Seeing the Image: Prioritizing Visual Correlation by Contrastive Alignment

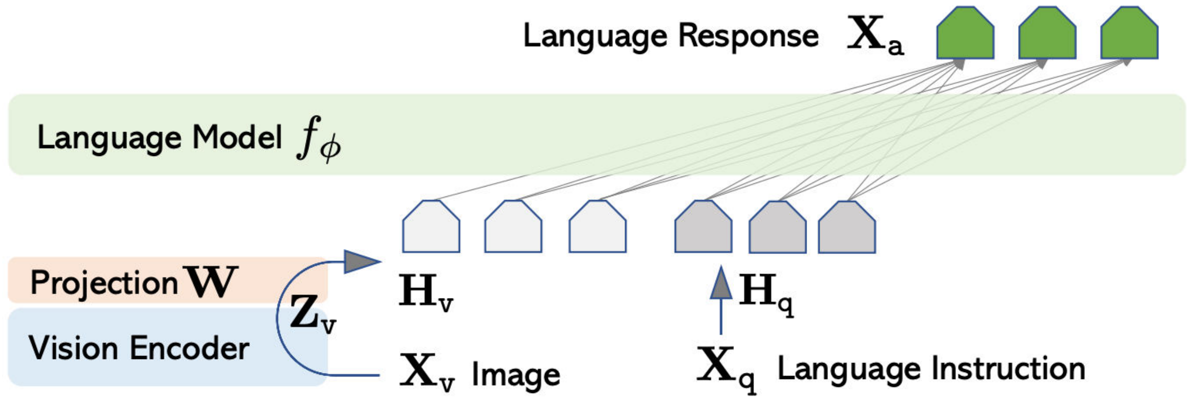
Xin Xiao, Bohong Wu, Jiacong Wang, Chunyuan Li, Xun Zhou, and Haoyuan Guo.

ByteDance.

Introduction to Vision Language Models

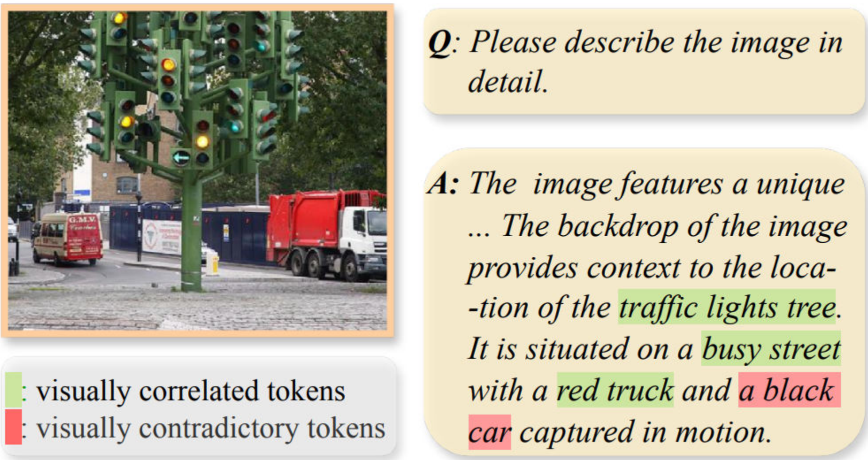
Motivation for Visual Correlation in Alignment

- **Vision Language Models (VLMs)**: Emerging multimodal systems that combine visual and textual data.
- **Challenges in Alignment:** Existing VLMs often treat all text tokens equally, ignoring the varying relevance of tokens to image content

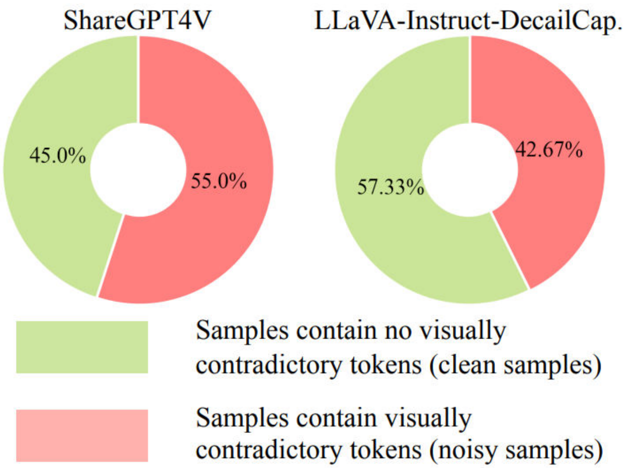


Problem Statement: Existing Alignment Strategies

Limitations in Token Correlation in VLMs



(a) Tokens correlate differently with the image.



(b) Human evaluations.

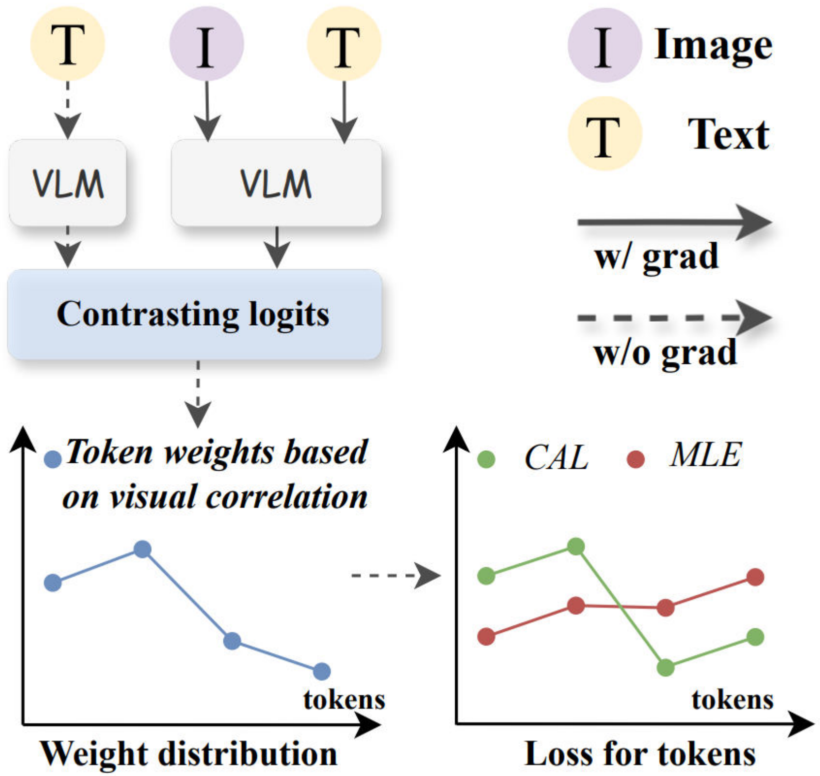
Different text tokens have varying degrees of correlation with the image and should not be treated equally.

Irrelevant or contradictory tokens can lead to poor alignment and degraded model performance.

Proposed Method: Contrastive Alignment (CAL)

A Novel Token Re-weighting Strategy

- **CAL**: A method to prioritize visually correlated tokens.
- **Process:** Re-weighting strategy based on prediction logit contrasts, distinguishing between visually relevant, irrelevant, and contradictory tokens.
- **Objective**: Enhances alignment with minimal computational overhead.



(b) Training procedures of *CAL*

Experiments and Results

NEURAL INFORMATION PROCESSING SYSTEMS

CAL achieved higher accuracy in OCR and image caption benchmarks.

Method	LLM	OCRB.	VQA^{Doc}	VQA^{Chart}	VQA^{Text}	SQA	MMS.	MMT.	Win/All
LLaVA-NeXT	Vicuna-7B	542	75.1	62.2	64.2	68.5	33.7	49.5	
LLaVA-NeXT+CAL	Vicuna-7B	**561**	**77.3**	**64.3**	**65.0**	**70.1**	**35.5**	**50.7**	7 / 7
LLaVA-NeXT	Vicuna-13B	553	78.4	63.8	67.0	**71.8**	37.5	50.4	
LLaVA-NeXT+CAL	Vicuna-13B	**574**	**80.1**	**67.2**	**67.1**	71.5	**38.1**	**52.4**	6 / 7

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments and Results

CAL achieved cleaner attention map and better text alignment

(a) a white van on a highway from Monks & Crane.

Conclusion and Future Directions

- **Summary**: CAL enhances image-text alignment by focusing on visually correlated tokens.
- **Contribution**: Helps VLMs concentrate on relevant data, advancing multimodal performance.
- **Future Work**: Further refinement of token-weighting strategies and adaptive settings for token bounds.