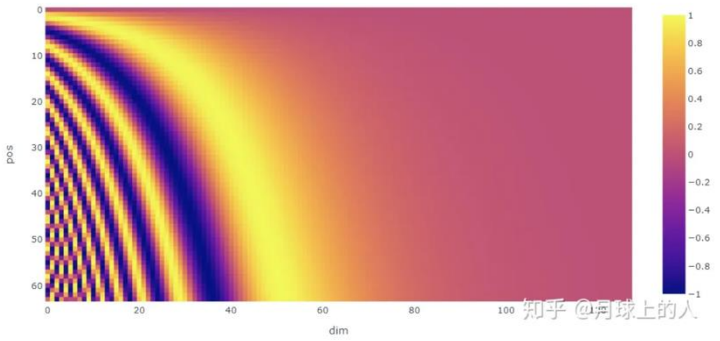
Figure 2. **The framework of VISTA-LLAMA**. The visual encoder and large language model are both frozen (❄) during training, while the projector is trainable (🔥) to map video into the language's space. The attention operation in each layer is present on the right part. Only the text tokens are applied with rotary position embedding to include relative distance information. The attention weights between visual and language tokens are calculated without the rotary position embedding. The casual mask is applied to the bottom-right attention weights.

位置编码

- 绝对位置编码
  - 与q、k相加

$$\boldsymbol{p}_{i,2t} = \sin\left(k/10000^{2t/d}\right)$$

$$\boldsymbol{p}_{i,2t+1} = \cos\left(k/10000^{2t/d}\right)$$



128维位置编码2D示意图

- 旋转位置编码（RoPE）
  - 显式的引入相对位置信息
  - 对q、k乘旋转矩阵

$$f_q(\boldsymbol{x}_m, m) = (\boldsymbol{W}_q\boldsymbol{x}_m)e^{im\theta}$$

$$f_k(\boldsymbol{x}_n, n) = (\boldsymbol{W}_k\boldsymbol{x}_n)e^{in\theta}$$

$$g(\boldsymbol{x}_m, \boldsymbol{x}_n, m-n) = \mathrm{Re}[(\boldsymbol{W}_q\boldsymbol{x}_m)(\boldsymbol{W}_k\boldsymbol{x}_n)^* e^{i(m-n)\theta}]$$

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}$$

RoPE存在长距离衰减

- RoPE得到的attention具有长距离衰减特性
  - 这种特性对于长文本效果好，广泛应用于llama，Chat-GLM等LLM

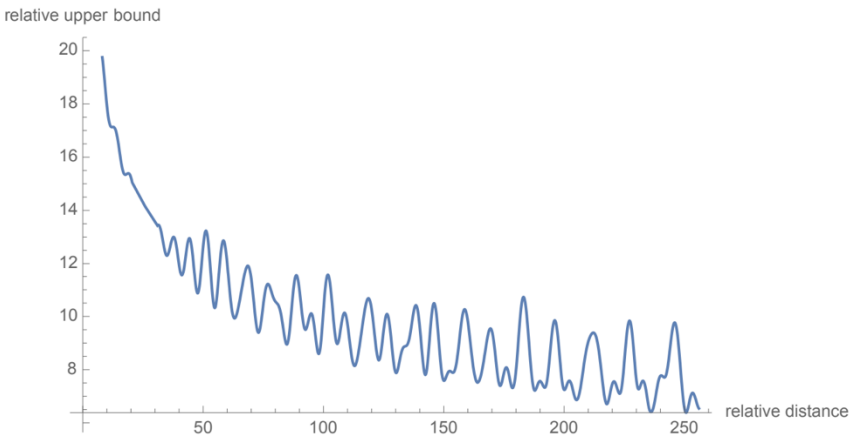
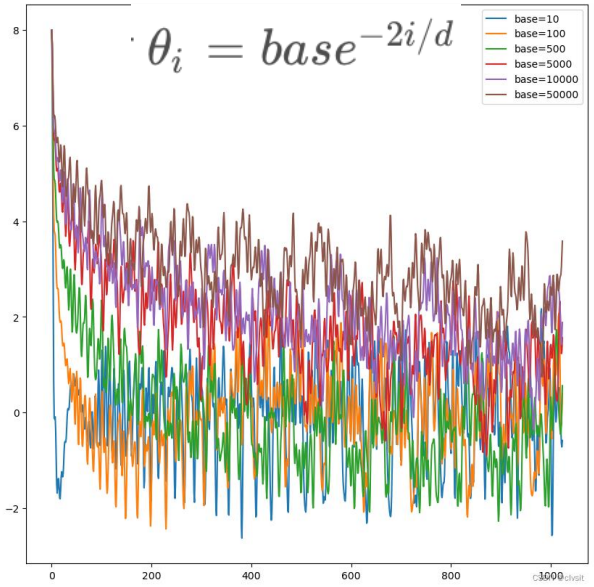


Figure 2: Long-term decay of RoPE.



研究动机

- 无论文本多长，MLLM中的视觉模态注意力权重不应该衰减

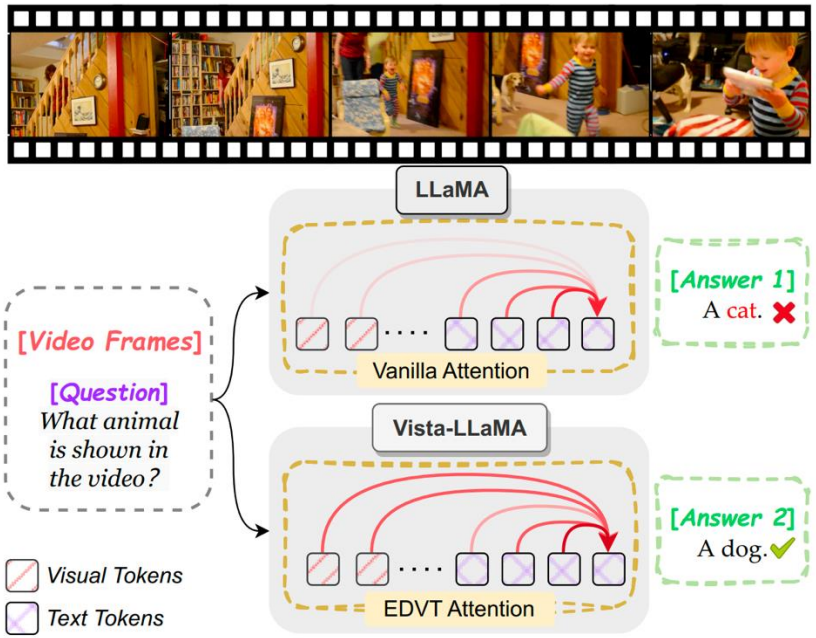
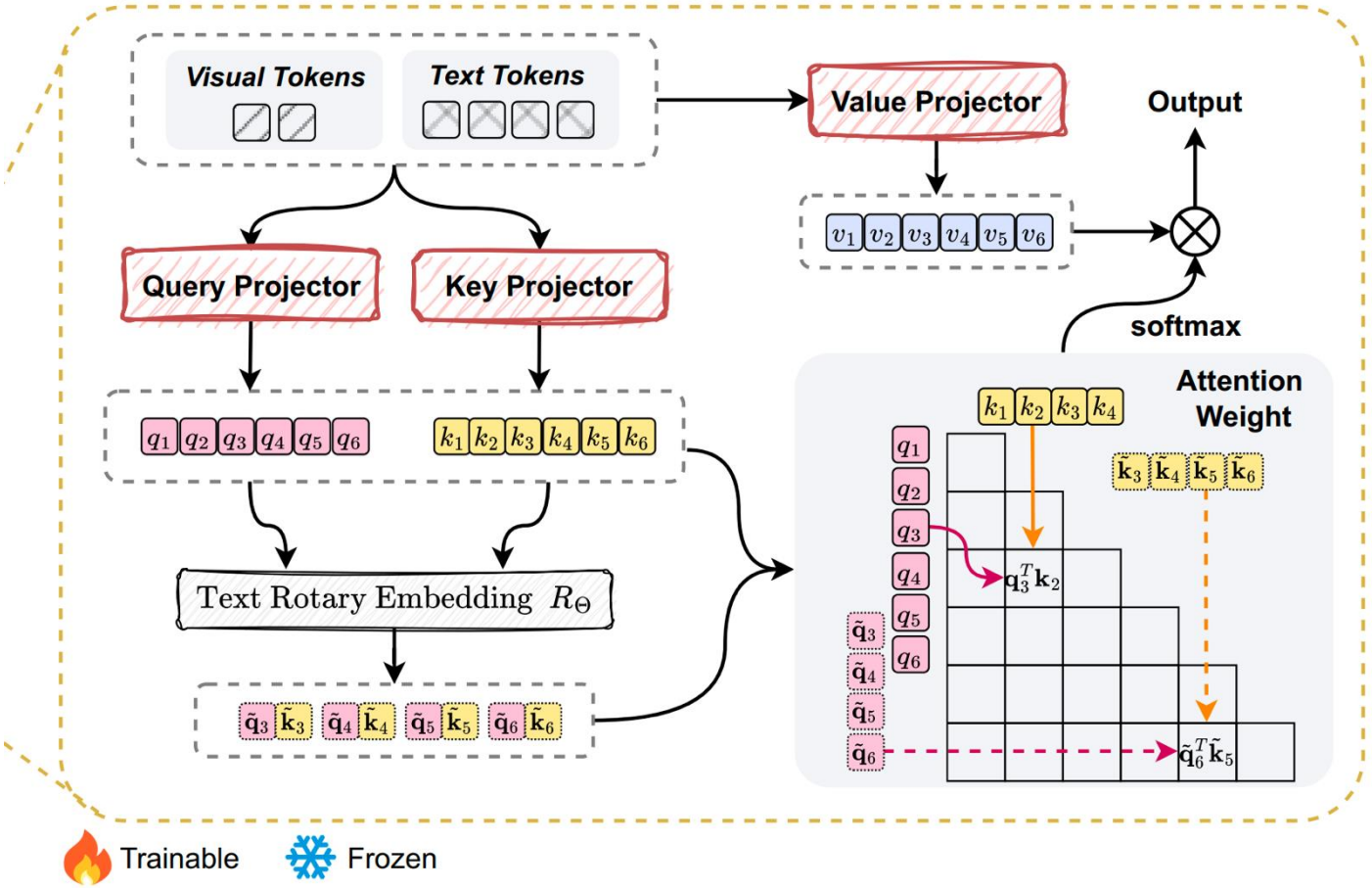


Figure 1. **Video language processing with LLaMA [26] and our VISTA-LLAMA.** The vanilla LLaMA treats visual tokens (▨) the same as other language tokens (▩), weakening the impact for tokens in long distance. Our model retains the same mechanism for language tokens and strengthens the impact of the visual tokens. The intensity of the impact of each token is conveyed through the depth of the line color (→). Our model provides the accurate response for the presented scenario.

Equal Distance to Visual Token

- 仅text token使用RoPE（图中前两列无RoPE，其余有）



Equal Distance to Visual Token

- 传统的attention

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})_j = \frac{\sum_{i=1}^{j} sim(\mathbf{q}_j,\mathbf{k}_i)\mathbf{v}_i}{\sum_{i=1}^{j} sim(\mathbf{q}_j,\mathbf{k}_i)},$$

- RoPE attention

$$\text{Attention}_{rope}(\mathbf{Q},\mathbf{K},\mathbf{V})_j = \frac{\sum_{i=1}^{j} sim(\mathbf{R}_j\mathbf{q}_j,\mathbf{R}_i\mathbf{k}_i)\mathbf{v}_i}{\sum_{i=1}^{j} sim(\mathbf{R}_j\mathbf{q}_j,\mathbf{R}_i\mathbf{k}_i)}.$$

- 组合得到EDVT

$$\text{Attention}_{edvt}(\mathbf{Q},\mathbf{K},\mathbf{V})_j = \frac{\sum\limits_{k_i\in\mathcal{T}} sim(\tilde{\mathbf{q}}_j,\tilde{\mathbf{k}}_i)\mathbf{v}_i + \sum\limits_{k_i\in\mathcal{V}} sim(\mathbf{q}_j,\mathbf{k}_i)\mathbf{v}_i}{\sum\limits_{k_i\in\mathcal{T}} sim(\tilde{\mathbf{q}}_j,\tilde{\mathbf{k}}_i) + \sum\limits_{k_i\in\mathcal{V}} sim(\mathbf{q}_j,\mathbf{k}_i)},$$

Algorithm 1 Pseudocode of EDVT attention in a PyTorch-like style.

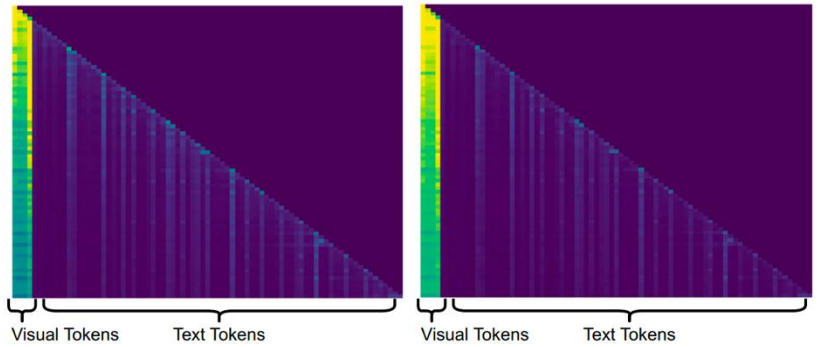
```
# x: hidden input in each attention laye
# v_mask: indcitate which input are from visual tokens

def edvt_attention_layer(x, v_mask):
    # query, answer, value projection
    q, k, v = qkv_proj(x)

    # apply RoPE for query and key inputs
    r_q, r_k = rope(q, k)

    # attention weights without RoPE
    attention = bmm(q.T, k)

    # attention weights with RoPE
    r_attention = bmm(r_q, r_k)

    # Merge attention weights based on visual token
    attention = v_mask * attention + (1 - v_mask) *
        r_attention
    attention = Softmax(attention, dim=-1)

    # Update representation based on attention weights
    v = bmm(attention, v)
    out = linear_proj(v)
    return out
```

`qkv_proj`: linear projection layer; `bmm`: batch matrix multiplication; `rope`: apply rotary position embedding.

消融实验-EDVT-attn

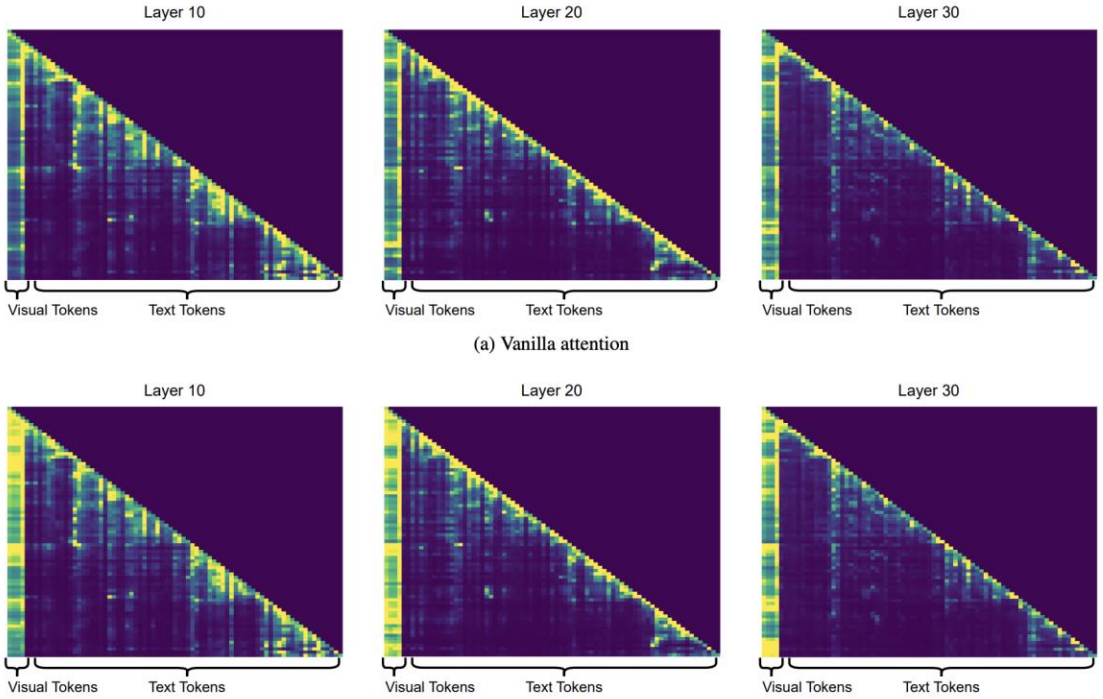
Method	NExT-QA [30]			
	Tem.	Cau.	Des.	Avg.
Video-ChatGPT [18]	37.6	65.1	54.9	54.6
+ **EDVT-Attention**	39.5 (+1.9)	72.8 (+7.7)	54.8	59.3 (+4.7)
VISTA-LLAMA (ViT-L-14)	34.0	69.1	42.2	53.6
+ **EDVT-Attention**	36.8 (+2.8)	72.2 (+3.1)	47.7 (+5.5)	56.5 (+2.9)
VISTA-LLAMA (EVA-CLIP-g)	34.3	65.8	55.9	54.1
+ **EDVT-Attention**	40.7 (+6.4)	72.3 (+6.5)	57.0 (+1.1)	59.7 (+5.6)

Table 3. **Comparison of EDVT-Attention design** with diffent



(a) Vanilla attention (b) EDVT-Attention (Ours)

Figure 5. **Comparison of attention weights** for varing context



(a) Vanilla attention

Sequential Visual Projector

- 在q-former中引入了自回归的序列建模，更好的提取帧间信息

$$\mathbf{x}_t = f_Q(\mathbf{o}_i, \mathbf{x}_{t-1}),$$

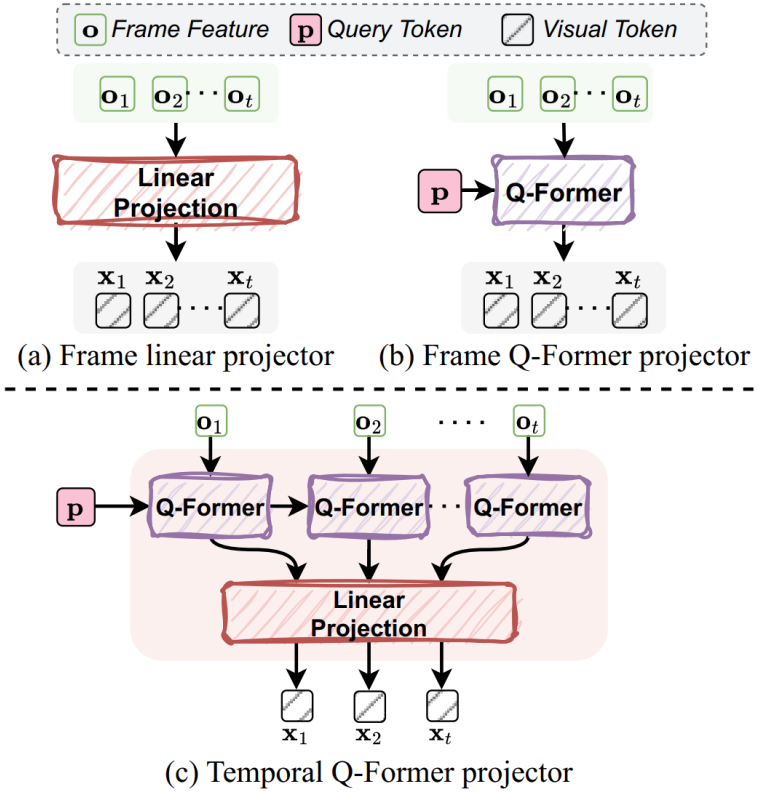
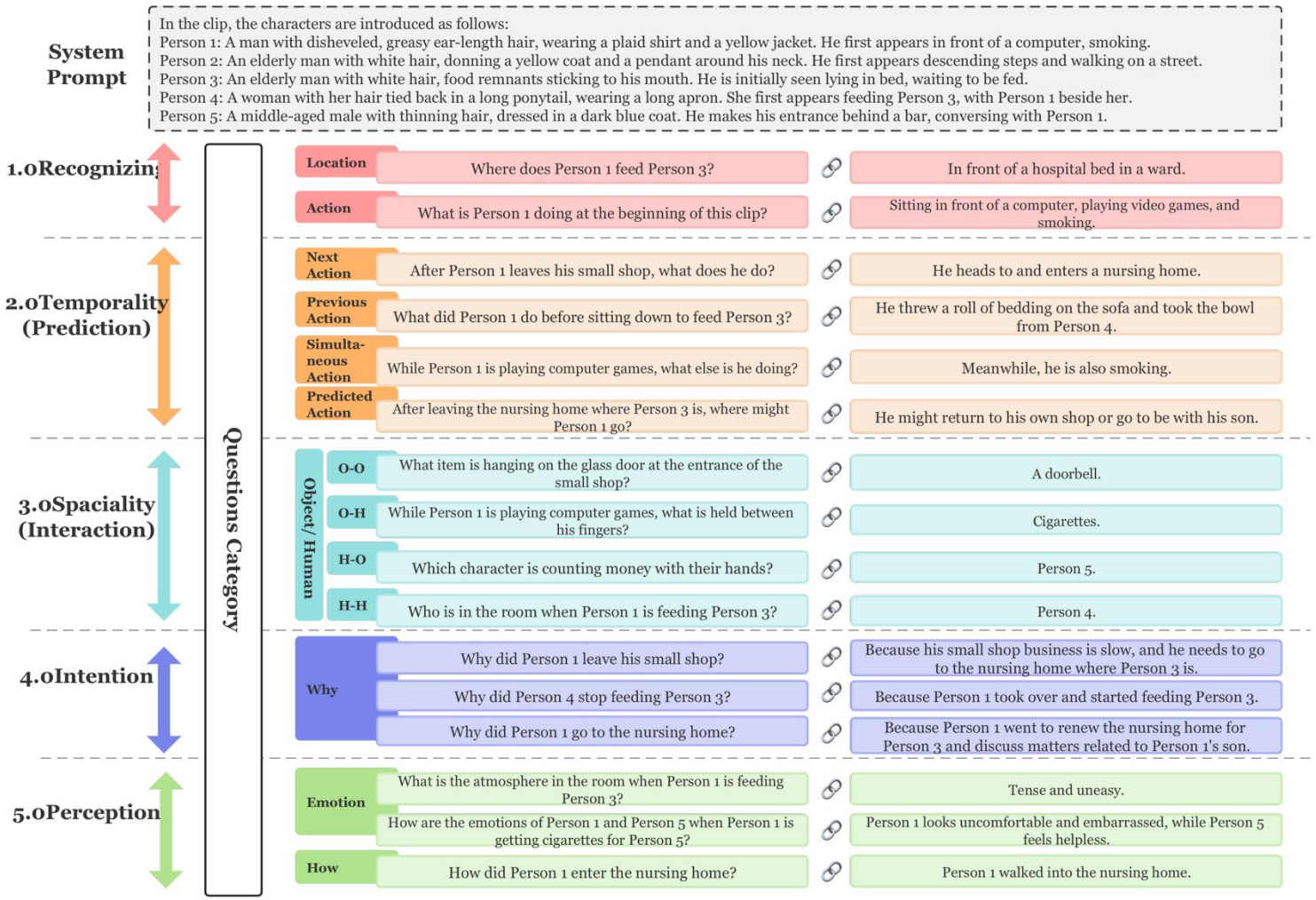


Figure 3. **Comparison of three visual projectors.** (a): Each frame feature is projected into the visual tokens independently with the linear projection. (b): Q-Former uses shared learnable query tokens to separately map each frame into fixed-length tokens. (c): The sequential Q-Former with linear projection layer to enable temporal modeling.

CineClipQA Dataset

- 多



The question consists of two parts: **System Prompt** and **Questions**. The **System Prompt** contains basic information about key characters in the current video clip and provides prompts for the initial actions of characters when necessary. The **Questions** are primarily divided into five categories: *Recognizing*, *Temporality (Prediction)*, *Spaciality (Interaction)*, *Intention*, and *Perception*. Specifically, *Recognizing* includes questions about Location and Action; *Temporality* encompasses questions about the next action, the previous action, simultaneous actions, and predicted actions; *Spaciality* involves questions about spatial information between Object and Human; *Intention* involves three similar types of questions about the purpose of actions; finally, *Perception* examines the recognition of emotions and inquiries about the "how" (approaches, manners ...).

- 研究背景
- 研究方法
- 实验效果
- 后续工作
- 总结

实验性能

Method	NExT-QA [30]		MSVD-QA [32]		MSRVTT-QA [32]		ActivityNet-QA [37]	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM [34]	-	-	32.2	-	16.8	-	24.7	-
Video Chat [13]	56.2	3.2	56.3	2.8	45.0	2.5	26.5	2.2
LLaMA Adapter [40]	-	-	54.9	3.1	43.8	2.7	34.2	2.7
Video LLaMA [39]	-	-	51.6	2.5	29.6	1.8	12.4	1.1
MovieChat [22]	49.9	2.7	61.0	2.9	49.7	2.8	**51.5**	3.1
Video-ChatGPT [18]	54.6	3.2	64.9	3.3	49.3	2.8	35.2	2.7
VISTA-LLAMA (Ours)	**60.7**	**3.4**	**65.3**	**3.6**	**60.5**	**3.3**	48.3	**3.3**

Table 1. **Comparison with SoTA methods** on zero-shot VideoQA. See §4.2 for more details.

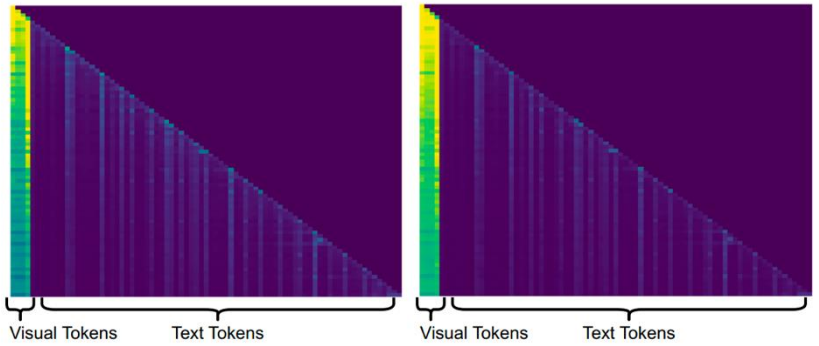
Method	Overall		Description		Temporality		Spaciality		Intention		Perception	
	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy
MovieChat	2.11	20.86	2.41	23.67	1.97	16.32	1.98	16.40	2.41	30.19	1.97	21.80
Video-LLAMA	2.27	23.17	2.31	19.30	2.12	16.35	2.19	21.95	2.47	31.94	2.35	27.70
Video-ChatGPT	2.60	34.11	2.55	26.24	2.60	34.11	2.50	30.62	2.94	46.36	2.43	31.77
VISTA-LLAMA (Ours)	**2.98**	**44.90**	**2.79**	**31.46**	**2.92**	**46.22**	**2.73**	**35.63**	**3.38**	**61.89**	**3.12**	**47.49**

Table 6. **Performance Comparison on CineClipQA** of different methods on various classifications.

消融实验-EDVT-attn

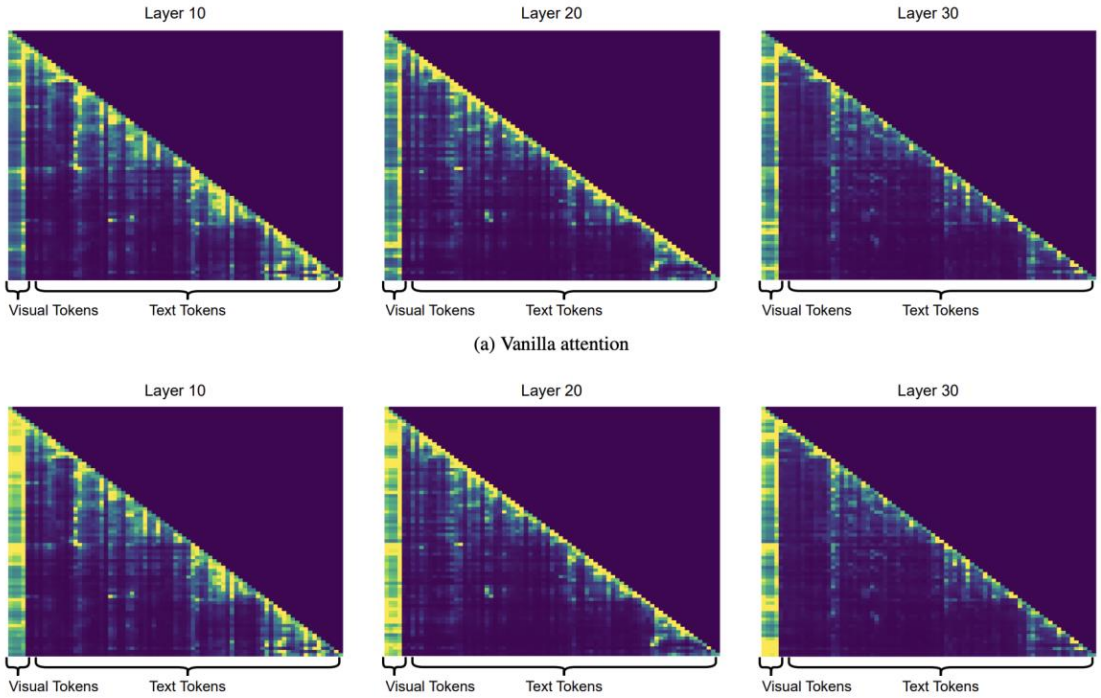
Method	NExT-QA [30]			
	Tem.	Cau.	Des.	Avg.
Video-ChatGPT [18]	37.6	65.1	54.9	54.6
+ **EDVT-Attention**	39.5 (+1.9)	72.8 (+7.7)	54.8	59.3 (+4.7)
VISTA-LLAMA (ViT-L-14)	34.0	69.1	42.2	53.6
+ **EDVT-Attention**	36.8 (+2.8)	72.2 (+3.1)	47.7 (+5.5)	56.5 (+2.9)
VISTA-LLAMA (EVA-CLIP-g)	34.3	65.8	55.9	54.1
+ **EDVT-Attention**	40.7 (+6.4)	72.3 (+6.5)	57.0 (+1.1)	59.7 (+5.6)

Table 3. **Comparison of EDVT-Attention design** with diffent



(a) Vanilla attention (b) EDVT-Attention (Ours)

Figure 5. **Comparison of attention weights** for varing context



(a) Vanilla attention

消融实验-visual projector

Visual projector	NExT-QA [30]			
	Tem.	Cau.	Des.	Avg.
Linear Projector	37.6	65.1	54.9	54.6
Q-Former (*BERT init.*)	35.2	62.7	49.2	51.8
Q-Former (*BLIP-2 init.*)	34.3	65.8	55.9	54.1
SeqQ-Former (*BLIP-2 init.*)	36.2	68.5	51.1	55.4

Table 4. **Comparison of different visual projectors** on NExT-QA [30]. The linear projector is initialized with pre-trained weights in LLaVa [14]. *BERT init.* and *BLIP-2 init.* indicate that the visual projector is initialized with weights from BERT [5] and BLIP-2 [12]. SeqQ-Former is the proposed sequential visual projector. See §4.3 for more details.

消融实验

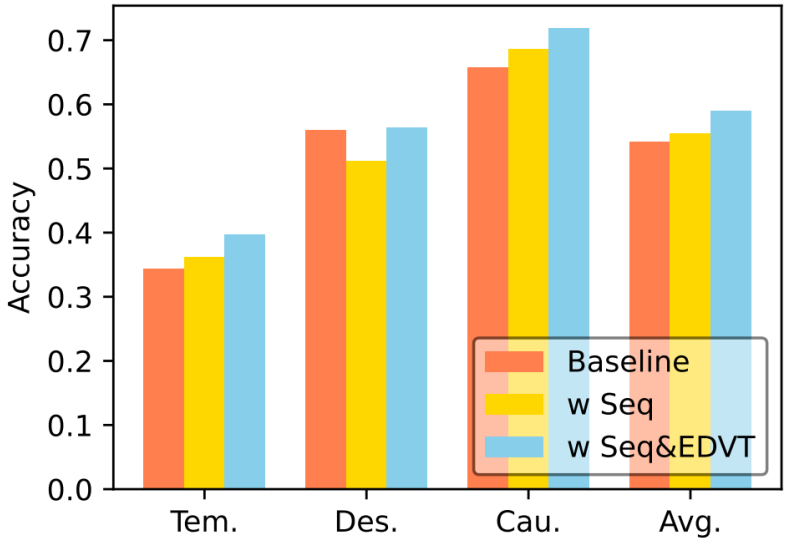


Figure 4. **Comparison of different design choices** on NExT-QA [30](§4.3). For clarity, accuracy of base model and two variants are given ("Baseline": the frozen LLaVA [14] with trainable Q-Former [12], "*w Seq*": base model with sequential visual projector, "*w Seq&EDVT*": base model with both sequential visual projector and EDVT-Attention).

VideoQA展示

Question: Which side of the phone did the person show the camera at the end ?

Answer: Show brand of phone

Video-ChatGPT : The person shows the camera at the bottom left side of the phone.

Ours: The person in the video shows the camera side of the phone at the end.

Question: What is the person in pink shorts doing when the man in white is playing with dog ?

Answer: Stand and watch

Video-ChatGPT: The person in pink shorts is sitting on the ground and watching the man play with the dog.

Ours: The person in pink shorts is watching the man in white play with the dog.

Question: What did the man do after he reached the cameraman ?

Answer: Stop

Video-ChatGPT: The man in the video is walking down a snow-covered mountain and is seen walking on a snowy mountain. He is wearing a backpack and appears to be walking on a snowy path. After reaching the cameraman, he turns around and continues walking down the mountain.

Ours: After the man reached the cameraman, he stopped and turned to face the camera.

Figure 7. **Visualization results** on different video questions. The questions and annotated answers are located on the left side. The generated text from Video-ChatGPT [31] and our model is presented in the green and orange boxes, respectively. See §4.3 for more details.

- 研究背景
- 研究方法
- 实验效果
- 总结

总结

- MLLM幻觉问题有两方面原因：
  - 视觉信息捕捉不足
  - LLM固有的幻觉问题
- Viusal token注意力：本文针对长序列中RoPE距离衰减问题，通过提高对视觉token的注意力权重缓解幻觉问题
- Text token注意力？

OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation

- 注意力图中的 aggregation现象与幻觉问题高度相关，提出aggregation导致过度自信

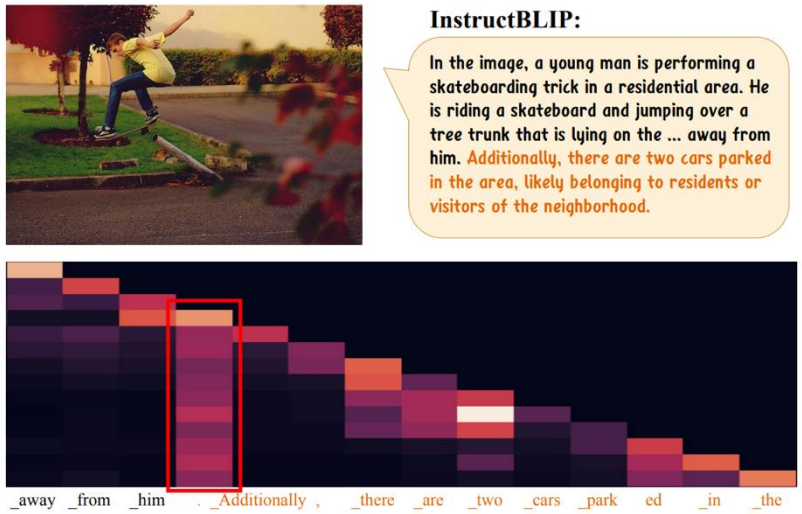


Figure 2. A case of relationship between hallucinations and knowledge aggregation patterns. Hallucinations are highlighted.

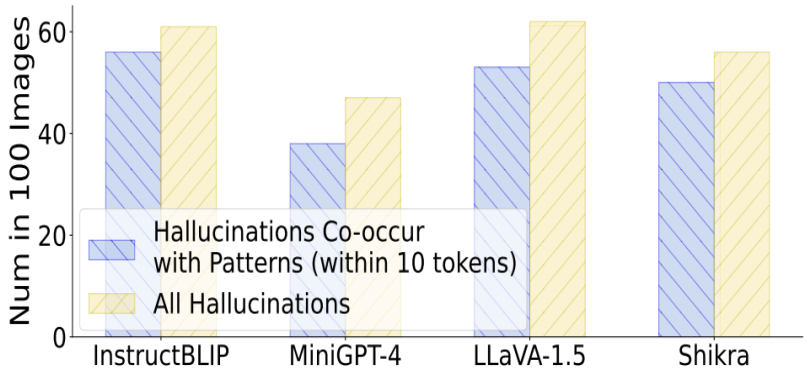
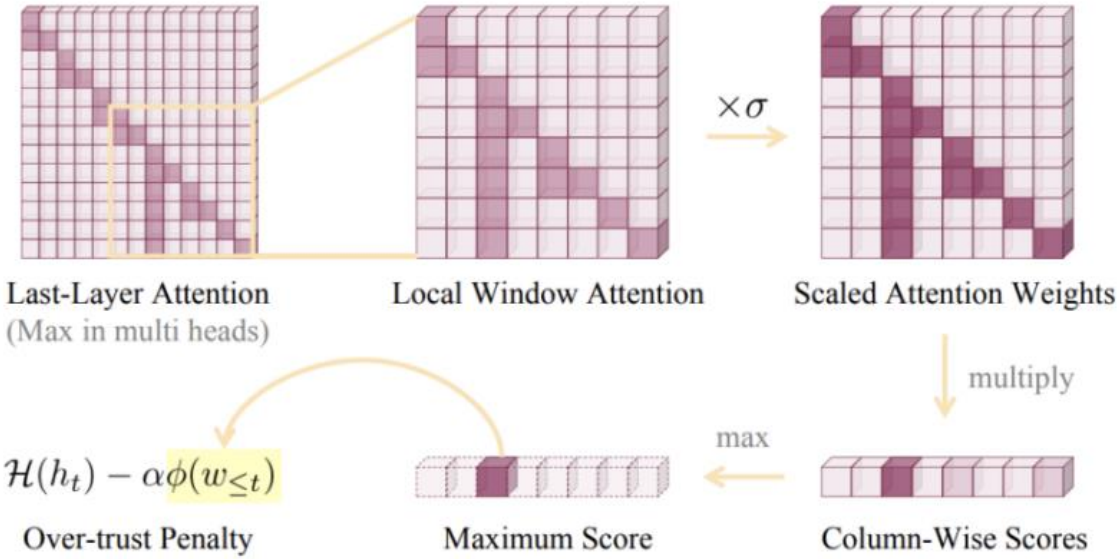


Figure 3. Hallucinations often start within the first 10 tokens after knowledge aggregation patterns.

Over-trust penalty



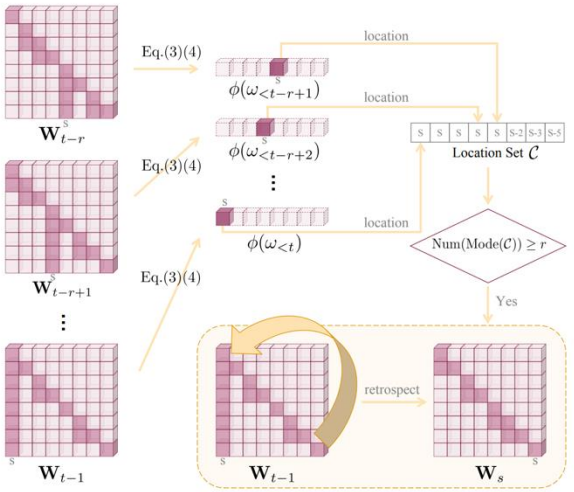


Figure 6. The scheme of the proposed Retrospection strategy. We compute the maximum value coordinates of the past several token's column-wise scores and check if the overlap time is larger than r. If yes, we retrospect the decoding procedure and reselect the next token x_{s+1}.

Thanks!