FineMatch: Aspect-based Fine-grained Image and Text Mismatch Detection and Correction

Hang Hua [1], Jing Shi [2], Kushal Kafle[2], Simon Jenni[2], Daoan Zhang[1], John Collomosse[2], Scott Cohen[2], and Jiebo Luo[1]

[1] University of Rochester
[2] Adobe Research
{hhua2, jluo}@cs.rochester.edu, dzhang52@ur.rochester.edu
{jingshi, kkafle, jenni, collomos, scohen}@adobe.com

Abstract. Recent progress in large-scale pre-training has led to the development of advanced vision-language models (VLMs) with remarkable proficiency in comprehending and generating multimodal content. Despite the impressive ability to perform complex reasoning for VLMs, current models often struggle to effectively and precisely capture the compositional information on both the image and text sides. To address this, we propose **FineMatch**, a new aspect-based fine-grained text and image matching benchmark, focusing on text and image mismatch detection and correction. This benchmark introduces a novel task for boosting and evaluating the VLMs' compositionality for aspect-based fine-grained text and image matching. In this task, models are required to identify mismatched aspect phrases within a caption, determine the aspect's class, and propose corrections for an image-text pair that may contain between 0 and 3 mismatches. To evaluate the models' performance on this new task, we propose a new evaluation metric named **ITM-IoU** for which our experiments show a high correlation to human evaluation. In addition, we also provide a comprehensive experimental analysis of existing mainstream VLMs, including fully supervised learning and in-context learning settings. We have found that models trained on FineMatch demonstrate enhanced proficiency in detecting fine-grained text and image mismatches. Moreover, models (e.g., GPT-4V, Gemini Pro Vision) with strong abilities to perform multimodal in-context learning are not as skilled at fine-grained compositional image and text matching analysis. With FineMatch, we are able to build a system for text-to-image generation hallucination detection and correction. Resources are available at https://hanghuacs.github.io/finematch/.

Keywords: Pre-trained Vision-Language Models · Aspect-based Image and Text Analysis · Compositionality

1 Introduction

Pretrained vision-language models, such as GPT-4V [1], LLaVA [51], MiniGPT-4 [7], and BLIP [22, 23], have demonstrated impressive ability to perform complex reasoning. Benefiting from large pretrained VLMs, a series of VLM-based

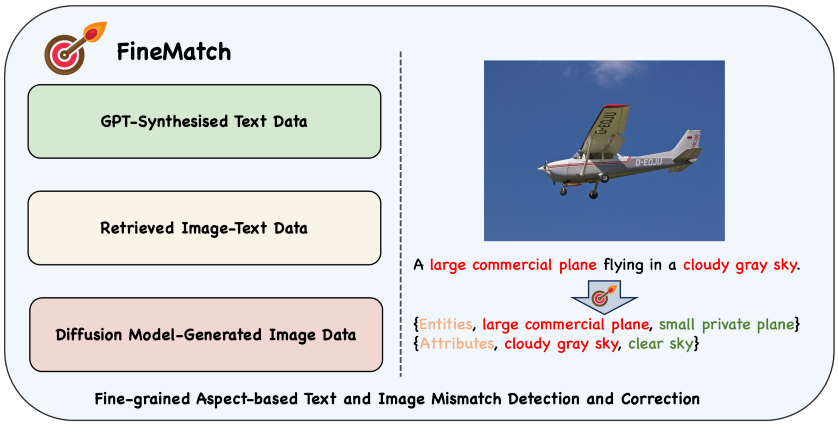

Fig. 1: Given a text and image pair, FineMatch enables VLMs to detect the **mismatched aspects** and the aspect classes in the caption and then give the corresponding **corrections**.

methods [14,17,24,54,60] have emerged and achieved remarkable results on various vision-language (VL) tasks. However, contemporary state-of-the-art VLMs still struggle with fine-grained compositional information understanding [19,61]. Prior works have pointed out that pretrained VLMs face challenges in comprehending fine-grained visual and textual compositional information [9, 61]. This issue poses a significant limitation to the reliability and performance of VLMs.

In recent years, there has been a growing focus on evaluating and improving the compositionality in large VL models [9,35,36]. Most of these approaches focus on constructing hard negative text-image pairs to evaluate the models' compositionality [13, 31, 50, 56, 61, 65]. These evaluations typically require models to identify the hard negative samples at the sentence level but ignore evaluating the model's capability to localize the mismatched phrases and provide the corresponding corrections. Nevertheless, evaluating the compositionality of pretrained VL models from the perspective of sentence level may seem almost too trivial a task [9]. Additionally, research into the ability to identify and rectify discrepancies between images and text has been overlooked. Based on this background, we propose **FineMatch**, a new challenging benchmark for boosting VLMs' ability to identify and address the fine-grained discrepancies between visual and textual data.

With FineMatch, we analyze and address the semantic discrepancies between visual and textual data from four aspects: **Entity**, **Relation**, **Attribute**, **Number**. And we provide examples for these four aspects in supplementary material. We build FineMatch data from both the image side and text side, aggregating data from multiple sources. The FineMatch benchmark comprises 49,906 high-quality, human-annotated image-text pairs, distributed as 43,906 in the training set, 1,000 in the validation set, and 5,000 in the test set. This data originates from various sources, including: **GPT-Synthesized Text Data**, **Retrieved Image and Text Data**, and **Diffusion Model-Generated Image**

Data. Each image-text pair encompasses a varying number of mismatched aspects, ranging from 0 to 3. The collection methods for each data source are described in Section 3.

To evaluate the ability to fine-grained text and image mismatch analysis of pretrained VLMs, we conduct experiments in both supervised learning and in-context learning settings. To verify the rationality of our proposed benchmark, we also provide the human performance results on the FineMatch test set.

Our main contributions are three-fold:

- We propose a novel task for aspect-based fine-grained image and text mismatch detection and correction. To support this endeavor, we have constructed a large-scale dataset **FineMatch** with human annotations tailored to the proposed task. We put forth a new evaluation metric, **ITM-IoU**, which evaluates model predictions against ground truth on both the character and semantic levels. The experimental results show a high correlation between the ITM-IoU and human evaluation.
- We evaluate various state-of-the-art pre-trained VL models on our proposed benchmark. The empirical results show that training on FineMatch can effectively improve the models' capability of identifying and rectifying text and image mismatches.
- With FineMatch, we are able to build a novel and simple self-correction text-to-image generation system, which can detect the detailed mismatch information between a generated image and text prompt and then automatically generate image editing instructions to edit the image to be semantically consistent with the text prompt. The generation examples indicate the system can effectively reduce hallucinations in text-to-image generation.

2 Related Work

2.1 Compositionality Evaluation

Compositional image and text understanding is a critical capability for VLMs. Research indicates that VLMs struggle with distinguishing the hard negative examples, i.e., image text pairs that mismatch in at least one aspect (e.g., attribute, relation), since they have little incentive to learn to encode compositionality during contrastive pretraining [61]. Moreover, finetuning with generated hard negative examples can improve the performance of language models [13]. In recent years, numerous benchmarks have been proposed to assess the capability of VLMs for fine-grained compositional vision and language reasoning. VL-CheckList [65] is an explainable framework that generates fine-grained and disentangled evaluation reports about VLMs. ARO [61] evaluate models for fine-grained relation, attribution, and order understanding. Winoground [50] is a task for visio-logic compositional reasoning. SUGARCREPE [13] aims to remove the artifact bias in model-synthesized visual compositional understanding evaluation benchmarks. Furthermore, there are several other benchmarks for compositionality evaluation, including SeeTRUE [56], CREPE [31], Cola [41], and T2I-

CompBench [18], etc. However, there are no VL compositionality benchmarks for aspect-based fine-grained text and image match detection and correction.

2.2 Pretrained Vision-Language Models

Large pretrained VL models such as OFA [53], BEiT-3 [55], CoCA [59], and mPlug [57] have successfully facilitated many cross-modal downstream tasks. CLIP [38] and its following works [23,33,44] learn to align image and text features through contrastive learning objectives on large-scale image-text pairs. BLIP [22, 23], LLaVA [28], and MiniGPT4 [7] show promising results by connecting vision encoders and LLMs through a compact intermediary model. These pretrained VL models with a stable fine-tuning strategy [15,16] can be easily adapted to a new downstream task. In addition, GPT-4V [34], Flamingo [3], and Emu2 [48] show strong abilities in zero-shot learning and multimodal in-context learning. Despite the remarkable achievements of large pretrained VL models, they struggle with capturing and understanding the fine-grained compositional information present in both text and images. The goal of this study is to provide a benchmark for evaluating and boosting the compositionality of pretrained VL models.

3 Aspect-based Fine-grained Image and Text Mismatch Analysis

3.1 Task Definition

FineMatch contains two subtasks: (1) **Mismatch Detection (MD)**; (2) **Mismatch Detection & Correction (MD&C)**. Let $\mathcal{D} = \{I_i, C_i, P_i\}_{i=1}^{|\mathcal{D}|}$ represents the dataset. (I_i, C_i) is the image and text pair, and $P_i = \{c_j, p_j, o_j\}_{j=1}^{M}$ is the mismatched aspects representation, where c_j is the aspect class, p_j is the aspect phrase that is extracted from the caption C_i, o_j is the corresponding correction, M is the number of mismatched aspect. Given an image-text pair (I_i, C_i), for the Mismatch Detection task, the models need to predict a set of tuples that contains the mismatched phrases in the captions and the corresponding class.

$$\textbf{MD}(I_i, C_i) = \{(c_j, p_j)\}_{j=1}^{M} \tag{1}$$

For the Mismatch Detection & Correction task, the models need to predict a set of triplets that contains the mismatched phrases, the class of the phrase, and the suggested corrections.

$$\textbf{MD\&C}(I_i, C_i) = \{(c_j, p_j, o_j)\}_{j=1}^{M} \tag{2}$$

FineMatch also contains the data for which the caption matches the image. In this case, the models' output should be *None*.

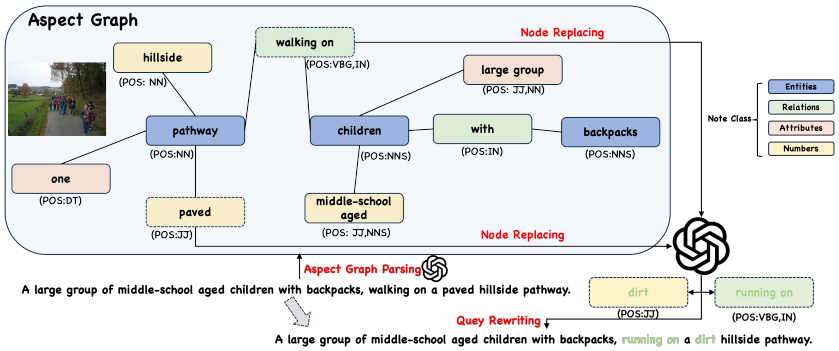

Fig. 2: Aspect graph parsing and node replacement for GPT-Synthesized text data.

3.2 GPT-Synthesized Text Data with Aspect Graph Parsing and Node Replacement

Data Generation For the GPT-Synthesised text data, we generate mismatched captions via aspect graph parsing and node replacement. To extract fine-grained compositional sub-phrases from the captions, we parse the captions into aspect graphs using In-Context Learning (ICL) with GPT-4. As depicted in Figure 2, an aspect graph consists of nodes representing aspect entities and edges illustrating the relationships between these entities, with each node being atomic. Then, we prompt GPT-4 to randomly replace the nodes with the counterfactual descriptions while maintaining the same Part of Speech (POS) tag with the initial nodes. Subsequently, the aspect graphs are translated back into new captions. This approach enables us to create mismatched captions without changing the structure of the initial captions. We filter the generated captions with the CLIP score, to remove any mismatched captions that are significantly incongruent with the images.

Data Debiasing Previous work [13] pointed out that the rule-based mismatched caption generation procedures may introduce two major types of undesirable artifacts: (1) nonsensible artifacts (irrational contents), and (2) non-fluent artifacts (grammar issues). Additionally, the significance of the semantic gap between mismatched captions and their associated images is another artifact bias for the generated data, making these captions easy to discern by models. To fix these biases and ensure the quality of the GPT-Synthesised text data, we first use the combined Vera score [29], grammar score [32], and CLIP score [38] to filter the data. This approach allows us to exclude examples that exhibit grammatical errors, contradict common sense, or present significant discrepancies between the image and caption. Subsequently, the filtered data is annotated by human experts. In the annotation, the workers are required to check and revise the GPT-synthesised captions and the corresponding mismatched aspects. The quality analysis indicates this aspect-graph parsing and node-replacing method

combined with human annotation can effectively reduce the artifact biases in the GPT-synthesised data.

3.3 Retrieved Image-Text Data

The semantic discrepancy between text queries and images is a common issue in text-to-image retrieval systems. We can utilize this property of text-to-image retrieval systems to obtain the mismatched text and image pairs. We select text queries with rich compositional structures from various datasets, including NoCaps [2] and WizViz [12]. To get the text queries with rich compositional information, we first parse each text query into a constituency tree using syntactic parsing tools SpaCy and then filter the queries according to the depth of the tree. Queries with a deeper constituency tree indicate the more complex syntax of the sentence, and these sentences contain more compositional information. We sample 10k diverse and complex queries for image retrieval. The retrieved images source include: LAION-400M [43], COYO-700M [5], and Smithsonian Open Access [45]. We first use the ViT-G/14 CLIP model with the weight from Open CLIP [20] to retrieve 10 candidates from the image datasets and then filter the images with the combination of aesthetic score, similarity score, and image size. We finally obtain 10K high-quality text and image pairs for human annotation.

3.4 Stable Diffusion Generated Image Data

The text query comes from T2I-CompBench [18], a benchmark for compositional text-to-image generation. All the text queries in the benchmark are meticulously designed in 3 categories (attribute binding, object relationships, and complex compositions). We use the text queries from the T2I-CompBench training set to prompt Stable Diffusion 2.1 [42] to generate images. We finally obtain 2.5K high-quality text and image pairs for human annotation.

3.5 Human Annotation

To standardize the labeling scheme across different data sources and ensure data quality while eliminating potentially harmful content and addressing ethical concerns, we employ a consistent annotation team composed of the same group of workers. The annotation interfaces of different sources of data are shown in Appendix Section 7.

3.6 Comparison of FineMatch with Previous Works

We summarize the novelty of our work compared with previous works. Our work introduces a novel task centered on aspect-based, fine-grained detection and correction of text and image mismatches. Previous research mainly focuses on evaluating the pre-trained models' ability to identify the hard negative examples [13, 31, 61] via retrieval accuracy and ignores evaluating models' ability

Table 1: Comparison of Different Datasets

Benchmark	# Images-Text Pair	Fine-grained Mismatch Detection & Correction	Human Annotation	Multiple Source/Domain
ARO [61]	28,748	✗	✗	✗
SUGARCREPE [13]	7512	✗	✔	✗
Winoground [50]	800	✗	✔	✗
CREPE [31]	370,000	✗	✗	✔
VL-Checklist [65]	410,000	✗	✗	✔
SeeTrue [56]	31,855	✗	✔	✔
FineMatch	49,906	✔	✔	✔

to detect which part of the text is mismatched with the image and correct the mismatched aspects. Moreover, earlier studies, such as ARO [61], predefined the instance classes and collected images with 48 relations and 117 attribute pairs, thereby constraining the diversity of instances. In contrast, our proposed benchmark achieves the open set in text and image mismatch detection. Furthermore, previous works that perform fine-grained text and image mismatch detections depend on the VQA-based method, which needs to generate a list of questions first and then employ the VQA model to answer the questions. These methods, however, are not flexible and may suffer from bias accumulation issues. Our method trains models to generate the mismatched aspect phrase and the correct aspect phrase in an end-to-end manner. In Table 1, we also compare the difference of FineMatch with other related works from the perspective of the size of the dataset, whether fine-grained IMT detection and correction is supported, and if human annotation is employed to improve the data quality and remove harmful content.

3.7 Evaluation

To evaluate models' performance on FineMatch, we propose a novel metric called ITM-IoU. IoU (Intersection over Union) is a standard metric in computer vision for measuring the accuracy of object detection or segmentation, based on the overlap between predicted and ground truth boundaries or pixels. In this study, since each caption may contain multiple mismatched aspects, we compute the IoU between the models' predicted set of aspect triplets and the set of ground truth triplets. For triplets matching, we draw inspiration from the generic structured prediction evaluation methods in the NLP field [30]. To evaluate the accuracy of the predicted mismatched aspect classes, we adopt the exact match (EM) [39] metrics. For mismatched aspect phrase detection, we evaluate from both the character level and semantic level. For the lexical similarity evaluation, we use chrF [37], an F1-score for character n-gram matches (we use the default setting of n = 6). For the semantic level evaluation, we use the BERT score [63]. Given a predicted mismatched aspect representation $P_i = \{c_j, p_j, o_j\}_{j=1}^{M}$ $(i \in \{1, 2, ..., |\mathcal{D}|\})$ and the corresponding ground truth $G_i = \{c'_j, p'_j, o'_j\}_{j=1}^{M'}$ (M' is the number of ground truth mismatched aspects), the combined detection score $Score_{Dj}$ is calculated as:

$$Score_{Dj} = \frac{BERTScore(p_j, p'_j) + chrF(p_j, p'_j)}{2}, \tag{3}$$

As the aspect phrase correction is an open-ended generation task, we calculate the BERT score to evaluate the semantic similarity of the generated corrections o_j and the ground truth o'_j. The correction score $Score_{Cj}$ is calculated as:

$$Score_{Cj} = BERTScore(o_j, o'_j), \tag{4}$$

The total score of a predicted aspect is the weighted sum of the three elements' scores in the mismatched aspect representation:

$$Score_{Aspect_j} = W_{Ca} \cdot EM(c_j, c'_j) + \quad W_{De} \cdot Score_{Dj} + W_{Co} \cdot Score_{Cj} \tag{5}$$

where the W_{Ca}, W_{De}, and W_{Co} is the weight of the EM, $Score_{Dj}$, and $Score_{Cj}$, respectively. In this study, we set the weights $W_{Ca} = 0.2$, $W_{De} = 0.4$, and $W_{Co} = 0.4$. To compute the IoU of the aspect representation, we set a threshold T to match the predictions with ground truth, if $\max(\{Score_{Aspect_k}\}_{k=1}^{M'}) \geq T$ then the predicted triplet matches the ground truth. For each aspect representation prediction, we compute the final score as:

$$Score_{Aspect_j} = \begin{cases} Score_{Aspect_k} & \text{if} \quad \max(\{Score_{Aspect_k}\}_{k=1}^{M'}) \geq T, \\ 0 & \text{else} \end{cases} \tag{6}$$

Then the final calculation of the mismatched aspect ITM-IoU is calculated as:

$$\text{ITM-IoU} = \frac{\sum_{j=0}^{M} Score_{Aspect_j}}{M} \times \frac{|P_i \cap G_i|}{|P_i \cup G_i|}, \tag{7}$$

where $|P_i \cap G_i|$ denotes the number of matched triplets of data i, and $|P_i \cup G_i| = |P_i| + |G_i| - |P_i \cap G_i|$. We also provide the pseudocode in supplementary material for the ITM-IoU calculation.

A common way to show the goodness of an evaluation metric is to show its correlation with human evaluations. We conduct a human evaluation for the experimental results in Section 4.3, the results indicate the high correlation of ITM-IoU with human evaluation, which reflects the rationality of our proposed evaluation metrics.

3.8 Quantitative Analysis

Distribution of Data Source and Domain We provide the text and image distribution analysis from different perspectives. Figure 3 shows the data source distribution in the inner circle and domain distribution in the otter circle of FineMatch training and test set.

Distribution of Mismatched Aspects Figure 4 shows the number of mismatched aspects distribution across different data sources in FineMatch.

Distribution of Aspects Classes Figure 5 shows the distribution of aspect classes across different data sources in FineMatch.

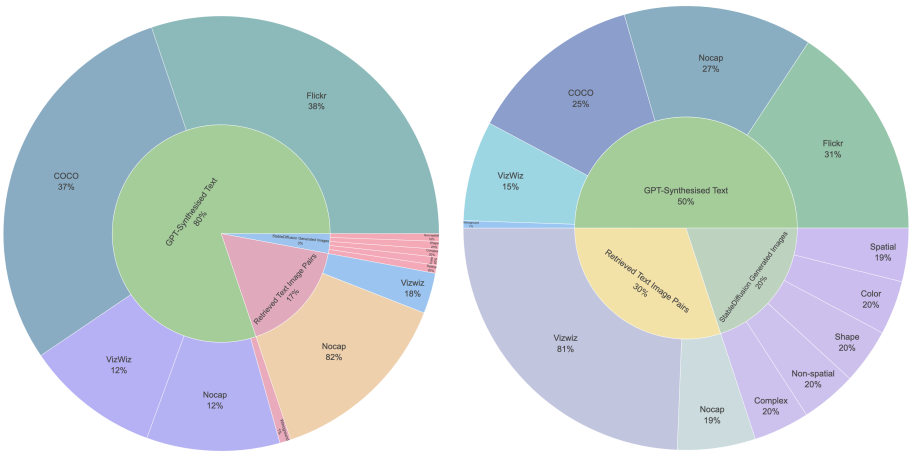

Fig. 3: The initial data source distribution (inner circle) and domain distribution (outer circle) for the FineMatch training set (left) and test set (right).

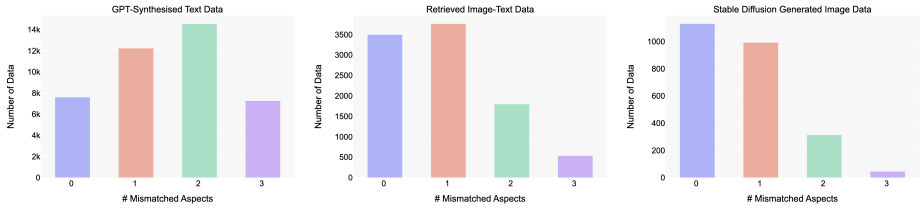

Fig. 4: Data distribution of varying numbers of mismatched aspects across different data sources in FineMatch.

Fig. 5: Data distribution of the mismatched aspect classes across the training, validation, and test sets in FineMatch.

3.9 Qualitative Analysis

We present the analysis of human-annotated and GPT-Synthesized data, focusing on the Vera Score Gap, Grammar Score Gap, and CLIP Score Gap. These metrics compare the quality changes of human-annotated mismatched captions against the original captions. Previous research [13] analyzes the artifact bias through the Vera Score Gap and Grammar Score Gap, and employs the adversarial refinement to fix this bias. In this research, we utilize human annotation to fix this bias and add the CLIP Score Gap to quantify the semantic changes of the human-annotated mismatched captions. The score gap is typically calculated as $S^P - S^N$, where the S^P and S^N are the Vera/Grammar/CLIP scores of the initial captions and the human-annotated GPT-Synthesised data. The findings, illustrated in Figure 6, reveal that score gap distribution lies on the positive spectrum, indicating that hard negative samples in the GPT-Synthesized data exhibit a higher likelihood of being nonsensical. The results also reflect a marginally reduced similarity to the initial matching captions. While these effects are within an acceptable range. Moreover, GPT rewrites do not significantly impact grammar fluency or introduce errors compared to the original captions. We also provide more quality analysis from different perspectives about other data sources in Appendix Section 2.

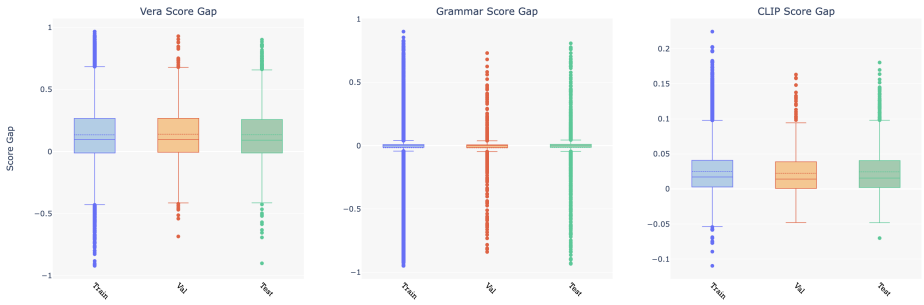

Fig. 6: Score gap distribution in FineMatch. Since we use human annotation to replace the adversarial refinement, the score gap distribution lies on the positive spectrum, but these effects are within an acceptable range.

4 Experiments

4.1 Visual Instruction Tuning

We train the models on FineMatch in an image-to-text generation setting. Given an image I_i and the corresponding caption C_i, and the output aspect representations $P_i = \{c_j, p_j, o_j\}_{j=1}^{M}$, the training objective is:

$$\mathcal{L} = -\sum_{\mathcal{D}} \sum_{t=1}^{M} \log p(P_t \mid [C_i : I_i], P_{\leq t-1}).$$

We conduct experiments on various state-of-the-art pretrained VL models including OFA [53], [57], InternLM-Xcomposer2-VL [10] LLaMA-Adapter2 [11], MiniGPT-4 [7], ShareGPT4V [8], and LLaVA series [28]. The experiment results are shown in Table 2.

Table 2: Visual instruction tuning performance (ITM-IoU) of different VL models on the FineMatch test set.

Models	Size	Mismatch Detection ↑	Mismatch Detection&Correction ↑
OFA-Large [53]	472M	19.72	21.35
LLaMA-Adapter2 [11]	7B	35.84	40.76
mPLUG-Owl2 [57]	8.2B	46.70	48.28
MiniGPT-4-V2 [7]	7B	51.18	55.95
InternLM-Xcomposer2-VL [10]	7B	58.70	61.07
LLaVA-1.5-LoRA [28]	7B	62.18	63.80
LLaVA-1.5 [28]	7B	62.25	63.62
LLaVA-1.5-LoRA [28]	13B	65.51	66.73
LLaVA-1.5 [28]	13B	66.02	67.13
ShareGPT4V [8]	13B	66.06	67.21
LLaVA-1.6-Vicuna [27]	13B	66.10	67.31
Human Performance	-	**88.32**	**89.19**

From the results presented in Table 2, it is evident that VLMs integrated with larger language models (LLaVA-1.5 7B vs. LLaVA-1.5 13B) exhibit superior performance on FineMatch. Language models that have been pretrained on more data or those that have been finetuned with carefully designed data (e.g., ShareGPT4V vs. LLaVA-1.5) tend to achieve enhanced performances. Furthermore, models with improved reasoning capabilities, image encoder, and world knowledge (LLaVA-1.6 vs. LLaVA-1.5) also demonstrate performance improvements.

4.2 In-Context Learning

In-Context Learning (ICL) explores training-free few-shot learning, where models are encouraged to "learn to learn" from limited tasks and generalize to unseen tasks. In this study, we conduct experiments on public accessible pretrained VL models including MMICL [64], Otter [21], OpenFlamingo [4], Emu2 [48], Gemini Pro Vision [49], and GPT4-V [34]. The experiment results are shown in Table 3. Since GPT-4V and Gemini cannot process some specific contents in the image of FineMatch test set, which will be judged as illegal input, we discard these examples, and we obtain 4278/5000 for GPT-4V and 4824/5000 for Gemini Pro to calculate the ITM-IoU.

It can be summarized from Table 3 that even the models such as GPT-4V and Gemini Pro Vision, which have a strong ability of multimodal in-context learning, are not as skilled at fine-grained compositional image and text matching analysis as we might have expected. Compared with the results in Table 2, models trained on FineMatch demonstrate enhanced proficiency in detecting fine-grained text and image mismatches.

Table 3: In-context learning results (in terms of ITM-IoU) of different VL models on the FineMatch test set. (* indicates subset of the FineMatch test set).

Models	Size	Mismatch Detection ↑	Mismatch Detection&Correction ↑
Otter [21]	7B	0.03	0.09
MMICL [64]	7B	0.11	0.25
OpenFlamingo [4]	9B	0.34	0.96
Emu2 [48]	37B	6.10	11.23
Gemini Pro Vision* [49]	-	9.07	11.14
GPT-4V* [34]	-	**21.92**	**21.58**

4.3 Human Evaluation

To evaluate the rationality of our proposed ITM-IoU, we carried out a human evaluation of model predictions on the sampled FineMatch test set. The sample size is 500. The human annotators are required to rate the quality of the generated mismatched aspect representations, the score ranges from 1-5, and the higher the better. The findings, presented in Table 4, demonstrate a high correlation between human evaluation results and the automatic evaluation metric ITM-IoU detailed in Table 2.

Table 4: Average human evaluation scores (ranging form 1-5) for fully supervised-learning (the upper row) and in-context learning methods (the lower row).

Model	Mismatch Detection ↑	Mismatch Detection&Correction ↑
mPLUG-Owl2	3.56	3.41
MiniGPT-4	3.77	3.65
LLaVA-1.5-7B	4.02	3.89
LLaVA-1.5-13B	4.41	4.15
OpenFlamingo	2.01	1.63
Emu2	2.55	2.32
Gemini Pro Vision	2.95	2.86
GPT4-V	3.35	3.27

4.4 Failure Cases of Language Models on FineMatch

In Table 2 and Table 3, we show the performance (in terms of ITM-IoU) of current state-of-the-art language models on FineMatch. The results indicate that the models still have difficulties in identifying the mismatched aspects of the given data. We also show examples of the GPT4-V's predictions, the finetuned LLaVA-1.6 predictions, and the ground truth in Figure 7 to illustrate that fine-grained image-text matching is a challenging task for VLMs.

5 FineMatch for Text-to-Image Generation Hallucination Detection and Correction

Hallucination issues are prevalent in text-to-image (T2I) generation models, particularly for the text prompts that describe detailed scenes and intricate relationships [26, 40]. Many approaches are proposed for reducing hallucination for

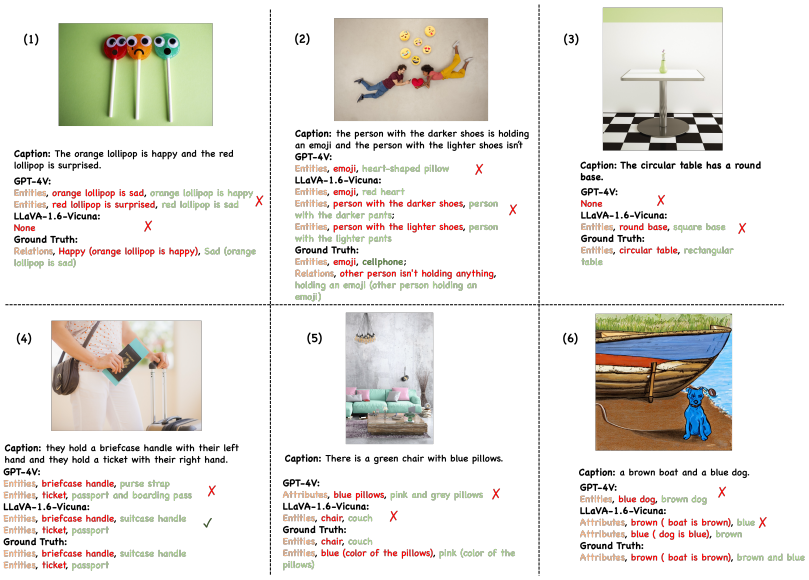

Fig. 7: Failure cases ({aspect classes,**mismatched aspects**,corrections}) of GPT-4V and LLaVA-1.6 on FineMatch.

T2I models [6, 25, 46, 47, 52, 58]. In this study, we illustrate how FineMatch can help detect and address the mismatch between the text prompts and the generated images for T2I generation models.

To achieve this goal, we develop a framework named AutoAlign that incorporates FineMatch with T2I models and VLMs for T2I automatic hallucination detection and correction. This system comprises four key modules: the T2I generation module, the hallucination detection module, the image editing prompt generation module, and the image editing module. Given a text prompt T and its corresponding generated image I, the text-image pair is evaluated by a VLM (LLaVA-1.6 Vicuna) fine-tuned with FineMatch to identify if there are any discrepancies between I and T. Upon detecting a mismatch, the system invokes the image editing prompt generation module (utilizing GPT-4 for this purpose) to generate an image editing prompt. Subsequently, the image editing module (employing MagicBrush [62] for image editing) is engaged to adjust the image. This process of hallucination detection, generation of editing prompts, and image editing is iteratively performed until the edited image satisfactorily aligns with the text prompt.

The architecture of AutoAlign is shown in Figure 8, as we can see from the diagram that the system is designed to effectively reduce the hallucination for T2I generation. More cases are shown in Appendix 6.

6 Limitation and Future Work

This study has a few limitations that present opportunities for further research. First, we provide only one possible correction for each mismatched aspect in

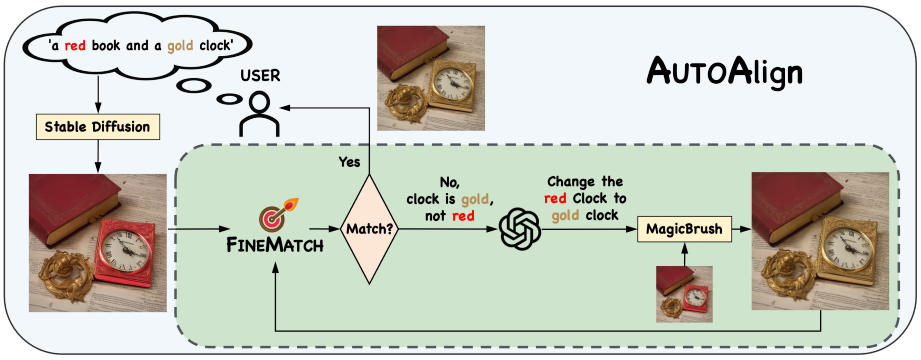

Fig. 8: Pipeline of the AutoAlign system: The modules within the system work collaboratively in a loop process until the edited image achieves satisfactory alignment with the text prompt.

the captions. Nevertheless, each mismatched aspect can have several viable corrections. For instance, if an image depicting a woman with blond hair, while the caption is "a woman holding a golden flower", then the correction could be "a golden flower" -> "no golden flower", or it can be "holding"->" with" and "a golden flower"->"blond hair". Recognizing the potential for multiple corrections, we treat the task of mismatched aspect correction as one of open-ended generation. We employ the BERT Score to evaluate the semantic similarity for models' generated corrections and the ground truth. In addition, introducing human evaluation to check if the generated corrections address the mismatch between the image and caption helps evaluate the quality of the predicted corrections of models. Second, finetuning VLMs directly with FineMatch has yet to yield results on par with human performance. In the future, we can explore designing better instruction following data for VLMs using FineMatch, such as introducing aspect graphs in the text prompt.

7 Conclusion

In this paper, we present FineMatch, a novel benchmark designed to address aspect-based, fine-grained mismatches between text and images. We design an aspect graph parsing and node-replacing method combined with human annotation to effectively reduce the artifact biases in the GPT-synthesised data. We also present comprehensive experiments on current state-of-the-art VLMs and found that FineMatch can help enhance the ability of the models to perform detailed analyses of text and image mismatches. In addition, we evaluate the current black-box models with strong multimodal in-context-learning capability and find that these models are not skilled at addressing fine-grained mismatches between text and images. With FineMatch, we build a text-to-image generation hallucination detection and correction system, and the system can effectively reduce the hallucination for T2I generation. We believe our efforts will benefit real-world applications involving text and image compositional analysis and generation.

References

1. Gpt-4v(ision) system card (2023), https://api.semanticscholar.org/CorpusID:263218031 1
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. International Conference on Computer Vision pp. 8947–8956 (2019), https://api.semanticscholar.org/CorpusID:56517630 6
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 4
4. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023) 11, 12
5. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022) 6
6. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. Advances in Neural Information Processing Systems **36** (2024) 13
7. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 1, 4, 11
8. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023) 11
9. Diwan, A., Berry, L., Choi, E., Harwath, D., Mahowald, K.: Why is winoground hard? investigating failures in visuolinguistic compositionality. arXiv preprint arXiv:2211.00768 (2022) 2
10. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., Zhang, W., Li, Y., Yan, H., Gao, Y., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420 (2024) 11
11. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023) 11
12. Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N.: Captioning images taken by people who are blind. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. pp. 417–434. Springer (2020) 6
13. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. arXiv preprint arXiv:2306.14610 (2023) 2, 3, 5, 6, 7, 10
14. Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N.A., Luo, J.: Promptcap: Prompt-guided task-aware image captioning. arXiv preprint arXiv:2211.09699 (2022) 2
15. Hua, H., Li, X., Dou, D., Xu, C.Z., Luo, J.: Noise stability regularization for improving bert fine-tuning. arXiv preprint arXiv:2107.04835 (2021) 4

16. Hua, H., Li, X., Dou, D., Xu, C.Z., Luo, J.: Fine-tuning pre-trained language models with noise stability regularization. arXiv preprint arXiv:2206.05658 (2022) 4
17. Hua, H., Tang, Y., Xu, C., Luo, J.: V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. arXiv preprint arXiv:2404.12353 (2024) 2
18. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. ArXiv **abs/2307.06350** (2023), https://api.semanticscholar.org/CorpusID:259847295 4, 6
19. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2
20. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773, if you use this software, please cite it as below. 6
21. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 11, 12
22. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 1, 4
23. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 1, 4
24. Lin, J., Hua, H., Chen, M., Li, Y., Hsiao, J., Ho, C., Luo, J.: Videoxum: Cross-modal visual and textural summarization of videos. arXiv preprint arXiv:2303.12060 (2023) 2
25. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023) 13
26. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024) 12
27. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/ 11
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. ArXiv **abs/2304.08485** (2023), https://api.semanticscholar.org/CorpusID:258179774 4, 11
29. Liu, J., Wang, W., Wang, D., Smith, N.A., Choi, Y., Hajishirzi, H.: Vera: A general-purpose plausibility estimation model for commonsense statements. arXiv preprint arXiv:2305.03695 (2023) 5
30. Lu, X.H., Kasner, Z., Reddy, S.: Weblinx: Real-world website navigation with multi-turn dialogue (2024), https://api.semanticscholar.org/CorpusID:267547883 7
31. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023) 2, 3, 6, 7

32. Morris, J.X., Lifland, E., Yoo, J.Y., Grigsby, J., Jin, D., Qi, Y.: Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. arXiv preprint arXiv:2005.05909 (2020) 5
33. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European Conference on Computer Vision. pp. 529–544. Springer (2022) 4
34. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023), https://api.semanticscholar.org/CorpusID:257532815 4, 11, 12
35. Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., Gatt, A.: Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. arXiv preprint arXiv:2112.07566 (2021) 2
36. Parcalabescu, L., Gatt, A., Frank, A., Calixto, I.: Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. arXiv preprint arXiv:2012.12352 (2020) 2
37. Popovic, M.: chrf: character n-gram f-score for automatic mt evaluation. In: WMT@EMNLP (2015), https://api.semanticscholar.org/CorpusID:15349458 7
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4, 5
39. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016) 7
40. Rawte, V., Sheth, A., Das, A.: A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922 (2023) 12
41. Ray, A., Radenovic, F., Dubey, A., Plummer, B.A., Krishna, R., Saenko, K.: Cola: A benchmark for compositional text-to-image retrieval (2023), https://api.semanticscholar.org/CorpusID:258546995 3
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022) 6
43. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. ArXiv **abs/2210.08402** (2022), https://api.semanticscholar.org/CorpusID:252917726 6
44. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022) 4
45. Smithsonian Institution: Smithsonian Open Access (2023), https://www.si.edu/openaccess 6
46. Song, L., Yin, G., Jin, Z., Dong, X., Xu, C.: Emotional listener portrait: Realistic listener motion simulation in conversation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20782–20792. IEEE (2023) 13
47. Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Aliaga, D.: Objectstitch: Generative object compositing. arXiv preprint arXiv:2212.00932 (2022) 13

48. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., et al.: Generative multimodal models are in-context learners. arXiv preprint arXiv:2312.13286 (2023) 4, 11, 12
49. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 11, 12
50. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022) 2, 3, 7
51. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023), https://api.semanticscholar.org/CorpusID:257219404 1
52. Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al.: Vigc: Visual instruction generation and correction. arXiv preprint arXiv:2308.12714 (2023) 13
53. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022) 4, 11
54. Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., Li, Z., Gao, M., Zhao, S., Shan, Y., et al.: Caption anything: Interactive image description with diverse multimodal controls. arXiv preprint arXiv:2305.02677 (2023) 2
55. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 (2022) 4
56. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., Szpektor, I.: What you see is what you read? improving text-image alignment evaluation. arXiv preprint arXiv:2305.10400 (2023) 2, 3, 7
57. Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv preprint arXiv:2311.04257 (2023) 4, 11
58. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. arXiv preprint arXiv:2310.16045 (2023) 13
59. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022) 4
60. Yu, Y., Zeng, Z., Hua, H., Fu, J., Luo, J.: Promptfix: You prompt and we fix the photo. arXiv preprint arXiv:2405.16785 (2024) 2
61. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2022) 2, 3, 6, 7
62. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36** (2024) 13

63. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019) 7
64. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., Chang, B.: Mmicl: Empowering vision-language model with multi-modal in-context learning. arXiv preprint arXiv:2309.07915 (2023) 11, 12
65. Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., Yin, J.: Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. arXiv preprint arXiv:2207.00221 (2022) 2, 3, 7