Seeing the Image: Prioritizing Visual Correlation by Contrastive Alignment

Xin Xiao[1,2*†], **Bohong Wu**[2*], **Jiacong Wang**[2,3†], **Chunyuan Li**[2], **Xun Zhou**[2], **Haoyuan Guo**[2]
[1]School of Computer Science, Wuhan University [2]ByteDance Inc.
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
https://github.com/foundation-multimodal-models/CAL

Abstract

Existing image-text modality alignment in Vision Language Models (VLMs) treats each text token equally in an autoregressive manner. Despite being simple and effective, this method results in sub-optimal cross-modal alignment by over-emphasizing the text tokens that are less correlated with or even contradictory with the input images. In this paper, we advocate for assigning distinct contributions for each text token based on its visual correlation. Specifically, we present by contrasting image inputs, the difference in prediction logits on each text token provides strong guidance of visual correlation. We therefore introduce **C**ontrastive **AL**ignment (*CAL*), a simple yet effective re-weighting strategy that prioritizes visually correlated tokens. Our experimental results demonstrate that *CAL* consistently improves different types of VLMs across different resolutions and model sizes on various benchmarks. Importantly, our method incurs minimal additional computational overhead, rendering it highly efficient compared to alternative data scaling strategies.

1 Introduction

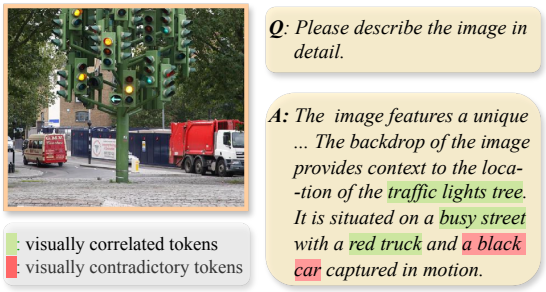
Recent advancements in Large Language Models (LLMs) [1–4] have opened up new avenues in multimodal understanding, giving rise to a novel model category known as Vision Language Models (VLMs) [5–8]. Many recent studies on VLMs are centered around enhancing their capabilities, either through increasing the resolution of input images [9–12] or incorporating higher-quality training datasets [13–16]. Additionally, research efforts have been directed towards exploring variations of vision models, such as replacing or augmenting vision encoders beyond CLIP [17, 18], including approaches like SigLIP [19], DINO [20, 21] or ConvNeXt [22, 23]. The integration of these techniques has spurred the development of VLMs, continually enhancing their performance across various benchmarks including visual question answering [24–28], image captioning [29, 30], and visual grounding [31, 32].

Despite these advancements, whether the current alignment strategy on existing image-text datasets performs satisfactorily is often less studied. Existing alignment strategies simply treat all text tokens equally in an auto-regressive manner. Although such a method has been proven to be simple and effective, many text tokens exhibit limited relevance to the visual inputs, which contribute little to image-text modality alignment. Figure 1a presents a sample drawn from the ShareGPT4V [16] dataset, where a large proportion of text tokens including *unique, context* presents little visual correlation. Treating these text tokens in equal weights results in ineffective training and can introduce negative effects by prioritizing more on fitting the distribution of these visually irrelevant tokens, rather than the image-text modality alignment.

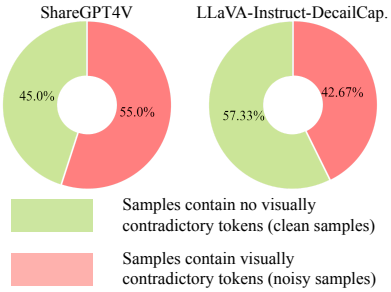
*Equal contribution
†Work done during internship in ByteDance

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0973



(a) Tokens correlate differently with the image.



(b) Human evaluations.

Figure 1: Figure 1a is one sample drawn from the ShareGPT4V dataset, which contains text tokens that are even contradictory with the given image. Figure 1b further presents our human evaluation results on the proportion of noisy samples that contain contradictory tokens.

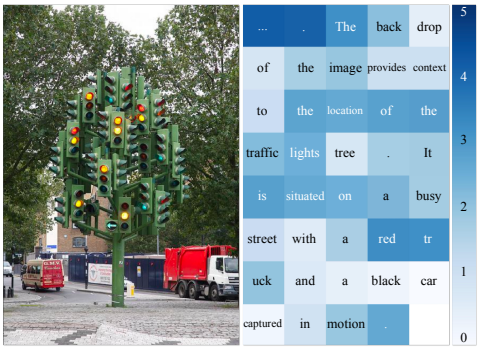
Moreover, there also exists a proportion of tokens that contradicts visual conditions, which is inevitable in model generated datasets [7, 16, 13, 15, 14], presented in Figure 1a. In particular, we conduct human evaluations on the broadly used GPT-assisted datasets including ShareGPT4V and the detail caption subset of LLaVA-Instruct in Figure 1b by sampling 100 samples from each dataset. We score each sample by 0 and 1 based on whether there exist text tokens that contradict with visual input, and the score is averaged by three annotators. We found approximately half of the sampled datasets contain visually contradictory tokens in both datasets. Imitating the text distribution on these contradictory tokens further harms the image-text modality alignment. Consequently, recent evaluations on existing VLMs [33, 34, 5, 35, 6, 8] present the shortcomings of current alignment strategy from various aspects, including hallucination [36, 37, 33] and responding without depending on visual conditions [27, 38].

Fortunately, inspired by recent training-free visual contrastive decoding researches [36, 33, 39], we present that the visual correlation can be directly indicated by contrasting input image conditions. In particular, we investigate the change in prediction logits of text tokens with or without the image input and observe strong relevance between the logit change of each text token and its visual correlation. We therefore propose **C**ontrastive **AL**ignment (*CAL*), which is a surprisingly simple re-weighting strategy to prioritize the training of text tokens that are highly correlated with the input image to enhance image-text modality alignment. Experiments have shown that our proposed method can improve leading VLMs of different kinds including LLaVA-1.5/LLaVA-NeXT [7, 6, 10], MiniGemini(MGM)/MGM-HD [11], across different resolution and model size on various types of benchmarks including visual question answering, captioning and grounding. Especially, *CAL* on LLaVA-Next-13B [10] can bring an impressive performance of 1.7 ANLS on VQA^{Doc} [26], 3.4 relaxed accuracy on $\text{VQA}^{\text{Chart}}$ [25], 2.2 CIDEr [40] on COCO [30] and 6.3 CIDEr on TextCaps [29], 0.6/0.7 IoU on validation/test set of RefCOCOg [32]. Moreover, our method introduces little computational overhead, with one auxiliary gradient-free forward operation in each training step. The lightweight feature while the impressive performance of *CAL* brings current VLMs to a new stage, highlighting the importance of a delicate image-text modality alignment strategy design. We further conduct extensive qualitative analysis for *CAL* and present the improved ability of OCR recognition and image-captioning.

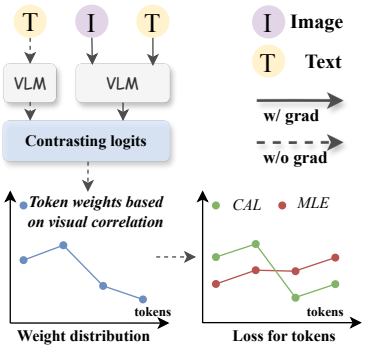
In summary, our contributions are listed as follows.

- We present that by contrasting image inputs, existing VLMs are able to distinguish visually correlated tokens both visually irrelevant and visually contradictory ones.
- We propose *CAL* , a contrastive image-text alignment method via token re-weighting, which is lightweight and effective. *CAL* requires little additional training cost and no additional inference cost.
- Experiments show that our *CAL* can consistently improve VLMs of different kinds, across different resolutions and sizes in various types of benchmarks.

https://doi.org/10.52202/079017-0973



(a) $\Delta \mathbf{o}^{i,j}_{[t_j]}$ heat map.



(b) Training procedures of *CAL*

Figure 2: Overview of *CAL* . Figure 2a presents a sample drawn from the ShareGPT4V dataset. We calculate the logit difference w/ or w/o image inputs and plot the heat map on partial text tokens. Figure 2b presents the training procedure of *CAL* , which re-weights the importance of label tokens based on the contrasting logits.

2 Contrastive Alignment

In this section, we describe the detailed design of *CAL*. First of all, we review the existing image-text modality alignment method and provide the notations in Section 2.1. Secondly, we show the token discrepancy in cross-modal datasets can be inferred via contrasting image inputs in Section 2.2. Finally, we present a detailed description of our proposed *CAL* in Section 2.3.

2.1 Preliminary and Notations

Preliminary Most existing VLMs adopt a two-stage strategy to align pre-trained image features with text embeddings in Large Language Models, i.e., a PreTraining (PT) stage that uses quantitative while noisy datasets for rough alignment, and an Instruction-Tuning (IT) stage that uses high-quality datasets to enhance the alignment. Both stages treat all tokens equally in an auto-regressive generation manner, i.e., the Maximum Likelihood Estimation (MLE) objective.

Notations In this paper, we denote the alignment dataset $\mathbf{D}$ consisting paired image-text samples $\mathbf{D} = \{(I^1, T^1), (I^2, T^2), ..., (I^n, T^n)\}$, and denote the logit computation function as $f(\theta)$ where θ is the weight of VLMs. For the i^{th} sample (I^i, T^i) in the training dataset, where T^i consists of a sequence of l tokens $T^i = [t^{i,1}, t^{i,2}, ..., t^{i,l}]$, we denote the prediction logit distribution without input I as $\mathbf{o}$, and prediction logit distribution with input I as $\tilde{\mathbf{o}}$, depicted in the following equation:

$$\mathbf{o}^{i,j} = f_\theta(T^{i,<j}), \tilde{\mathbf{o}}^{i,j} = f_\theta(I^i, T^{i,<j}) \tag{1}$$

where $T^{i,<j}$ represents all previous tokens before position j in the i^{th} sample. We further use $\mathbf{o}^{i,j}_{[t_j]}$ or $\tilde{\mathbf{o}}^{i,j}_{[t_j]}$ to represent the prediction logit in the i^{th} sample at token t_j. As a result, the MLE loss objective for the given i^{th} sample at token t_j is written in the following equation by treating the weight c of each token equally, where c is set to 1:

$$\mathcal{L}^{i,t_j}_{MLE} = c \cdot \log_{\text{softmax}}(\tilde{\mathbf{o}}^{i,j}_{[t_j]}) \tag{2}$$

2.2 Tokens Differ in Image-text Modality alignment

In this section, we present the necessity of token re-weighting in Section 2.2.1, and show that the re-weighting guidance could be naturally inferred by contrasting image inputs in Section 2.2.2.

2.2.1 Discrepancy exists in text tokens

Our proposed method begins with the discrepancy in the training label tokens. For image-text modality alignment, the training labels are usually natural texts, where not all text tokens have a

https://doi.org/10.52202/079017-0973

strong correlation with the image inputs. Moreover, due to the existence of model generated datasets, there also exist noisy text tokens that harm the alignment process, which is depicted in Figure 1.

Based on the relevance between corresponding input images, text tokens can be naturally divided into three kinds. (1) **Visually correlated tokens**, which contain clear visual concepts and movements depicted in the image. (2) **Visually irrelevant tokens**, which contain either irrelevant to the image inputs or could be easily inferred by previous text tokens. (3) **Visually contradictory tokens**, which contain hallucinated objects, especially in the model generated datasets.

2.2.2 Visually correlation can be inferred by contrasting image inputs.

Inspired by VCD [36] and IBD [33], which both enhance the generation by contrasting image inputs, we further present that the contrastive method can also provide clear guidance for visually correlation on each token.

We take the prediction logit of each label token under two circumstances, i.e., with or without the image inputs, which we denote as $\tilde{\mathbf{o}}^{i,j}_{[t_j]}$ and $\mathbf{o}^{i,j}_{[t_j]}$. Then denote $\Delta\mathbf{o}^{i,j}_{[t_j]} = \tilde{\mathbf{o}}^{i,j}_{[t_j]} - \mathbf{o}^{i,j}_{[t_j]}$ as the difference between the predictions logits across two circumstances, we plot $\Delta\mathbf{o}^{i,j}_{[t_j]}$ on each token in Figure 2a to visualize the effect of image conditions.

From Figure 2a, $\Delta\mathbf{o}^{i,j}_{[t_j]}$ performs impressively in distinguishing text tokens of three kinds. By $\Delta\mathbf{o}^{i,j}_{[t_j]}$, the visually correlated tokens *the traffic lights tree, busy street, red truck* are specially high-lighted while other tokens, especially the visually contradictory tokens *a black car* are light-colored.

In summary, the discrepancy in the label tokens motivates us to apply the token-wise dynamics on loss to enhance the image-text modality alignment. By contrasting the image inputs, the difference in the prediction logits $\Delta\mathbf{o}^{i,j}_{[t_j]}$ aids us with clear guidance for visually correlation of each text token.

2.3 Contrastive Alignment (*CAL*)

In this section, we present the details of our proposed *CAL*. *CAL* proposed to re-assign the contribution of each token based on their visually correlation weights. The overview of our method is shown in Figure 2b and the detailed algorithm is depicted in Algorithm 1.

Following the previous section, *CAL* first dynamically computes the visually correlation weight $\Delta\mathbf{o}^{i,j}_{[t_j]}$ of each token t_j by contrasting the image conditions. To avoid the effects of extreme values, we additionally introduce post-processing methods including clamping and average pooling. We clamp $\Delta logit$ by setting the upper bound to β and the lower bound to $\alpha(\alpha >= 0)$. By setting α to the extreme value 0, *CAL* neglects the visually irrelevant tokens and visually contradictory tokens. By setting β to the extreme value $+\infty$, *CAL* tolerates the circumstances where some visually correlated tokens occupy most of the importance weights in all label tokens:

Algorithm 1 Detail Procedure of $\mathcal{L}^{i,t_j}_{CAL}$

Input: i^{th} Image I^i, i^{th} Text $T = t_1, t_2, ..., t_n$, VLM f_θ

1: Compute contrastive logit.
$\mathbf{o}^{i,j} = f_\theta(T^{i,<j}), \tilde{\mathbf{o}}^{i,j} = f_\theta(I^i, T^{i,<j})$
2: Compute $\Delta logit$.
$\Delta\mathbf{o}^{i,j}_{[t_j]} = \tilde{\mathbf{o}}^{i,j}_{[t_j]} - \mathbf{o}^{i,j}_{[t_j]}$
3: Compute weights by post-processing $\Delta\mathbf{o}^{i,j}_{[t_j]}$.
$\tilde{\mathbf{w}}^{i,t_j} = pooling_W(clamp_{\alpha,\beta}(\Delta\mathbf{o}^{i,j}_{[t_j]}))$
4: Compute CAL loss by re-weighting tokens.
$\mathcal{L}^{i,t_j}_{CAL} = -\frac{1}{\sum_{k=1}^{l}\tilde{\mathbf{w}}^{i,t_k}}\tilde{\mathbf{w}}^{i,t_j} \cdot \log_{\text{softmax}} f_\theta(\tilde{\mathbf{o}}^{i,j}_{[t_j]})$

Output: $\mathcal{L}^{i,t_j}_{CAL}$

$$\mathbf{w}^{i,t_j} = clamp_{\alpha,\beta}(\Delta\mathbf{o}^{i,j}_{[t_j]}) \tag{3}$$

We further introduce average pooling with a window size of W to smooth the visually correlation weights of each token, where we denote as:

$$\tilde{\mathbf{w}}^{i,t_j} = pooling_W(\mathbf{w}^{i,t_j}) \tag{4}$$

The final loss objective of *CAL* is the weighted average of the original MLE objective based on w and is defined as:

$$\mathcal{L}^{i,t_j}_{CAL} = -\frac{1}{\sum_{k=1}^{l}\tilde{\mathbf{w}}^{i,t_k}}\tilde{\mathbf{w}}^{i,t_j} \cdot \log_{\text{softmax}} f_\theta(\tilde{\mathbf{o}}^{i,j}_{[t_j]}) \tag{5}$$

Table 1: Visual Question Answering benchmarks of *CAL* on leading methods including LLaVA-1.5, LLaVA-NeXT[1], and MGM/MGM-HD. Our results are marked with . VQA^{Text} is evaluated without OCR tokens. Abbreviations: OCRB. (OCR-Bench), MMS. (MMStar), MMT. (MMT-Bench).

Method	LLM	OCRB.	VQA			SQA^{I}	MMS.	MMT.	Win/All
			Doc	Chart	Text				
Low resolution setting									
MGM	Gemma-2B	335	39.8	23.4	48.1	60.6	**25.5**	43.4	6 / 7
MGM+*CAL*	Gemma-2B	**360**	**44.8**	**27.0**	**51.8**	**64.0**	25.4	**45.4**	
MGM	Vicuna-7B	431	57.7	**43.2**	61.1	69.9	32.8	50.3	6 / 7
MGM+*CAL*	Vicuna-7B	**443**	**58.0**	42.8	**63.0**	**70.4**	**35.5**	**51.4**	
MGM	Vicuna-13B	452	**61.7**	**48.8**	62.6	69.1	30.4	49.1	5 / 7
MGM+*CAL*	Vicuna-13B	**466**	61.6	48.0	**63.8**	**71.9**	**33.7**	**51.9**	
LLaVA-1.5	Vicuna-7B	315	28.5	17.5	**47.6**	68.2	32.4	48.6	5 / 7
LLaVA-1.5+*CAL*	Vicuna-7B	**328**	**30.6**	17.5	47.5	**68.7**	**32.9**	**48.8**	
LLaVA-1.5	Vicuna-13B	341	31.1	18.3	49.0	72.1	33.5	51.1	7 / 7
LLaVA-1.5+*CAL*	Vicuna-13B	**347**	**32.6**	**18.4**	**49.6**	**72.7**	**35.7**	**51.2**	
High resolution setting									
MGM-HD	Vicuna-7B	477	72.0	49.3	65.5	68.4	**31.0**	47.9	6 / 7
MGM-HD+*CAL*	Vicuna-7B	**503**	**73.4**	**49.6**	**67.1**	**69.2**	30.1	**50.5**	
MGM-HD	Vicuna-13B	502	77.7	55.8	67.2	**73.5**	34.2	50.9	6 / 7
MGM-HD+*CAL*	Vicuna-13B	**535**	**78.0**	**57.2**	**68.8**	73.1	**38.5**	**51.4**	
LLaVA-NeXT	Vicuna-7B	542	75.1	62.2	64.2	68.5	33.7	49.5	7 / 7
LLaVA-NeXT+*CAL*	Vicuna-7B	**561**	**77.3**	**64.3**	**65.0**	**70.1**	**35.5**	**50.7**	
LLaVA-NeXT	Vicuna-13B	553	78.4	63.8	67.0	**71.8**	37.5	50.4	6 / 7
LLaVA-NeXT+*CAL*	Vicuna-13B	**574**	**80.1**	**67.2**	**67.1**	71.5	**38.1**	**52.4**	

3 Experiments

3.1 Experimental Setup

Implementation Details In this paper, we verify our proposed *CAL* on two leading model structures: LLaVA-1.5/LLaVA-NeXT [6, 10] and Mini-Gemini/Mini-Gemini-HD [11]. LLaVA-1.5 uses CLIP-pretrained ViT-L as the visual encoder. For resolution scaling, LLaVA-NeXT employs a simple while adaptive image cropping strategy, encodes each image and concatenates them in one single sequence. Mini-Gemini (MGM) further introduces a LAION-pretrained ConvNeXt-L [22, 41] for high-resolution refinement. For MGM/MGM-HD/LLaVA-1.5, we follow the same setting as the original paper as it is public available, where the learning rate for the PT stage is set to $1e^{-3}$ and the IT stage is set to $2e^{-5}$ for both Vicuna-7B and Vicuna-13B. For LLaVA-NeXT, where only the evaluation code is made public, we reproduce LLaVA-NeXT with the same learning rate as MGM, and set the learning rate of ViT to 1/10 of the base learning rate (our reproduction presents on-par performance with the original paper/blog. We present a comparison of our reproduction results with those of the original papers in Appendix A.1). We also set the lower bound α and upper bound β in Equation (3) to 1 and 5 respectively, and we set l in Equation (4) to 3 for all experiments. We use 16 A100 for experiments, except for 8 GPUs in LLavA-1.5/Gemma-2B and 32 GPUs in MGM-HD-13B.

Datasets For experiments on LLaVA-NeXT [10], since the detailed composition of training datasets is not publicly available, we use a slightly different training dataset combination, where we include the mixture of $LLaVA_{665k}$ [6], VQA^{Doc} [26], VQA^{Chart} [25] and the ShareGPT4V [16]. For experiments of LLaVA-1.5 [6] and MGM/MGM-HD [11], we use the same dataset combination with original paper. The training datasets include LLaVA-filtered CC3M [42], ALLaVA [14], ShareGPT4V [16], LAION-GPT-4V [43], LIMA [44], OpenAssistant2 [45], VQA^{Doc} [26], VQA^{Chart} [25], DVQA [46] and AI2D [47]. Finally, we report results on widely-adopted VLM benchmarks, including VQA^{Text} [48](without providing OCR tokens), VQA^{Doc} [26], VQA^{Chart} [25], OCR-Bench [49], MMT [28], MMStar [27], SQA^{I} [50], COCO Caption [30], TextCaps [29], and RefCOCOg [32] in our main experiments which observe significant improvement in majority settings, and additional benchmarks MME [24], POPE [37], SEED-I [51], VQA^{Text*} [48](with OCR tokens given), with comparable performance in Appendix A.2.

[1]We reproduce the results on LLaVA-NeXT on a slightly different training data combination, since the instruction-tuning data is not made public.

https://doi.org/10.52202/079017-0973

Table 2: Image captioning and visual grounding benchmarks on LLaVA-1.5, LLaVA-NeXT, and MGM/MGM-HD[2]. Our results are marked with .

Method	LLM	COCO Caption	TextCaps	$\text{Refcocog}_{\text{val}}$	$\text{Refcocog}_{\text{test}}$	Win/All
Low resolution setting						
MGM	Gemma-2B	8.0	14.1	46.2	46.7	4 / 4
MGM+*CAL*	Gemma-2B	**29.7**	**25.1**	**50.9**	**51.1**	
MGM	Vicuna-7B	18.2	31.4	66.4	66.7	4 / 4
MGM+*CAL*	Vicuna-7B	**21.8**	**33.6**	**68.5**	**68.6**	
MGM	Vicuna-13B	17.6	27.1	73.2	73.5	4 / 4
MGM+*CAL*	Vicuna-13B	**18.6**	**32.1**	**73.7**	**74.1**	
LLaVA-1.5	Vicuna-7B	111.1	100.7	72.2	72.2	4 / 4
LLaVA-1.5+*CAL*	Vicuna-7B	**111.8**	**106.7**	**73.5**	**72.6**	
LLaVA-1.5	Vicuna-13B	115.9	102.4	74.6	75.2	3 / 4
LLaVA-1.5+*CAL*	Vicuna-13B	**119.4**	**105.6**	**75.3**	75.2	
High resolution setting						
MGM-HD	Vicuna-7B	**33.4**	42.4	70.2	70.6	3 / 4
MGM-HD+*CAL*	Vicuna-7B	31.3	**52.6**	**70.9**	**71.5**	
MGM-HD	Vicuna-13B	22.2	42.1	76.8	**77.7**	3 / 4
MGM-HD+*CAL*	Vicuna-13B	**30.0**	**51.6**	**77.2**	77.1	
LLaVA-NeXT	Vicuna-7B	112.0	115.0	77.6	77.5	4 / 4
LLaVA-NeXT+*CAL*	Vicuna-7B	**114.7**	**124.7**	**78.4**	**78.1**	
LLaVA-NeXT	Vicuna-13B	118.5	118.2	79.8	79.6	4 / 4
LLaVA-NeXT+*CAL*	Vicuna-13B	**120.6**	**124.4**	**80.4**	**80.3**	

3.2 Main Results

***CAL* effectively improves various VLMs in Visual Question Answering scenarios** Table 1 presents the performance of VLMs on Visual Question Answering. Our *CAL* consistently improves the performance on most of the understanding benchmarks with impressive margin, on different resolution settings on both MGM/MGM-HD and LLaVA-1.5/LLaVA-NeXT which have different vision model architectures. Especially on the high resolution setting, our *CAL* presents impressive performance improvement on OCR centric benchmarks. *CAL* can bring improvement in 7 out of 8 benchmarks on LLaVA-NeXT-7B, MGM-HD-7B/13B, and 6 out of 8 benchmarks on LLaVA-NeXT-13B. Especially, on LLaVA-NeXT-13B, *CAL* improves VQA^{Doc} by 1.7 ANLS, $\text{VQA}^{\text{Chart}}$ by 3.4 relaxed accuracy, and OCR-Bench by 21 points.

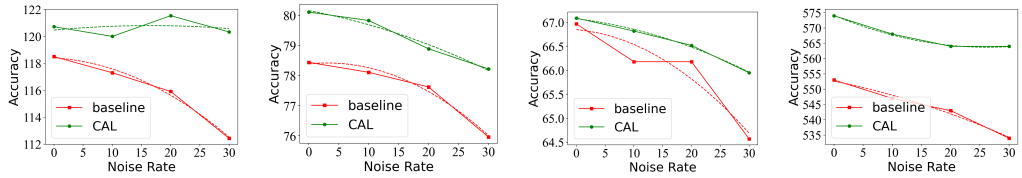
Effectiveness of *CAL* is consistent on other benchmarks including Image captioning and Grounding More promisingly, we found the improvement of *CAL* is not limited to Visual Question Answering benchmarks, but even more impressive on image-caption and visual grounding benchmarks. Table 2 presents the performance of VLMs on COCO Caption, TextCaps and both validation and test set of RefCOCOg. Especially, *CAL* improves LLaVA-NeXT-13B 2.1 CIDEr score on COCO Caption benchmark, 6.2 CIDEr score on TextCaps, 0.6/0.7 IoU on the validation/test set of RefCOCOg.

3.3 Ablation Studies

Stages of Conducting *CAL* within VLM Training *CAL* can be integrated into both the PreTraining (PT) stage and the Instruction Tuning (IT) stage in existing VLMs. In this section, we investigate which stage benefits the most from *CAL* in Table 3. Integrating *CAL* into the IT stage contributes most of the improvement in all listed benchmarks, and *CAL* in the PT stage further enhances the performance with notable improvement on MMT-Bench and OCR-Bench.

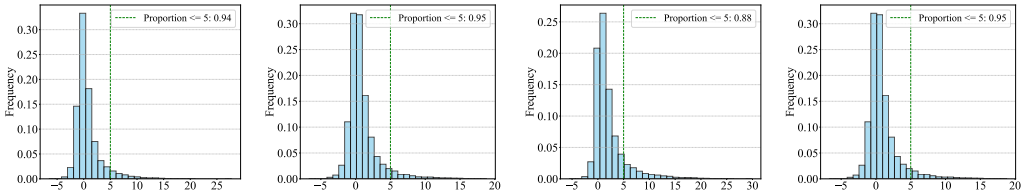
***CAL* relieve the effect of noisy labels in the training data** *CAL* can impressively reduce the effect of the contradictory label tokens in training data. To distinctly presents the denoising ability of *CAL* , we pollute the original training data with a ratio of 10 %, 20 % and 30 % by exchanging the training images and their labels, i.e., we select (I^i, T^i) and (I^j, T^j) with a probability of p and replace them with (I^i, T^j) and (I^j, T^i). We plot the performance difference when the noise rate is varied on four

[2]Note that the caption score of MGM/MGM-HD is significantly lower than LLaVA-NeXT. The training split of MGM does not include the training set for both COCO Caption and TextCaps, where we provide the zero-shot performance for MGM in all settings on these two benchmarks.



(a) COCO Caption (b) VQA$^{\text{Doc}}$ (c) VQA$^{\text{Text}}$ (d) OCR-Bench

Figure 3: Accuracy difference when different noise ratios applied. The performance of the baseline is marked with red lines, and *CAL* is marked with green lines. The dashed line represents the asymptote.



(a) LLaVA-1.5-7B (b) LLaVA-1.5-13B (c) LLaVA-NEXT-7B (d) LLaVA-NEXT-13B

Figure 4: $\Delta\mathbf{o}$ distribution for LLaVA models on 100 random sampled cases.

Table 3: Performance difference when *CAL* is applied at different training stages.

PT	IT	VQA$^{\text{Doc}}$	TextCaps	MMT.	OCRB.
		78.4	118.1	50.4	553
✓		78.5	116.7	50.9	554
	✓	79.9	124.8	51.5	564
✓	✓	80.1	124.4	52.4	574

Table 4: Performance difference when applying different weights $[\alpha, \beta]$ for clamping.

$[\alpha, \beta]$	VQA$^{\text{Doc}}$	VQA$^{\text{Chart}}$	OCRB.	Refcocog$_{\text{val}}$
$[0, +\infty]$	79.4	66.6	563	79.4
$[0, 5]$	80.4	66.3	570	79.7
$[1, +\infty]$	79.1	66.8	571	78.5
$[1, 5]$	80.1	67.2	574	80.4
Baseline	78.4	63.8	553	79.8

different benchmarks, including COCO caption, VQA$^{\text{Doc}}$, VQA$^{\text{Text}}$ and OCR-Bench. From Figure 3, the performance of baseline drops significantly when the noise rate is increased, while *CAL* performs more robust to noise.

Hyper-parameters for $[\alpha, \beta]$ in clamping We further conduct ablation study on α and β to study the effect of the hyperparameters in our clamping operation.

First, we plot the $\Delta\mathbf{o}$ distribution on LLaVA series in Figure 4. Tokens whose $\Delta\mathbf{o}$ is lower than 5 nearly occupy approximately 90% of the total label sequences. Therefore, we select 5 as the upper bound β. To prevent the context dependent tokens being totally neglected, we set the lower bound α to 1. We then extend both lower bound and upper bound to extreme values, i.e., 0 and $+\infty$. The results are shown in Table 4. (1) Both setting the lower bound to 0 and setting the upper bound to $+\infty$ causes performance degradation. (2) Even when setting the lower bound to 0, *CAL* still achieves better performance compared with the original baseline, indicating that totally neglecting the context dependent/visually contradictory tokens while focusing on the visually dependent ones still brings steady improvement.

Complementary Ablations We further provide ablations on the image contrasting conditions in Appendix A.3, where we compare *CAL* when different kinds of augmentation strategy is used.

3.4 Analysis

***CAL* presents better capability in OCR recognition and capturing details** We provide qualitative analysis in Table 5 to analyze the improved caption ability of *CAL*. In the table, *CAL* presents much better ability in OCR recognition by accurately distinguishing bewildering OCR cases, e.g., *CAL* accurately tells the difference between hand-written *consciousness* and *construction*, *world* and *would*. Such ability contributes to the superior performance of *CAL* on the TextCaps benchmark.

https://doi.org/10.52202/079017-0973

Meanwhile, *CAL* also presents better ability in capturing visually-conditioned details. For instance, compared with baseline, *CAL* captures the material details *cd cover*, and the numerical details by telling *two men on horses* from *a man on a horse*. The capability of *CAL* to capture intricate details leads to sustained enhancements in the COCO caption benchmark. *CAL* empowers the model to identify more accurate elements within images, including objects like *a trolley* and *the number 8*, which might otherwise be incorrectly recognized or overlooked.

Complementary Analysis We first provide statistics for computational overhead in Appendix A.6. And we further provide more qualitative analysis on studying the quality of image-text modality alignment. The attention map scores are visualized in Appendix B.1, and the aligned image features are visualized in Appendix B.2.

4 Related Work

Vision Language Models LLMs [1, 52, 2, 53, 4] have made significant strides in Natural Language Processing (NLP) tasks, including text generation and question-answering, paving the way for VLMs that integrate vision ability with LLMs. In the realm of visual language learning, CLIP [17, 54] has set a milestone by employing extensive image-text pair contrastive learning to achieve multimodal alignment. Recently, numerous VLMs [5, 7, 8, 55, 34, 56–59, 35, 60, 61] have leveraged the robust capabilities of LLMs for cross-modal understanding and generation tasks. Models like BLIP-2 [5] and MiniGPT-4 [8] have improved cross-modal alignment through comprehensive image-text pair pre-training. LLaVA [7] has further advanced its comprehension of complex prompts via refined instruction fine-tuning. Additionally, recent research [9–11] has incorporated higher resolution input images and longer sequences to enhance VLMs' understanding capabilities. Mini-Gemini (MGM) [11] introduces a LAION-pretrained ConvNeXt-L [22] for high-resolution refinement.

Image-text Modality Alignment Image-text modality alignment has long been regarded as the core problem in cross-modal understanding and generation tasks. Traditional image-text alignment strategies include both contrastive learning across different modalities and generative learning that train text tokens in an autoregressive manner [17, 62–64]. The combination of both techniques is also proven to be effective in the early era of VLMs, where BLIP [5] proposes a multi-stage alignment strategy, with contrastive learning in early alignment and generative learning in the latter stage. However, in recent researches [7, 10], contrastive learning is discarded for being redundant in image-text modality alignment of VLMs, and researchers propose to enhance the cross-modal alignment through dataset scaling and image resolution scaling [53, 9, 65, 10]. Despite being simple in application, the existing generative alignment method simply treats each text token with equal importance, resulting in sub-optimal alignment performance.

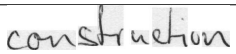
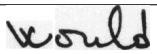
More recently, due to the great success in Reinforcement Learning (RL) in the alignment of LLMs, many recent works [66, 37, 67] have also integrated Reinforcement Learning methods to align existing VLMs with human preference. However, these RL-based methods require high-quality human-labeled pair-wise data and focus more on aligning with human preference rather than modality alignment.

Training-free Contrastive Decoding Recently, many researchers have proposed to improve the generation quality via contrastive decoding [68–70, 36, 33]. Especially, in the field of LLMs, CID [71] utilizes contrastive decoding on paired text inputs for model de-biasing. Such method is also proven to be effective in enhancing the reasoning ability of LLMs in various aspects [72–74]. Similar investigations have also been taking in VLMs. Recently, both VCD [36] and IBD [33] propose to enhance the generation of VLMs by contrasting the prediction logits between the original visual input and the perturbed ones. CRG [39] further proposes to improve the grounding ability without training via contrasting differently masked images. However, these training-free methods require additional computation during the decoding stage, making it highly ineffective for application.

5 Limitation

Despite the superiority of *CAL* in various model structures, resolution settings and model scales on various benchmarks, limitations still exist in our proposed method. First of all, there lacks a clear and

Table 5: OCR and captions generation comparison based on LLaVA-NEXT-13B model.

OCR cases		
		
Question:	What is written in the image?	What is the number in the image?
Baseline:	The word "consciousness" is written in the image.	The word "world" is written in the image.
CAL :	The word "construction" is written in the image.	The word "would" is written in the image.
		
Question:	What is the Mosman Manly exit going to?	How many items purchased from Amazon?
Baseline:	The mosman manly exit is going to mosman.	There are 2.43 million items purchased from amazon.
CAL :	The mosman manly exit is going to chatswood epping.	902k.
Short caption cases		
		
Question:	Provide a one-sentence caption for the provided image.	Provide a one-sentence caption for the provided image.
Baseline:	A red background with a sun and birds and the words Sibelius Symphonies No 2 & 5.	A book is open to a page with a picture of a man on a horse.
CAL :	A cd cover for Sibelius Symphonies Nos 2 & 5.	A book is open to a page with a picture of two men on horses.
		
Question:	Provide a one-sentence caption for the provided image.	Provide a one-sentence caption for the provided image.
Baseline:	A silver car is parked in front of a fence and a bus.	A person riding a horse with a cart attached to it.
CAL :	A silver car is parked behind a fence in front of a trolley.	A horse pulling a cart with the number 8 on it.

https://doi.org/10.52202/079017-0973

quantitative discrepancy between the three kinds of label tokens. More discussion of the importance weights guidance can be further investigated in future works.

The selection of lower bounds and upper bounds in Equation (3) are empirically decided based on the frequency of the prediction logits, which could be extended to more adaptive settings in further explorations. Nevertheless, the simple while broadly effective nature of *CAL* indicates the importance of a delicate image-text modality alignment strategy for leading VLM structures.

6 Conclusion

In this paper, we investigate the in-completeness of current image-text alignment in leading VLMs by treating all text tokens with equal weights. We present by contrasting input images, the difference in the prediction logits for each token naturally reveals their visual correlation. We therefore propose a token re-weighting strategy that prioritize the training of highly visually correlated tokens. Our proposed strategy, *CAL* is simple while impressively effective, achieving consistent performance gain across various benchmarks including visual question answering, image-captioning and grounding.

Our work raises a question about the potential optimal learning strategy of image-text modality alignment. Both the imperfectness of training data and over concentration on visually irrelevant/visually contradictory tokens hinder the performance of current VLMs. We hope the proposed *CAL* can inspire more investigation on better alignment strategy to enhance the capabilities of existing VLMs.

https://doi.org/10.52202/079017-0973

References

[1] OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2023. 1, 8

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 8

[3] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[4] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 1, 8

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2, 8

[6] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. 2, 5

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 8

[8] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2, 8

[9] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024. 1, 8

[10] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/. 2, 5, 8

[11] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 2, 5, 8

[12] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 1

[13] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2

[14] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 2, 5

[15] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024. 2

[16] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 1, 2, 5

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 8

https://doi.org/10.52202/079017-0973

[18] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[19] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1

[20] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 1

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1, 5, 8

[23] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 1

[24] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 1, 5

[25] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2, 5

[26] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 778–792. Springer, 2021. 2, 5

[27] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 2, 5

[28] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 1, 5

[29] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 1, 2, 5

[30] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2, 5

[31] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

[32] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 2, 5

[33] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024. 2, 4, 8

[34] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 8

[35] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 2, 8

[36] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023. 2, 4, 8

[37] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2, 5, 8

[38] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. 2

[39] David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. *arXiv preprint arXiv:2403.02325*, 2024. 2, 8

[40] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[43] LAION eV. Laion/gpt4v-dataset · datasets at hugging face. URL https://huggingface.co/datasets/laion/gpt4v-dataset. 5

[44] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[45] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[46] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018. 5

[47] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 5

[48] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 5

[49] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 5

[50] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 5

[51] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 5

[52] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8

 https://doi.org/10.52202/079017-0973

[53] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 8

[54] Yunhao Gou, Tom Ko, Hansi Yang, James Kwok, Yu Zhang, and Mingxuan Wang. Leveraging per image-token consistency for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19155–19164, 2023. 8

[55] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 8

[56] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 8

[57] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[58] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[59] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 8

[60] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 8

[61] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 8

[62] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 8

[63] Weicheng Kuo, AJ Piergiovanni, Dahun Kim, Benjamin Caine, Wei Li, Abhijit Ogale, Luowei Zhou, Andrew M Dai, Zhifeng Chen, Claire Cui, et al. Mammut: A simple architecture for joint learning for multimodal tasks. *Transactions on Machine Learning Research*, 2023.

[64] Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*, 2024. 8

[65] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 8

[66] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023. 8

[67] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 8

[68] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022. 8

[69] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424, 2021.

[70] Mian Zhang, Lifeng Jin, Linfeng Song, Haitao Mi, Wenliang Chen, and Dong Yu. Safeconv: Explaining and correcting conversational unsafe behavior. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22–35, 2023. 8

[71] Gal Yona, Or Honovich, Itay Laish, and Roee Aharoni. Surfacing biases in large language models using contrastive input decoding. *arXiv preprint arXiv:2305.07378*, 2023. 8

[72] Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*, 2023. 8

[73] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

[74] Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A Smith. Tuning language models by proxy. *arXiv preprint arXiv:2401.08565*, 2024. 8

 https://doi.org/10.52202/079017-0973

Outline

https://doi.org/10.52202/079017-0973

A Complementary Experimental Analysis

A.1 Ablations for Comparison with Official Results

Table 6: Comparisons between our reproduced results and official results. $*$ denotes providing OCR tokens for evaluation. Official results are marked with †.

Method	LLM	VQA				SQA^I	MME	POPE	SEED-I	COCO Caption	TextCaps
		Doc	Chart	Text	Text*						
Low resolution setting											
MGM	Gemma-2B	39.8	23.4	48.1	56.2	60.6	1335	85.7	64.6	8.0	14.1
MGM†	Gemma-2B	-	-	-	56.2	-	1341	-	-	-	-
MGM	Vicuna-7B	57.7	43.2	61.1	65.7	69.9	1551	85.7	68.2	18.2	31.4
MGM†	Vicuna-7B	-	-	-	65.2	-	1523	-	-	-	-
MGM	Vicuna-13B	61.7	48.8	62.6	66.1	69.1	1550	86.2	69.5	17.6	27.1
MGM†	Vicuna-13B	-	-	-	65.9	-	1565	-	-	-	-
LLaVA-1.5	Vicuna-7B	28.5	17.5	47.6	58.2	68.2	1468	86.2	66.6	111.1	100.7
LLaVA-1.5†	Vicuna-7B	28.1	18.2	46.1	-	69.5	1508	85.9	66.2	110.4	98.1
LLaVA1.5	Vicuna-13B	31.1	18.3	49.0	60.6	72.1	1574	85.7	68.6	115.9	102.4
LLaVa1.5†	Vicuna-13B	30.3	18.2	48.7	-	72.9	1523	85.9	68.2	115.5	104.0
High resolution setting											
MGM-HD	Vicuna-7B	72.0	49.3	65.5	68.2	68.4	1521	85.7	65.7	33.4	42.4
MGM-HD†	Vicuna-7B	-	-	-	68.4	-	1546	-	-	-	-
MGM-HD	Vicuna-13B	77.7	55.8	67.2	69.8	73.5	1633	86.3	70.1	22.2	42.1
MGM-HD†	Vicuna-13B	-	-	-	70.2	-	1597	-	-	-	-
LLaVA-NeXT	Vicuna-7B	75.1	62.2	64.2	67.1	68.5	1514	86.8	69.6	112.0	115.0
LLaVA-NeXT†	Vicuna-7B	74.4	54.8	64.9	-	70.2	1519	86.4	70.2	99.9	71.8
LLaVA-NeXT	Vicuna-13B	78.4	63.8	67.0	70.0	71.8	1574	87.2	71.8	118.5	118.2
LLaVA-NeXT†	Vicuna-13B	77.5	62.2	66.9	-	73.6	1575	86.3	71.9	102.0	67.4

We compare our reproduced results with those reported in the paper or from reliable sources in Table 6. The results for Mini-Gemini are taken from their paper, while the results for LLaVA are sourced from this sheet provided by the LLaVA authors. We can observe that our reproduced results are comparable to the official ones.

A.2 Results on other Visual Question Answering benchmarks.

Table 7: Results on additional Visual Question Answering benchmarks. $*$ denotes providing OCR tokens for VQA$^{\text{Text}}$. Our results are marked with .

Method	LLM	MME	POPE	SEED-I	VQA$^{\text{Text}*}$
Low resolution setting					
MGM	Gemma-2B	**1335**	85.7	64.6	56.2
MGM+*CAL*	Gemma-2B	1333	**85.8**	**65.0**	**58.1**
MGM	Vicuna-7B	**1551**	85.7	68.2	65.7
MGM+*CAL*	Vicuna-7B	1509	**87.5**	**69.0**	**66.7**
MGM	Vicuna-13B	**1550**	86.2	**69.5**	66.1
MGM+*CAL*	Vicuna-13B	1506	**86.4**	69.4	**67.1**
LLaVA-1.5	Vicuna-7B	1468	**86.2**	**66.6**	58.2
LLaVA-1.5+*CAL*	Vicuna-7B	**1500**	85.5	66.5	**58.6**
LLaVA1.5	Vicuna-13B	**1574**	85.7	68.6	**60.6**
LLaVa1.5+*CAL*	Vicuna-13B	1567	**85.8**	**69.3**	60.1
High resolution setting					
MGM-HD	Vicuna-7B	1521	85.7	65.7	68.2
MGM-HD+*CAL*	Vicuna-7B	**1531**	**87.0**	**68.6**	**69.5**
MGM-HD	Vicuna-13B	**1633**	86.3	70.1	69.8
MGM-HD+*CAL*	Vicuna-13B	1582	**86.4**	**70.5**	**70.3**
LLaVA-NeXT	Vicuna-7B	**1514**	**86.8**	69.6	67.1
LLaVA-NeXT+*CAL*	Vicuna-7B	1490	86.7	**70.7**	67.1
LLaVA-NeXT	Vicuna-13B	1574	87.2	**71.8**	**70.0**
LLaVA-NeXT+*CAL*	Vicuna-13B	**1606**	87.2	71.6	68.9

https://doi.org/10.52202/079017-0973

Table 8: Ablations for contrasting image conditions on Visual Question Answering benchmarks using LLaVA-NeXT/13B.

Method	Mask	σ	MMS.	MMT.	SQA^I	VQA				OCRB.
						Text	$Text^*$	Doc	Chart	
LLaVA-NeXT	-	-	37.5	50.4	71.8	67.0	70.0	78.4	63.8	553
LLaVA-NeXT+*CAL*	-	-	38.1	**52.4**	71.5	67.1	68.9	**80.1**	**67.2**	574
LLaVA-NeXT+CAL_{mask}	0.5	-	37.9	51.9	70.5	67.1	69.5	79.0	66.0	555
LLaVA-NeXT+CAL_{mask}	0.7	-	**38.9**	51.3	**72.1**	66.1	69.2	79.3	65.5	541
LLaVA-NeXT+CAL_{mask}	0.9	-	38.3	52.0	71.8	67.0	69.4	78.9	65.9	**580**
LLaVA-NeXT+$CAL_{gaussian}$	-	1	37.9	51.7	71.2	66.8	69.6	78.7	65.7	554
LLaVA-NeXT+$CAL_{gaussian}$	-	10	38.5	52.0	70.2	**67.3**	**70.2**	79.1	64.6	564

Table 9: Ablations for image contrasting conditions on image captioning and visual grounding benchmarks using LLaVA-NeXT/13B.

Method	Mask	σ	COCO Caption	TextCaps	$Refcocog_{val}$	$Refcocog_{test}$
LLaVA-NeXT	-	-	118.5	118.2	79.8	79.6
LLaVA-NeXT+*CAL*	-	-	**120.6**	**124.4**	80.4	80.3
LLaVA-NeXT+CAL_{mask}	0.5	-	114.6	122.9	**80.9**	80.8
LLaVA-NeXT+CAL_{mask}	0.7	-	116.0	122.9	80.6	**81.7**
LLaVA-NeXT+CAL_{mask}	0.9	-	117.7	121.5	80.6	81.1
LLaVA-NeXT+$CAL_{gaussian}$	-	1	117.3	118.9	80.3	80.1
LLaVA-NeXT+$CAL_{gaussian}$	-	10	116.8	123.4	79.2	79.8

In the primary sections of this manuscript, *CAL* achieves better performance on benchmarks of OCR centric VQA, image-captioning, visual grounding, and other image-dependent benchmarks. In this section, we extend our analysis by presenting additional results on MME, POPE and SEED-I and VQA^{Text} (with OCR token given). *CAL* presents comparable performance on these benchmarks, without a clear distinction on performance in different training settings. With steady improvement on OCR centric or image dependent VQA benchmarks, *CAL* also presents satisfying quality on the remaining VQA tasks.

A.3 Ablations for Image Contrasting Conditions

In this section, we further conduct additional analysis on the image contrasting conditions. Except from masking the whole image token sequence, which is the method we use in our paper, we further study two kinds of contrasting conditions, including random patch masking and Gaussian blurring. For random patch masking, we cut images into 20×10 grids, with 10 pieces in the width and 20 pieces in the height. Therefore each images is cut into 200 pieces. We then randomly select 50 %, 70 % and 90 % of the patches to mask, and contrast them with the original image. For Gaussian blurring, we set the kernel σ to 1 and 10 to simulate the extreme blurring circumstances. By using 1, we adopt light blurring which only brings significant effect on OCR centric images. By using 10, all images will be heavily affected.

Table 8 presents the results on Visual Question Answering benchmarks of different image contrasting conditions on LLaVA-NeXT-13B, and Table 9 further presents results on image-caption benchmarks and visual grounding benchmarks. From these two tables, either random patch masking or adding Gaussian blurring performs sub-optimal in different benchmarks. By CAL_{mask}, except for the significant improvement on the test split of RefCOCOg, performance of CAL_{Gaussian} drops significantly, especially on the OCR-centric benchmarks. By CAL_{Gaussian}, although the performance on OCR centric benchmarks (VQA^{Doc}, VQA^{Chart}, OCR-Bench) is less affected, it performs less effective in nearly all benchmarks than *CAL* .

A.4 Ablations for Average Pooling

We fix the window size (W) to 3 in Equation (4) to smooth the weight distribution. An additional experiment in Table 10, where we removed the average pooling step, shows slightly inferior performance, supporting our chosen.

Table 10: Comparison of benchmarks with and without Average Pooling.

	ChartVQA	**DocVQA**	**SQA^I**	**COCO Caption**	**TextCaps**	**OCRB.**	**Refcocog$_{val}$**
w/ AvgPool	67.2	80.1	71.5	120.6	124.4	574	80.4
w/o AvgPool	66.3	79.5	72.4	116.7	123.8	581	79.5

A.5 Ablations for Pre-trained Model

Table 11: Comparison of pre-trained models for *CAL* on LLaVA-Next-13B.

Model	Pretrain	**ChartVQA**	**DocVQA**	**SQA^I**	**COCO Caption**	**TextCaps**	**OCRB.**	**Refcocog$_{val}$**
Baseline	Original	63.8	78.4	71.8	118.5	118.2	553	79.8
CAL	Original	67.2	80.1	71.5	120.6	124.4	574	80.4
CAL	Baseline	66.3	79.5	72.4	116.7	123.8	581	80.2

One assumption of *CAL* is that after training a simple projector only, the VLM is capable of distinguishing visually correlated tokens. In the first phase of VLM training, it is common practice to freeze the ViT and the LLM, and only train a projector to align their features. In the second phase, we finetune the model using high-quality data to enable it to answer image-related questions. To validate this assumption, we finetune *CAL* models using two types of pre-trained versions: one with the original pre-trained model (only train the projector) and one with the fully trained baseline model (after the finetuning phase). As shown in Table 11, we compare the performance of these two versions and found no significant differences in performance.

A.6 Computational Overhead.

Table 12: Training time comparison with and without *CAL* in the instruction-tuning stage on LLaVA-NeXT.

Pretrain	LLaVA-NeXT-7B	LLaVA-NeXT-13B
Baseline	15.6*h*	27.4*h*
Baseline + *CAL*	19.1*h*	33.7*h*

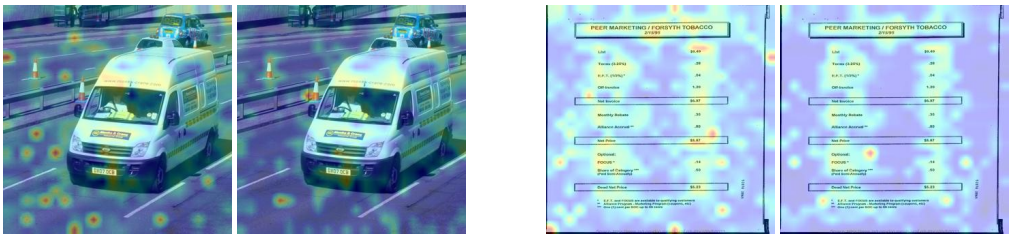
Each iteration of our proposed method requires forwarding text tokens twice to effectively enhance image-text modality alignment. Table 12 presents the training time for the instruct-tuning stage on LLaVA-NeXT, with each experiment conducted on 16 A100 GPUs. *CAL* introduces approximately 20% computational overhead on both LLaVA-NeXT-7B and LLaVA-NeXT-13B. Our implementation currently uses an attention mask to neutralize the effect of the image tokens, meaning that the image tokens are still forwarded twice. This additional burden could be further reduced by completely removing image tokens.

https://doi.org/10.52202/079017-0973

B Complementary Qualitative Analysis

B.1 Attention Map Visualization

In this section, we visualize the attention maps generated by both the baseline model and our proposed model. These visualizations provide insight into how each model focuses on different parts of the input image. The attention weights are calculated by accumulating the attention score between image tokens and text tokens across all layers. As shown in Figure 5, model w/ *CAL* provides clearer attention maps with less noisy points in the background area, indicating a more precise focus on relevant regions.



(a) a white van on a highway from Monks & Crane. (b) $8.49.

Figure 5: Comparison of attention maps with and without *CAL* on LLaVA-NeXT-13B. The left side of each sub-figure shows LLaVA-NeXT-13B without *CAL* , while the right side shows LLaVA-NeXT-13B with *CAL* .

B.2 Image-text Modality Alignment Visualization

In this section, we visualize the image-text modality alignment by retrieving the nearest text words to each image patch feature from the LLM vocabulary. We plot the results in Figure 6. *CAL* brings better image-text modality by accurately retrieving the OCR information from the language vocabulary. E.g., LLaVA1.5-7B + *CAL* correctly retrieves *Prices, expert, 0, shadow*, which is exactly the OCR information in the original image. We do not select LLaVA-NeXT in this section due to the presence of token overlaps from sub-crops.

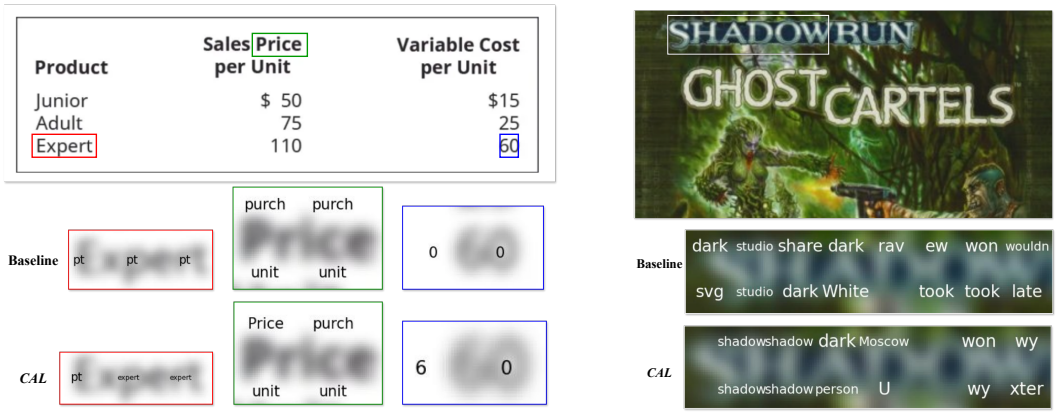


Figure 6: Visualization of image-text modality alignment for each image patch. We filtered out some nonsensical patches for better visualization.

https://doi.org/10.52202/079017-0973

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims fully reflect our paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-0973

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the specific parameters used in the experiment and the information needed to reproduce it.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release our code after the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We show all the training and test details in our paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the same guideline in our experiment as those previous works.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

https://doi.org/10.52202/079017-0973

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of our work performed for our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the creators or original owners of assets used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected. All datasets we use are from internet open source datasets under CC-BY licenses and are cited properly.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

https://doi.org/10.52202/079017-0973

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.