Extract Free Dense Misalignment from CLIP

JeongYeon Nam[1*], **Jinbae Im**[1], **Wonjae Kim**[2], **Taeho Kil**[1]

[1] NAVER Cloud AI
[2] NAVER AI Lab

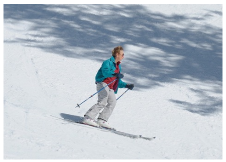
Abstract

Recent vision-language generative models still frequently produce outputs misaligned with their inputs, evidenced by object hallucination in captioning and prompt misalignment in the text-to-image generation model. Recent studies have explored methods for identifying misaligned elements, aiming not only to enhance interpretability but also to improve model performance. However, current approaches primarily rely on large foundation models in a zero-shot manner or fine-tuned models with human annotations, which limits scalability due to significant computational costs. This work proposes a novel approach, dubbed CLIP4DM, for detecting dense misalignments from pre-trained CLIP, specifically focusing on pinpointing misaligned words between image and text. We carefully revamp the gradient-based attribution computation method, enabling negative gradient of individual text tokens to indicate misalignment. We also propose F-CLIPScore, which aggregates misaligned attributions with a global alignment score. We evaluate our method on various dense misalignment detection benchmarks, covering various image and text domains and misalignment types. Our method demonstrates state-of-the-art performance among zero-shot models and competitive performance with fine-tuned models while maintaining superior efficiency. Our qualitative examples show that our method has a unique strength to detect entity-level objects, intangible objects, and attributes that can not be easily detected for existing works. We conduct ablation studies and analyses to highlight the strengths and limitations of our approach.

Code — https://github.com/naver-ai/CLIP4DM

Introduction

While recent advancements in generative models have garnered unprecedented progress, large-scale models still produce outputs misaligned with their inputs, exemplified by object hallucination (Li et al. 2023; Gunjal, Yin, and Bas 2024) in image-to-text (captioning) models and misalignment with text description (Rassin, Ravfogel, and Goldberg 2022; Chefer et al. 2023) in text-to-image generation models. It is crucial to effectively detect these misalignments in order to develop a more reliable system.



Caption: A man riding snowboard down a snow covered slope.
CLIPScore: 61.3
Ours: A man riding *snowboard* down a snow covered slope.
Misaligned word: *snowboard*

Figure 1: **Overview of our work.** CLIPScore indicates the alignment between the image and text in a single scalar score, limiting the interpretation of the score. Our approach extracts both positive and negative attributions to identify misaligned tokens between the image and text caption.

To measure the alignment between an image and text, the similarity score from CLIP (Radford et al. 2021; Hessel et al. 2021) has become a de facto approach. However, as shown in Figure 1, this simple score lacks the granularity needed to identify specific misaligned words, limiting interpretability (Hu et al. 2023; Cho et al. 2024). To address this limitation, recent studies (Petryk et al. 2024; Gordon et al. 2024) have focused on detecting misalignments at a dense level (e.g., word, phrase) and provide feedback to the models (Yu et al. 2022; Yan et al. 2024) for further enhancement. These approaches either employ a pipeline comprising multiple foundation models in zero-shot or fine-tuned configurations, leveraging costly human-annotated data. While these methods show promising results, their computational expense limits their applicability in practical scenarios.

In this paper, we introduce a novel approach that leverages pre-trained CLIP for detecting dense misalignments efficiently. Specifically, our work aims at pinpointing words inconsistent with the image, offering richer explanations for text-image misalignments. While CLIP's final output is a single similarity score, we hypothesize that rich token-specific information is embedded within the model's intermediate representations, such as attention maps and gradients with respect to them. We propose a new method, dubbed as CLIP4DM(**CLIP** for **d**ense **m**isalignment), which carefully modifies existing gradient-based attribution assignment techniques (Selvaraju et al. 2017; Chefer, Gur, and Wolf 2021a,b). We compute attribution scores for each text token primarily based on relevance propagation methods, where our method is modified so that each relevance score can also be a negative attribution value. Then, we predict

[*] Corresponding author: jy.nam@navercorp.com

misaligned tokens by identifying text tokens with negative attribution lower than the threshold as shown in Figure 1. We also introduce F-CLIPScore, which combines the overall score with calculated attributions of misaligned tokens.

We thoroughly evaluate our method on diverse dense misalignment detection benchmarks (FOIL (Shekhar et al. 2017), nocaps-FOIL (Petryk et al. 2024), HAT (Petryk et al. 2024), SeeTRUE-Feedback (Gordon et al. 2024), and Rich-HF (Liang et al. 2024)), encompassing various text, image, and misalignment types. The results consistently demonstrate that our method achieves state-of-the-art performance among zero-shot models and competitive performance with fine-tuned models. Qualitative assessments reveal that our method robustly handles various misalignments, such as entity-level object class, intangible objects, and attributes. Moreover, our method demonstrates significantly higher efficiency compared to baselines, which utilize large foundation models, suggesting its potential for practical applications.

Related Work

Dense Misalignment Detection

There has been a growing emphasis on detecting dense misalignments between image and text, which focuses on identifying specific misaligned regions or tokens within the text. This approach provides detailed feedback that improves the evaluation of image-text alignment. Shekhar et al. (2017) introduce FOIL benchmark for detecting and correcting misaligned words, where replacing one noun of COCO caption with a semantically similar one. ALOHa (Petryk et al. 2024) extends its coverage to various objects while leveraging multiple foundation models. ALOHa makes a candidate object pool with an extracted noun phrase from reference captions and the results of object detectors (Carion et al. 2020), then perform bipartite matching based on scores derived from a language semantic similarity model (Reimers and Gurevych 2019). SeeTRUE-Feedback (Gordon et al. 2024) utilizes LLMs and a visual grounding model to create a dataset of textual and visual misalignment descriptions, which are then used to train a vision-language model for automatic feedback generation. Rich-HF (Liang et al. 2024) focuses on misalignments in a text-to-image generation model while collecting human annotations on misaligned keywords and implausible image regions and trains a multimodal language model to show the dense image-text alignment automatically.

Beyond simply detecting dense misalignments, there have been studies leveraging dense misalignment labels to enhance model performance or reduce object hallucinations, particularly in the context of reinforcement learning-based approaches. As the length of sequences generated by LLMs increases, the problem of hallucination becomes pronounced, making dense feedback that reduces ambiguity inherent in single scalar reward more critical. Yu et al. (2024) and Xiao et al. (2024) tackle object hallucination in large vision-language models by incorporating dense-level (e.g., sub-sentence, sentence) human feedbacks. ViGoR (Yan et al. 2024) additionally employs a pipeline combining named entity recognition models with open vocab object detector (Liu et al. 2024) to detect hallucinations automatically. However, its scope is limited to object hallucinations, and human annotations are still needed to detect comprehensive misalignments.

In summary, the increasing emphasis on dense misalignment detection underscores its crucial role in developing more interpretable and reliable vision-language models. While current work demonstrates promising results in providing dense misalignment detection, they predominantly rely on costly human annotations or incorporation of foundation models, resulting in substantial cost overhead. In this work, we propose a cost-efficient dense misalignment detection method, leveraging the pre-trained CLIP in a zero-shot manner. The result demonstrates its efficiency and competitive performance over other cost-expensive zero-shot baselines.

Explainable AI Methods

Understanding the decision-making process of complex machine learning models is crucial for building trust and ensuring reliable performance. Explainable AI (XAI) methodologies address this need by providing insights into how models arrive at their predictions. XAI methods can be broadly categorized into two groups: input manipulation methods and mechanistic approaches. Input manipulation methods, such as SHAP (Lundberg and Lee 2017), occlusion analysis (Zeiler and Fergus 2014), and LIME (Ribeiro, Singh, and Guestrin 2016), perturb or mask input features to observe their impact on model output. While intuitive, these methods are often computationally expensive, especially for large models and datasets.

Mechanistic approaches, on the other hand, delve into the internal workings of the model to directly analyze feature contributions. Grad-CAM (Selvaraju et al. 2017) uses class-specific gradients to highlight relevant input regions but can produce coarse visualizations. LRP (Bach et al. 2015), grounded in the Deep Taylor Decomposition framework (Montavon et al. 2017), propagates relevance scores backward through the network layers, ensuring conservation of relevance. LRP has been successfully applied to various tasks, including image classification (Bach et al. 2015), NLP (Arras et al. 2017), and vision-and-language tasks (Chefer, Gur, and Wolf 2021b), showcasing its versatility and effectiveness.

The widespread adoption of Transformer networks (Vaswani et al. 2017) in NLP and vision-and-language tasks brought new challenges for XAI. Rollout (Abnar and Zuidema 2020) and Attention Flow (Abnar and Zuidema 2020) attempt to address complexities arising from self-attention, but limitations persist. Chefer, Gur, and Wolf (2021b) adapted LRP for single-modality Transformers, later extending it to multi-modal settings using a combination of attention scores and gradients for head averaging (Chefer, Gur, and Wolf 2021a). Unlike these approaches, which rely on positive-only relevance propagation, our work introduces the interpretation of negative attributions as indicators of misalignment in CLIP. Recent studies (Zhou, Loy, and Dai 2022; Wang, Rudner, and Wilson 2023; Zhao et al. 2024) apply XAI techniques to CLIP; however, they also focus on identifying only relevant image regions corresponding to the text.

Method

Preliminary: CLIP

We provide a brief overview of the key elements of the CLIP architecture. We also define the relevant terminology to consistently notate our method.

CLIP employs a dual-encoder structure, processing image and text modalities through separate encoders. The text encoder takes a sequence of tokens padded or truncated to a fixed length n,

$$t = [t_0, t_1, ..., t_z, ..., t_{n-1}], \tag{1}$$

where z is the index of the [EOS] token in the sequence. The image encoder processes the input image as a sequence of patches, including a special [CLS] token.

$$v = [v_0, v_1, ..., v_m], \tag{2}$$

where v_0 is the [CLS] token and $v_1, ..., v_m$ are image patches. The input image patches v and text tokens t are first forwarded through the image encoder (V) and text encoder (T), respectively, after which the representations are pooled and projected from the [CLS] and [EOS] tokens:

$$e_v = W_v(V(v)[0,:]), \quad e_t = W_t(T(t)[z,:]), \tag{3}$$

where W_v and W_t are projection matrices. The final score is computed by the cosine similarity (dot product with L2 normalization):

$$\text{score}_{v,t} = \frac{e_v}{||e_v||_2} \cdot \frac{e_t}{||e_t||_2}. \tag{4}$$

This score indicates the degree of semantic alignment between the image and text inputs.

Our Method

In this section, we introduce our attribution calculation method, which is inspired by Generic Attention-model Explainability (GAE) (Chefer, Gur, and Wolf 2021a). We first introduce GAE briefly and how our method is different from GAE. We then introduce fine-grained CLIPScore (F-CLIPScore), a drop-in replacement of CLIPScore by aggregating word attributions.

Generic Attention-model Explainability To determine the direction and magnitude of each token's attribution to the final output, GAE computes the gradients of the final score with respect to the attention map:

$$\nabla A_l^h = \frac{\partial \text{score}_{v,t}}{\partial A_l^h}, \tag{5}$$

where $A_l^h \in \mathbb{R}^{n \times n}$ denotes the attention map at l-th layer and h-th head. To aggregate its gradient, GAE calculates the element-wise product of this gradient with the corresponding attention map:

$$R_l^h = \text{ReLU}(\nabla A_l^h \odot A_l^h). \tag{6}$$

Note that relevance propagation methods (Selvaraju et al. 2017; Chefer, Gur, and Wolf 2021a,b) typically employ ReLU operation in ∇A_l^h to remove negative attribution.

The relevancy for layer l is obtained by averaging across attention heads:

$$R_l = \frac{1}{H} \sum_{h=1}^{H} R_l^h. \tag{7}$$

The relevancy in the final layer is initialized as an identity matrix and updated layer by layer. At each layer l, R is updated by adding the product of the current layer's relevancy R_l and the carried R as in the relevance propagation methods. This process propagates the attribution information through the network, accumulating each layer's attribution. Finally, the relevancy is aggregated along the [EOS] token row, $R[z,:]$.

Allowing Negative Gradient Flow. Unlike GAE (Chefer, Gur, and Wolf 2021a), which focuses on only the positive value of gradient, our work aims to identify misaligned words by incorporating negative gradients. We simply remove the ReLU operation on Equation (6), allowing negative gradients to explain the model's behavior.

$$R_l^h = \nabla A_l^h \odot A_l^h. \tag{8}$$

By adopting this formulation, our approach leverages both positive and negative gradients to capture a comprehensive spectrum of attributions.

Layer Aggregation. Since our method incorporates gradients of both signs, matrix multiplication could lead to ambiguous interpretations. To address this, we average the attribution map R_l across layers, preserving the interpretability of both positive and negative attributions.

$$R = \frac{1}{(L - \tilde{l} + 1)} \sum_{l=\tilde{l}}^{L} R_l, \tag{9}$$

where L is the total number of layers in the transformer model, $\tilde{l}$ is the index of the starting layer for accumulation, which is a hyperparameter.

Token Aggregation and F-CLIPScore. To identify misaligned words, we calculate the word-level attribution w_j by averaging the relevancy of its constituent tokens. We then predict a word as misaligned if its attribution falls below a threshold ϵ.

$$\text{mis}(w_j) = \begin{cases} 1, & \text{if } w_j < \epsilon \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

To get a global fine-grained misalignment score between images and text, similar to CLIPScore (Hessel et al. 2021), we devise a simple aggregation method to derive a single score, which is dubbed as F-CLIPScore, as follows:

$$\text{F-CLIPScore}(v,t) = (1 - \text{score}_{v,t}) \cdot \sum_j \text{mis}(w_j) \cdot w_j. \tag{11}$$

This aggregation integrates both overall semantic alignment and fine-grained misalignments for each token.

Benchmark	Source	Text / Image domain	Misalign Type	Num of Misaligns	Annotation Type	Dense Misalign	Global Misalign
FOIL	COCO caption	natural / natural	object	single	rule-based	accuracy	AP
nocaps-FOIL	nocaps	natural / natural	object	single	rule-based	accuracy	AP
HAT	COCO caption	generated / natural	various	multiple	human	accuracy	AP
SeeTRUE-Feedback	COCO-con	natural / natural	various	multiple	human	NLI score	-
	COCO-T2I	natural / generated	various	multiple	human	NLI score	-
	Drawbench	natural / generated	various	multiple	human	NLI score	-
	Pick-a-pic-con	generated / generated	various	multiple	human	NLI score	-
Rich-HF	Pick-a-pic	natural / generated	various	multiple	human	precision, recall, F1	correlation

Table 1: **Comprehensive overview of benchmarks for dense misalignment detection.** "Generated" in the Text / Image domain column indicates that the text or image was created by a captioning model or a text-to-image generation model, respectively. In contrast, "natural" signifies that the text or image originates from a human source.

Method	FPS	needs annotations	FOIL		nocaps-FOIL							
					Overall		In-Domain		Near-Domain		Out-of-Domain	
			LA	AP	LA	AP	LA	AP	LA	AP	LA	AP
CHAIR	-	✓	0.790	**0.925**	0.144	0.583	0.135	0.578	0.176	0.591	0.122	0.581
CLIPScore (ViT-B/32)	13.4		-	*0.707*	-	*0.692*	-	*0.651*	-	*0.675*	-	*0.743*
CLIPScore (ViT-H/14)	8.72		-	0.763	-	0.722	-	0.690	-	0.707	-	0.764
RefCLIPScore (ViT-B/32)	8.75	✓	-	*0.748*	-	*0.736*	-	*0.683*	-	*0.718*	-	*0.791*
ALOHa	0.16	✓	0.400	0.614	0.452	0.695	0.474	0.718	0.473	0.667	0.488	0.709
Ours (ViT-B/32)	12.0		0.732	0.714	0.603	0.690	0.547	0.673	0.597	0.684	0.632	0.713
Ours (ViT-H/14)	7.06		**0.836**	0.806	**0.716**	**0.794**	**0.661**	**0.789**	**0.708**	**0.793**	**0.748**	**0.802**

Table 2: **Experiment results on FOIL and nocaps-FOIL**. LA: Localization Accuracy. AP: Average Precision. FPS is measured on the nocaps-FOIL dataset. *Italic* denotes that we remeasured the result with ViT-B/32.

Experiments

As summarized in Table 1, we comprehensively evaluate our method across a diverse range of dense misalignment detection benchmarks. Our evaluation spans text domains (natural and generated), image domains (natural and generated), misalignment types (object, attribute, relation, and action), and the number of misaligned words (single or multiple). This extensive testing demonstrates the robustness and versatility of our approach. For detailed information about the datasets and experiments on additional benchmarks, please refer to the supplementary materials.

We report two variants of CLIP: OpenAI CLIP ViT-B/32 (Radford et al. 2021), following Hessel et al. (2021), and ViT-H/14 trained on LAION-2B (Schuhmann et al. 2022) from OpenClip (Cherti et al. 2023), which yields our best score. Further analysis of other backbones is provided in the supplementary material. We use a template "A photo depicts " following Hessel et al. (2021). We set our hyperparameters by searching the development set of Rich-HF and a subset of the training set from the FOIL dataset. We use $\tilde{l}$ to 10 and 22 for ViT-B/32 and ViT-H/14, respectively, utilizing the final three layers in both cases. Unless otherwise specified, ϵ is set to -0.00005. Frames-Per-Second (FPS) is measured with a single V100. Finally, we use F-CLIPScore for the global misalignment classification task.

Quantitative Results

FOIL and nocaps-FOIL. FOIL (Shekhar et al. 2017) and nocaps-FOIL (Petryk et al. 2024) are benchmarks for detecting misaligned captions where one object is replaced by a conceptually similar word (e.g., car, bicycle). We assess performance on two protocols: (1) localization accuracy (LA) for dense misalignment detection and (2) average precision (AP) for global misalignment classification. Following existing works, our approach predicts a single word with the lowest attribution w_j. In nocaps-FOIL, we report results as in-domain, near-domain, or out-of-domain based on how similar the altered objects are to COCO object classes.

In Table 2, our ViT-B/32 variant demonstrates state-of-the-art performance on most dense misalignment detection (LA). It is worth noting that baselines such as CHAIR (Rohrbach et al. 2018) or ALOHa (Petryk et al. 2024) make use of ground truth segmentation labels or reference captions. The ViT-H/14 variant demonstrates significantly enhanced performance, showing improved results consistently across all domains. It shows the robustness of our approach, which utilizes CLIP model pre-trained on various alt-text. Furthermore, our F-CLIPScore boosts up global misalignment classification (AP) by a significant margin, even surpassing reference-based methods. Lastly, our proposed approach demonstrates significantly superior computational efficiency compared to ALOHa, achieving a 44-fold reduction in inference time.

HAT. The HAT dataset (Petryk et al. 2024) comprises 400 human-annotated samples featuring captions generated by VLM models (Li et al. 2022; Wang et al. 2022; Chan et al. 2023; Zhu et al. 2024). For evaluation, we measured with the same metric as FOILs: LA and AP. For LA, correctly iden-

method	ref. captions	FPS	LA	AP
CHAIR	✓	-	0.067	0.369
CLIPScore		18.8	-	*0.385*
RefCLIPScore	✓	9.03	-	*0.429*
ALOHa	✓	0.24	0.203	**0.486**
Ours (ViT-B/32)		9.64	0.193	0.355
Ours (ViT-H/14)		6.56	**0.348**	0.360

Table 3: **Evaluation results on HAT test set.** ref. captions denotes that the method utilizes reference captions. *Italic* denotes that we remeasured the result with ViT-B/32.

Model	ft.	FPS	NLI score
LLaVa-1.5 (Vicuna-7B)		0.24	0.173
PaLI 5B		-	0.226
mPLUG-Owl (LLaMa-7B)		0.24	0.297
InstructBLIP (FlanT5$_{\text{XL}}$)		0.51	0.555
MiniGPT-v2 (LLaMa2-7B)		0.28	0.560
Ours (ViT-B/32)		7.90	0.605
Ours (ViT-H/14)		5.81	0.660
PaLI 5B	✓	-	0.765
PaLI 17B	✓	-	**0.785**

Table 4: **Evaluation results on SeeTRUE-Feedback test set.** ft. denotes that the model is fine-tuned.

tifying any hallucinated object in a sentence is considered accurate. To compare with ALOHa, which extracts a noun phrase, we concatenate neighboring misaligned words and average scores within a phrase. The phrase with the lowest aggregate score was predicted as the erroneous segment.

In Table 3, our ViT-H/14 variant demonstrates superior performance in LA with significantly improved FPS. In terms of AP, our method shows a performance gap compared to models that utilize reference captions. Further analysis for AP is presented in the supplementary.

SeeTRUE-Feedback. SeeTRUE-Feedback (Gordon et al. 2024) comprises a test set of 2K samples covering various images, text domains, and misalignment types. We specifically focus on textual misalignment detection, which aims to extract mismatched spans from caption. Following established protocols, we report the natural language inference (NLI) (Bowman et al. 2015) score obtained from a BART-NLI model (Lewis et al. 2020). We calculate the entailment score where the premise is the ground truth label, and the hypothesis is the predicted word span. To form a single sequence, we take the same strategy as the one used in evaluating the HAT dataset.

As shown in Table 4, our method surpasses the zero-shot models, showing its efficiency and robustness in various domains. It is also worth noting that, as a non-generative model, ours offers faster inference times compared to larger vision language models.

Rich-HF. Rich-HF (Liang et al. 2024) comprises 955 prompt and image pairs with word-level misalignment annotations and overall alignment score. Since prompts are collected by real users (Kirstain et al. 2023), its captions cover various lengths, styles, and contents. We evaluate the performance of misalignment labels using precision, recall, and F1 scores at the word level. We also measure Pearson and Spearman's correlation between our aggregated score and the Likert score for alignment. We further report the performance of ALOHa for comparison.

Model	ft.	F1	precision	recall
ALOHa		0.344	0.311	0.385
Ours (ViT-B/32)$_{\epsilon=-0.00001}$		0.398	0.328	0.504
Ours (ViT-H/14)$_{\epsilon=-0.00001}$		0.427	0.365	**0.516**
Ours (ViT-H/14)$_{\epsilon=-0.00005}$		0.314	0.487	0.231
Rich-HF (multi-head)	✓	0.433	**0.629**	0.330
Rich-HF (augmented prompt)	✓	**0.439**	0.613	0.341

Table 5: **Experiment results on Rich-HF test set**. ft. denotes that the model is fine-tuned with the Rich-HF training set.

Model	ft.	pearson	spearman
CLIPScore (ViT-B/32)		0.185	0.130
PickScore (ViT-H/14)		0.346	0.340
Ours (ViT-B/32)$_{\epsilon=-0.00001}$		0.279	0.332
Ours (ViT-H/14)$_{\epsilon=-0.00001}$		0.368	0.433
CLIPScore (ViT-B/32)	✓	0.398	0.390
Rich-HF (multi-head)	✓	**0.487**	**0.500**
Rich-HF (augmented prompt)	✓	0.474	0.496

Table 6: **Evaluation results on Rich-HF misalignment score correlation.** ft. denotes that the model is fine-tuned with the Rich-HF training set.

In Table 5, our method demonstrates promising performance as a zero-shot method. While precision is limited, it shows higher recall, resulting in a substantial F1 score. In Table 6, our aggregated score shows a significantly enhanced score in two correlation coefficients. Our ViT-H/14 variant even shows comparative performance with PickScore (Kirstain et al. 2023), which is finetuned with 583K human preference scores. These results suggest that the selected negative attributions effectively capture misalignment, leading to superior performance in measuring text-image discrepancies.

Qualitative Results

We present qualitative examples on three representative datasets in Figure 2. Further examples of all datasets are shown in the supplementary materials. For the FOIL and nocaps-FOIL datasets, models need to predict a single word regardless of the presence or absence of misaligned words. When misaligned words exist, our model detects them well for images from various domains. In cases where misaligned words do not exist, our model predicts unimportant word '.' and 'medium', a word that is difficult for the model to distinguish, as misaligned words as shown in the second and fourth images for the FOIL dataset. For the Rich-HF dataset, our model demonstrates decent misaligned word detection

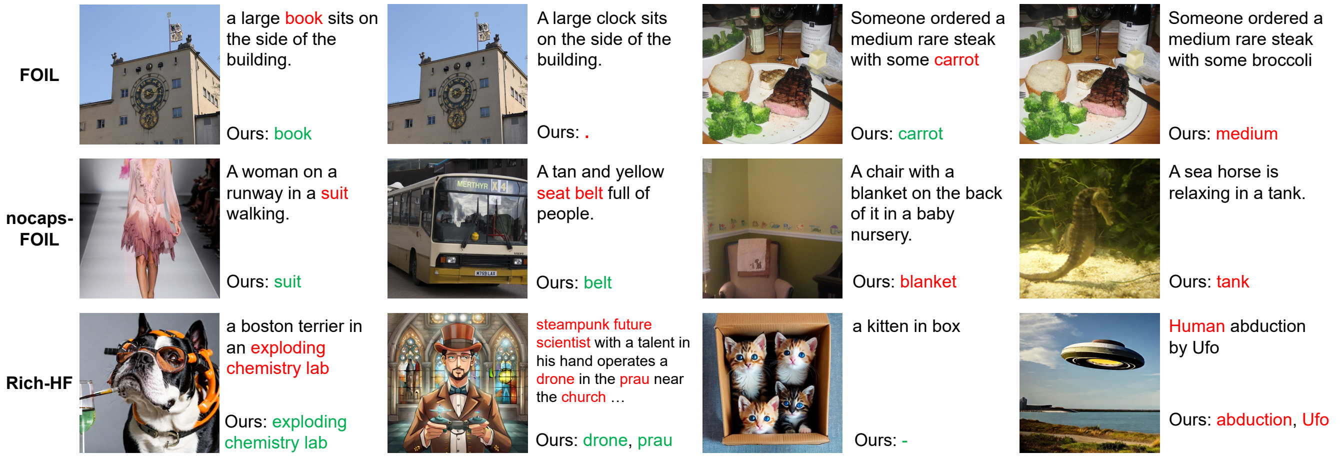

FOIL
a large book sits on the side of the building.
Ours: book
A large clock sits on the side of the building.
Ours: .
Someone ordered a medium rare steak with some carrot
Ours: carrot
Someone ordered a medium rare steak with some broccoli
Ours: medium
nocaps-FOIL
A woman on a runway in a suit walking.
Ours: suit
A tan and yellow seat belt full of people.
Ours: belt
A chair with a blanket on the back of it in a baby nursery.
Ours: blanket
A sea horse is relaxing in a tank.
Ours: tank
Rich-HF
a boston terrier in an exploding chemistry lab
Ours: exploding chemistry lab
steampunk future scientist with a talent in his hand operates a drone in the prau near the church …
Ours: drone, prau
a kitten in box
Ours: -
Human abduction by Ufo
Ours: abduction, Ufo

Figure 2: **Qualitative examples on FOIL, nocaps-FOIL, and Rich-HF datasets.** Misaligned words are highlighted in red in captions paired with images. Note that misaligned words may not exist. For predicted misaligned words, correct words are shown in green and incorrect words in red. If our model predicts that there are no misaligned words, it is indicated as '-'.

Method	FPS	LA	AP
occlusion-based	0.6	0.566	0.748
gradient-based			
∇A_l^h	5.8	0.423	0.741
$A_l^h \odot \nabla A_l^h$	5.8	**0.716**	**0.794**

Table 7: Ablation study of attribution calculation methods on the nocaps-FOIL test set.

ReLU($-\nabla A_l^h$)	ReLU($-\nabla A_l$)	LA	AP
✓		0.698	0.779
	✓	0.700	0.776
		0.716	**0.794**

Table 8: **Ablation study of the disabling positive gradients on nocaps-FOIL test set.** ReLU($-\nabla A_l^h$) and ReLU($-\nabla A_l$) indicate retaining only negative gradients before averaging across heads and layers, respectively.

performance for generated images. In addition, ours shows the ability to detect multiple misaligned words or not detect misaligned words when misaligned words do not exist.

Ablation Studies

To demonstrate the efficiency of the proposed method, we conduct ablation studies using the ViT-H/14 variant.

Attribution Calculation Method. We conduct an ablation study on the attribution calculation method. The occlusion (Goyal et al. 2016) method iteratively omits individual words from the input text and identifies the word whose removal leads to the highest increase in the score as the most likely erroneous element. Among gradient-based methods, we ablate components used in extracting relevance maps. Table 7 shows that the occlusion-based method demonstrated superior performance, but its efficiency was limited due to the requirement of multiple forward passes. In contrast, gradient-based methods, particularly when combined with attention maps, achieved a balance of high efficiency and performance.

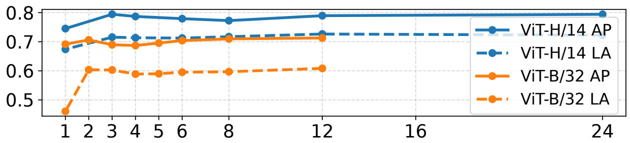


Figure 3: Ablation on the number of text encoder layers used for attribution calculation on nocaps-FOIL dataset.

Dataset	Method	AP	Pearson	Spearman
nocaps-FOIL	score$_{v,t}$	0.722	-	-
	$\sum_j \text{mis}(w_j) \cdot w_j$	0.776	-	-
	F-CLIPScore	**0.794**	-	-
Rich-HF	score$_{v,t}$	-	0.171	0.085
	$\sum_j \text{mis}(w_j) \cdot w_j$	-	0.352	0.419
	F-CLIPScore	-	**0.368**	**0.433**

Table 9: Ablation on components of F-CLIPScore.

Disabling Positive Gradient. We examine the effectiveness of using both positive and negative gradients in attribution calculation. We compare removing positive gradients before averaging across heads or layers, similar to conventional relevance map approaches (Selvaraju et al. 2017; Chefer, Gur, and Wolf 2021a). Table 8 demonstrates that utilizing full gradients yields the best performance, outperforming methods that isolate negative gradients. This finding underscores the importance of considering both positive and negative contributions in gradient-based relevance extraction.

Number of Layers. We perform an ablation study on $\tilde{l}$, the number of text encoder layers used for relevance map calculation. Figure 3 demonstrates that utilizing multiple layers, rather than solely the final layer, significantly enhances performance across both metrics. It demonstrates that utilizing intermediate features across layers enhances the detection of misalignments.

Components of F-CLIPScore. We conduct an ablation study on the components of F-CLIPScore using the nocaps-FOIL and Rich-HF datasets. As shown in Table 9, the mere summation of negative attribution w_j yields significantly improved AP and correlation coefficients. Moreover, integrating this with the overall similarity score further enhances performance, demonstrating our method's efficacy in capturing fine-grained misalignments. Additional analyses are presented in the supplementary material.

Analysis

Comparison with Baselines. In Figure 4, we present qualitative examples comparing our method to the baseline ALOHa (Petryk et al. 2024) on the HAT dataset. Our approach demonstrates robust and diverse detection capabilities, such as colors (e.g., white), numbers (e.g., two), entity-level objects (e.g., calf), and intangible objects (e.g., sunset), which can not be easily captured with combinations of foundation models. Different from ALOHa, which uses a language similarity module, CLIP, which is trained on diverse alt-text data, is sensitive to conceptually similar but visually distinct words (e.g., "blue" and "grey"). This underscores the effectiveness of our CLIP-based approach, which operates independently of additional foundation models. While showing promising results, our method also reveals some inherent limitations of CLIP, particularly in identifying discrepancies related to backgrounds (e.g., "wooden floor") or small objects (e.g., "birds"). Further examples and analyses are presented in the supplementary materials.

We provide qualitative examples comparing our method to the baseline MiniGPT-v2 (Chen et al. 2023) on SeeTRUE-Feedback dataset, as shown in Figure 5. As a large vision-language model, MiniGPT-v2 has the advantage of providing natural and rich responses. However, despite a prompt that requests the model to answer in short words, it provides lengthy and unformatted responses to almost all examples. Since it is quite difficult to accurately extract misaligned words from unstructured responses, its usability as a dense misalignment detector is low. As a non-generative model, our method achieves significantly higher FPS compared to MiniGPT-v2.

Part-of-Speech. We report Rich-HF word level metrics per part-of-speech (POS) for further analysis. In Table 10, our method predicts all overall POS, which shows that our method has the capability to predict misaligned words of various types, not limited to nouns. Still, we observe a trend that shows decent performance with nouns but limited performance with adverbs, adjectives, numbers, and adpositions. The result shows that our result corresponds with studies that reveal CLIP's weaknesses (Paiss et al. 2023; Nikolaus et al. 2022; Yuksekgonul et al. 2023). We leave it as future work to test with CLIP variants, which are further fine-tuned to tackle such shortcomings.

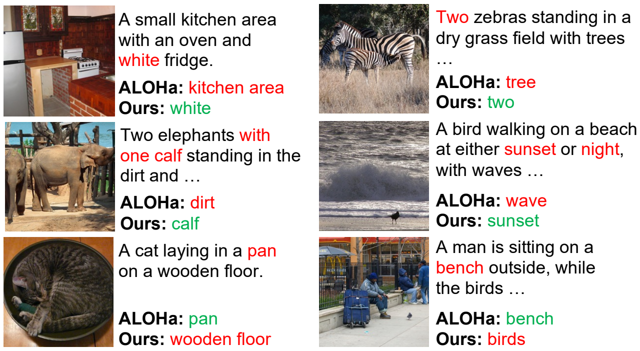

A small kitchen area with an oven and white fridge.
ALOHa: kitchen area
Ours: white
Two zebras standing in a dry grass field with trees …
ALOHa: tree
Ours: two
Two elephants with one calf standing in the dirt and …
ALOHa: dirt
Ours: calf
A bird walking on a beach at either sunset or night, with waves …
ALOHa: wave
Ours: sunset
A cat laying in a pan on a wooden floor.
ALOHa: pan
Ours: wooden floor
A man is sitting on a bench outside, while the birds …
ALOHa: bench
Ours: birds

Figure 4: **Qualitative examples compared to ALOHa on HAT dataset.** Our method demonstrates improved robustness in various misalignment types.

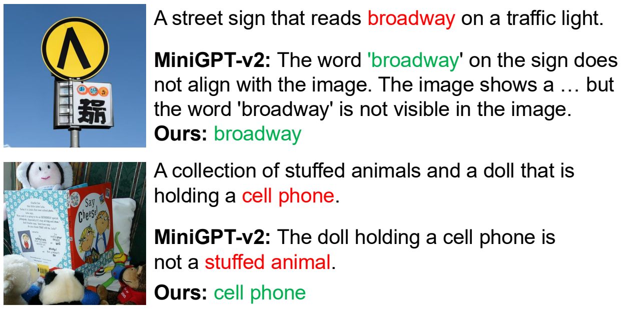

A street sign that reads broadway on a traffic light.
MiniGPT-v2: The word 'broadway' on the sign does not align with the image. The image shows a … but the word 'broadway' is not visible in the image.
Ours: broadway
A collection of stuffed animals and a doll that is holding a cell phone.
MiniGPT-v2: The doll holding a cell phone is not a stuffed animal.
Ours: cell phone

Figure 5: **Qualitative examples compared to MiniGPT-v2 on SeeTRUE-Feedback dataset.** Our method shows precise and concise misalignment detection ability while achieving significantly higher FPS.

	NOUN	PROPN	VERB	ADV	ADJ	NUM	ADP
F1	0.393	0.312	0.301	0.258	0.258	0.132	0.177
Precision	0.470	0.602	0.567	0.444	0.417	0.500	0.278
Recall	0.337	0.211	0.205	0.182	0.187	0.076	0.130

Table 10: Rich-HF test set results per part-of-speech.

Conclusion and Future Work

In this paper, we present a novel approach for detecting dense misalignments between text and image using a pre-trained CLIP model. By extracting attributions from CLIP's intermediate gradients, our method provides a scalable and efficient solution that achieves state-of-the-art performance in zero-shot settings and competitive results with fine-tuned models across multiple benchmarks. Also, our proposed F-CLIPScore shows enhanced performance to capture global misalignments. While showing effectiveness in capturing various misalignment types, our analysis reveals that our method inherits weaknesses observed in CLIP. Further examination is needed to improve the detection of misalignments using CLIP variants specifically trained to address these shortcomings.

Acknowledgements

We thank Byungju Han for the discussion and Donghyun Kim for proofreading. We also thank the Vision Understanding team for supporting our work and project namuniv.

References

Abnar, S.; and Zuidema, W. 2020. Quantifying Attention Flow in Transformers. In *ACL*, 4190–4197. Online: Association for Computational Linguistics.

Arras, L.; Montavon, G.; Müller, K.-R.; and Samek, W. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 159–168. Copenhagen, Denmark: ACL.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7): e0130140.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*. Association for Computational Linguistics.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229. Springer.

Chan, D.; Myers, A.; Vijayanarasimhan, S.; Ross, D.; and Canny, J. 2023. IC3: Image Captioning by Committee Consensus. In *EMNLP*, 8975–9003. Singapore: Association for Computational Linguistics.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Trans. Graph.*, 42(4).

Chefer, H.; Gur, S.; and Wolf, L. 2021a. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *ICCV*, 387–396. IEEE.

Chefer, H.; Gur, S.; and Wolf, L. 2021b. Transformer Interpretability Beyond Attention Visualization. In *CVPR*, 782–791. Computer Vision Foundation / IEEE.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2818–2829.

Cho, J.; Hu, Y.; Baldridge, J. M.; Garg, R.; Anderson, P.; Krishna, R.; Bansal, M.; Pont-Tuset, J.; and Wang, S. 2024. Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation. In *ICLR*.

Gordon, B.; Bitton, Y.; Shafir, Y.; Garg, R.; Chen, X.; Lischinski, D.; Cohen-Or, D.; and Szpektor, I. 2024. Mismatch quest: Visual and textual feedback for image-text misalignment. In *ECCV*, 310–328. Springer.

Goyal, Y.; Mohapatra, A.; Parikh, D.; and Batra, D. 2016. Towards transparent ai systems: Interpreting visual question answering models. In *ICML 2016 Workshop on Visualization for Deep Learning*.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *AAAI*, volume 38, 18135–18143.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 7514–7528. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 20406–20417.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-Pic: an open dataset of user preferences for text-to-image generation. In *NeurIPS*, 36652–36663.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*, 292–305.

Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; et al. 2024. Rich human feedback for text-to-image generation. In *CVPR*, 19401–19411.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *ECCV*, 38–55. Cham: Springer Nature Switzerland. ISBN 978-3-031-72970-6.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*, 4765–4774.

Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65: 211–222.

Nikolaus, M.; Salin, E.; Ayache, S.; Fourtassi, A.; and Favre, B. 2022. Do Vision-and-Language Transformers Learn

Grounded Predicate-Noun Dependencies? In *EMNLP*, 1538–1555.

Paiss, R.; Ephrat, A.; Tov, O.; Zada, S.; Mosseri, I.; Irani, M.; and Dekel, T. 2023. Teaching CLIP to Count to Ten. In *ICCV*, 3147–3157. IEEE.

Petryk, S.; Chan, D.; Kachinthaya, A.; Zou, H.; Canny, J.; Gonzalez, J.; and Darrell, T. 2024. ALOHa: A New Measure for Hallucination in Captioning Models. In *NAACL*, 342–357.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Rassin, R.; Ravfogel, S.; and Goldberg, Y. 2022. DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 335–345.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *SIGKDD*, 1135–1144. ACM.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *EMNLP*, 4035–4045. Brussels, Belgium: Association for Computational Linguistics.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C. W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*, 618–626. IEEE Computer Society.

Shekhar, R.; Pezzelle, S.; Klimovich, Y.; Herbelot, A.; Nabi, M.; Sangineto, E.; and Bernardi, R. 2017. FOIL it! Find One mismatch between Image and Language caption. In *ACL*, 255–265. Vancouver, Canada: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 23318–23340. PMLR.

Wang, Y.; Rudner, T. G.; and Wilson, A. G. 2023. Visual explanations of image-text representations via multi-modal information bottleneck attribution. In *NeurIPS*, volume 36, 16009–16027.

Xiao, W.; Huang, Z.; Gan, L.; He, W.; Li, H.; Yu, Z.; Jiang, H.; Wu, F.; and Zhu, L. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint*, arXiv:2404.14233.

Yan, S.; Bai, M.; Chen, W.; Zhou, X.; Huang, Q.; and Li, L. E. 2024. ViGoR: Improving Visual Grounding of Large Vision Language Models with Fine-Grained Reward Modeling. In *ECCV*, 37–53. Cham: Springer Nature Switzerland. ISBN 978-3-031-73030-6.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *TMLR*.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, 13807–13816.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*, 818–833. Springer.

Zhao, C.; Wang, K.; Zeng, X.; Zhao, R.; and Chan, A. B. 2024. Gradient-based visual explanation for transformer-based clip. In *ICML*, 61072–61091. PMLR.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *ECCV*, 696–712. Springer.

Zhu, D.; Chen, J.; Haydarov, K.; Shen, X.; Zhang, W.; and Elhoseiny, M. 2024. ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions. *TMLR*.