Mismatch Quest: Visual and Textual Feedback for Image-Text Misalignment

Brian Gordon*[1,2], Yonatan Bitton*[2], Yonatan Shafir[1,2], Roopal Garg[2], Xi Chen[2], Dani Lischinski[2,3], Daniel Cohen-Or[1,2], and Idan Szpektor[2]

Tel Aviv University[1], Google Research[2]
The Hebrew University of Jerusalem[3]

Abstract. While existing image-text alignment models reach high quality binary assessments, they fall short of pinpointing the exact source of misalignment. In this paper, we present a method to provide detailed textual and visual explanation of detected misalignments between text-image pairs. We leverage large language models and visual grounding models to automatically construct a training set that holds plausible misaligned captions for a given image and corresponding textual explanations and visual indicators. We also publish a new human curated test set comprising ground-truth textual and visual misalignment annotations. Empirical results show that fine-tuning vision language models on our training set enables them to articulate misalignments and visually indicate them within images, outperforming strong baselines both on the binary alignment classification and the explanation generation tasks. Our code and human curated test set are available at: `https://github.com/MismatchQuest/MismatchQuest`.

1 Introduction

Recently, text/image generative models [6, 12, 14, 26, 57, 63, 65, 78] achieved remarkable capabilities. However, they still often generate outputs that are not semantically-aligned to the input, both for text-to-image (T2I) and image captioning [40, 44]. They especially struggle with complex, nuanced, or out-of-distribution descriptions and fail to generate images which follow the prompt precisely [7, 60]. As long as alignment quality is insufficient, adoption of Vision-Language Models (VLMs) may be limited.

To automatically gauge the alignment performance of VLMs, alignment evaluation models were proposed [23, 71, 75]. These models provide binary classification scores for text/image pairs. However, they do not offer insights regarding the misalignment: *explanations* that could improve the understanding of VLM limitations and direct the training of better models. To bridge this gap, we propose that alignment models should not only predict misalignments but also elucidate the specifics of text-image misalignments via both textual explanations and visual feedback using bounding boxes, as demonstrated in Figure 1 and Figure 2. We hypothesize that this novel form of feedback would deepen the understanding of misalignment causes within text/image pairs and facilitates the improvement of generative models.

* Equal contribution

Fig. 1: Our alignment model steps: (1) the model predicts the alignment label between the input image/text pairs; (2) for misalignment labels, it then generates textual and visual feedback.

To this end, we introduce *ConGen-Feedback*, a method that, for an aligned image/textual-caption pair, generates plausible contradicting captions on aspects such as entities, actions, attributes, and relationships, together with corresponding textual and visual (bounding-box) explanations of the misalignments (see Figure 3). This is done by employing the capabilities of large language models (LLMs) and visual grounding models. The outcome training set, denoted **T**extual and **V**isual (TV) **Feedback**, is a comprehensive compilation of 3 million instances, crafted to simulate a wide array of text-image scenarios from diverse databases including COCO [37], Flickr30K [51], PickaPic [31], ImageReward [71], ADE20K [82, 83], and OpenImages [32]. We train an alignment evaluation model with this training set to both predict the alignment label and to generate feedback for misaligned image/text pairs.

To evaluate our alignment model, we construct and publish *SeeTRUE-Feedback*, a human-annotated test set. Human annotators provide textual explanations and approve visual bounding boxes to delineate misalignments, derived from a mixture of real and synthetic images and texts. Our model outperforms other baselines across all metrics: including 10% increase in alignment *Accuracy*, 20% increase in *Entailment* w.r.t gold (human-annotated) textual feedback, and a 2-13% increase in *F1* for visual feedback. *SeeTRUE-Feedback* will be publicly available on our project page. We complement our automated metrics with human ratings through an annotation study on Amazon Mechanical Turk [17], where our model outperforms the competing models by more than 100% improvements on all metrics. Our model also shows strong generalization capabilities with out-of-distribution images and prompts from various advanced T2I models such as Stable Diffusion (SD) v2.1 [63], SD XL [52], Composable Diffusion [41], and Adobe Firefly [2]. Finally, our ablation studies verify the advantage of our multi-task training, a single model generating both misalignment labels and feedback

Fig. 2: Qualitative analysis of out-of-distribution results: Showcasing image-text pairs generated by Stable-Diffusion XL [52], Stable-Diffusion 2.1 [63], Adobe Firefly [2] and Composable Diffusion [41] (credits to [7]) text-to-image models alongside the corresponding textual and visual feedback as predicted by the PaLI-X model finetuned on TV-Feedback

for different prompts, compared to training individual models for each task, as well as the effectiveness of our training set filtering strategy. We aim to encourage future works based on the presented methodology for various cross-domain applications, such as enhancing text-to-image processes by providing a feedback signal. Furthermore, it can be utilized to identify and correct incorrect annotations in text-image pairs datasets and refine image captioning models by detecting erroneous captions. In sum, our contributions are: (a) a feedback-centric data generation method (ConGen-Feedback); (b) a comprehensive training set (TV-Feedback); (c) a human-annotated evaluation set (SeeTRUE-Feedback), which we also make publicly available; (d) trained models that surpass strong baselines.

2 Related Work

Our research intersects with developments in T2I generative models, vision-language models (VLMs), and approaches to T2I evaluation, emphasizing on automatic and explainable methods.

Text-To-Image Generative Models. T2I generation has evolved from Generative Adversarial Network (GAN) based models [22, 43, 61, 62, 72] to visual transformers and diffusion models, like DALL-E [57, 58], Parti [78], Imagen [65] and Stable Diffusion [52, 63]. While these models showcase improved capabilities in image generation from textual prompts, they still grapple with challenges in accurately reflecting intricate T2I correspondences [15, 50, 60].

Fig. 3: The ConGen-Feedback data generation method: Top image shows a synthetic image from PickaPic with a predicted caption; Bottom image is a natural image from COCO with its longest available caption. Both undergo LLM processing to generate contradictions, feedback, textual misalignment labels, and visual misalignment labels, followed by visual bounding box generation.

Vision-Language Models. LLMs like the GPT series [1, 46, 55, 56] have revolutionized various fields but primarily focus on text, limiting their efficacy in vision-language tasks. Recent advancements [8, 10–12, 19, 34, 35, 38, 69, 73, 74, 76, 79] explore the synergy between visual components and LLMs to tackle tasks like image captioning and visual question answering (VQA), enhancing the understanding of visual content through textual descriptions.

T2I Automatic Evaluation. Traditional T2I evaluation methods utilize metrics like Fréchet Inception Distance (FID) [24] and Inception Score [66]. Alignment classification uses methods such as CLIP [54], CLIPScore [23], and CLIP-R [49], or via image-captioning model comparison [3, 47, 68]. Methods such as [31, 71] learn image quality reward models based on datasets with side-by-side human preferences and general ratings. In contrast, [75] focuses on image-text alignment, producing alignment scores without detailed feedback on what is wrong with the generated image. Some studies [15, 21, 25] dissect alignment into components like object detection and color classification. Both datasets and automatic metrics lack detailed misalignment feedback, a gap that our work addresses.

Image-Text Explainable Evaluation. Recent studies, such as TIFA [30] and VQ^2 [75], offer an interpretable evaluation scheme by generating question-answer pairs from the text. These pairs are then analyzed using Visual Question Answering (VQA) on the image. DSG [13] leverages this approach and creates a graph of questions, exploiting the dependencies between different questions and answers. These methods allow for detailed insights by contrasting expected text-based answers with image-derived responses, highlighting specific misalignments.

In a recent work, VPEval [16] generates a visual program using ChatGPT [45] and breaks down the evaluation process into a mixture of visual evaluation modules, which can be interpreted as an explanation.

Our method aims for the direct generation of explanations for image/text discrepancies without the need for an interrogative question-answering pipeline or breaking the evaluation task into sub-tasks.

Table 1: TV-Feedback dataset examples including aligned and misaligned text-image pairs, and textual and visual misalignment feedback.

Source Dataset	**PickaPic**	**ImageReward**	**COCO**	**Flickr30k**	**Open Images**	**ADE20K**
Images & Texts	Synthetic & Synthetic	Synthetic & Synthetic	Natural & Natural	Natural & Natural	Natural & Synthetic	Natural & Synthetic
# Instances	1,982,362	56,392	418,653	37,327	577,717	19,825
Image						
Positive Caption	A cartoon of a person dressed as a **joker** with a green coat and a blue tie against a gray background.	A close up of a glass of **blue** liquid on a table with a gray wall behind it .	A kitchen with cabinets, a stove, microwave and **refrigerator**.	Two men in Germany jumping **over** a rail at the same time without shirts.	A duck with a yellow beak is **swimming** in water.	A bed and a table with a lamp on it are in a room with a window and a view of **trees**.
Negative Caption	A cartoon of a person dressed as a **clown** with a green coat and a blue tie against a gray background	A close up of a glass of **red** liquid on a table with a gray wall behind it.	A kitchen with cabinets, a stove, microwave and a **toaster**.	Two men in Germany jumping **under** a rail at the same time without shirts.	A duck with a yellow beak is **flying** in the air.	A bed and a table with a lamp on it are in a room with a window and a view of a **lake**.
Misalignment Type	Object	Attribute	Object	Relation	Action	Object
Feedback	The person is dressed as a joker, not a clown	The liquid is blue, not red	The kitchen is missing a toaster, but has a refrigerator.	The men are jumping over a rail, not under it	The duck is swimming, not flying	The room has a view of trees, not a lake
Misalignment in Text	clown	red liquid	toaster	jumping under a rail	duck flying	a view of a lake
Visual Misalignment Detection	[2, 3, 996, 995] joker	[380, 308, 944, 666] blue liquid	[193, 327, 347, 553] refrigerator	[277, 26, 664, 477] two men and [608, 3, 729, 998] a rail	[339, 245, 581, 834] duck swimming	[409, 727, 559, 930] trees

3 Textual and Visual Feedback

Traditional image-text alignment evaluation models only provide alignment scores without detailed feedback. We propose to introduce a feedback mechanism, so that alignment models would not only score but also describe and visually annotate discrepancies between images and text.

In our multitask framework, as depicted in Figure 1, a single model handles two main tasks. In the first task, *Image-Text Entailment* [70], the model determines if an image corresponds to a given text description, outputs an alignment score to represent the likelihood of a "yes" answer[1]. The second task, *Textual and Visual Feedback*, is performed when misalignments are detected in an input image-text pair. The model is expected to provide three outputs: (a) a textual summary of discrepancies between the pair; (b) identification of misaligned text segments; (c) image visual misalignments, marked by bounding boxes.

To equip a model with the tasks outlined above, we perform VLM fine-tuning. To this end, an extensive training set encompassing all necessary information is required. The primary challenge lies in creating a sufficiently large training set with suitable examples. The following section provides a detailed description of the methodology we employed for generating such a set.

4 Training Dataset (TV Feedback) Generation

To construct our training set, which is designed to detect and interpret misalignments in image-text pairs, we first collect aligned image-text pairs. Then, utilizing LLMs and visual grounding models, we generate negative examples with misalignments accompanied by textual and visual feedback (see examples in Table 1). We next detail our approach, named ConGen-Feedback.

4.1 Collecting Positive Image-Text Pairs

We compile a set of over a million positive image-text pairs, consisting of synthetic and natural images. Approximately 65% of our examples consist of synthetic images, which were generated by a variety of T2I models from PickaPic [31] and ImageReward [71]. For these images, we employ the PaLI [12] model to predict captions that are aligned with the image.

We also include natural images sourced from two well-established datasets, COCO [37] and Flickr30k [51]. In these datasets, the images are already paired with human-annotated captions. When several captions are available per image, we select the longest to encourage textual richness.

Finally, we take localized narratives [53], captions offering a detailed point-of-view from the annotators) from ADE20k [82, 83] and OpenImages [32] and transform them into more conventional positive captions. To this end, we apply PaLM 2 [4] with a few-shot prompt (examples provided at the appendix) that rewrites the narratives into standardized captions.

4.2 LLM Generation of Misaligned Image-Text Pairs and Feedback

For each positive example from Section 4.1 we derive negative examples that include misaligned captions and relevant feedback. This is a four step approach (Figure 3):

[1] For direct comparison with other vision-language models, we present these outcomes as binary "Yes/No" responses instead of numerical scores.

1) Identify Misalignment Candidates. For each aligned image/caption pair, we tag the caption for part of speech tags with spaCy [27]. We then define four misalignment categories: object (noun), attribute (adjective), action (verb), and spatial relations. To ensure a balanced representation, we sample from these categories uniformly.

2) Generate Misalignment and Textual Feedback. Per chosen misalignment candidate, we instruct PaLM 2 [4] API with few-shot prompts to automatically generate: (a) a contradiction caption that introduces the target misalignment; (b) a detailed explanation of the contradiction; (c) a misalignment cue that pinpoints the contradictory element in the caption; and (d) a label for the visual bounding box to be placed on the image. Our instructions and few-shot prompts are presented in the appendix chapter.

3) Validate the Generation. Some LLM generations may be inaccurate. To increase the quality of the outputs, we filter out examples based on entailment validation as follows. Textual Entailment [18] models classify whether a *hypothesis* text is entailed by a *premise* text. We view this relationship as indicating the degree of semantic alignment. We use an entailment model by Honovich *et al*. [29] to assess the misalignment between our generated contradicting captions (hypothesis) and the original captions (premise), as well as the alignment between feedback (hypothesis) and caption (premise), as illustrated in the appendix chapter. Only valid contradictions and textual feedback, indicated by low and high entailment scores respectively, are retained.

4) Annotate Visual Feedback. To create visual feedback for the target misalignment, we employ GroundingDINO [42], which takes the textual label from PaLM 2's output and places a bounding box around the corresponding element in the image. To ensure consistent representation for different images, the bounding box coordinates are stored as a normalized range between 0 and 1000.

To assess the quality of our **T**extual and **V**isual (TV)-**Feedback** training set, we sampled 300 generated items for manual inspection. The outcome of this rigorous human validation is a high confidence score of 91%, which reflects the robustness of our automated generation process and the overall quality of the training dataset we have produced.

5 SeeTRUE-Feedback Benchmark

We present SeeTRUE-Feedback, a comprehensive alignment benchmark. It features 2,008 human-annotated instances that highlight textual and visual feedback.

5.1 Dataset Compilation

The SeeTRUE-Feedback Benchmark is based on the SeeTRUE dataset [75], featuring aligned and misaligned image-text pairs. Each misaligned pair includes three human-generated descriptions detailing the misalignment. Similar to our

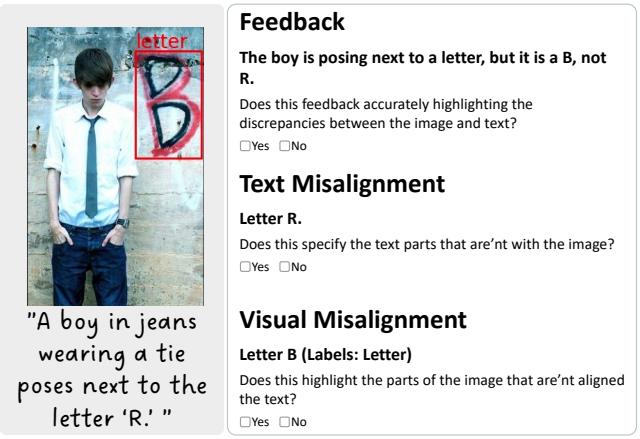


Fig. 4: SeeTRUE-Feedback annotation Amazon Mechanical Turk interface, questioning whether each part of the feedback, misalignment in text and misalignment in image are correct or not.

method in Section 4, we use PaLM 2 to generate a unified feedback statement at scale, covering both textual and visual misalignments. GroundingDINO then annotates these discrepancies on the images.

For verification, we conduct an annotation process on Amazon Mechanical Turk. Three annotators per instance, paid $18 per hour, evaluated the accuracy of feedback and visual annotations (Figure 4). Only unanimously agreed instances, 66% of the cases, were included in the final benchmark dataset.

5.2 Evaluation Metrics

We compare alignment evaluation models on SeeTRUE-Feedback using the following metrics:

- **Image-Text Alignment:** Binary Accuracy to gauge a model's ability to separate aligned and misaligned pairs.
- **Textual Feedback Quality:** Using BART NLI [33][2], we measure feedback quality by treating ground truth as the 'premise' and model predictions as the 'hypothesis', extracting an entailment score (0-1) as semantic alignment.
- **Misalignment in Text:** This metric evaluates the model's ability to identify specific segments within the text that are not aligned with the corresponding image. Similar to the metric above, we use BART NLI to measure the entailment between the predicted text and the ground truth. The goal is to pinpoint the exact parts of the input text that are sources of misalignment.
- **Visual Misalignment Detection:** We evaluate the model's bounding box generation using F1-Score@0.75 (indicating an IoU threshold of 0.75). This assessment combines precision and recall metrics to measure the accuracy

[2] `huggingface.co/facebook/bart-large-mnli`

of localization and object detection, ensuring a balance between avoiding missed objects (high precision) and minimizing false positives (high recall).

We note that *Image-Text Alignment* is applied to all 8,100 instances from the SeeTRUE dataset. The other metrics are computed on SeeTRUE-Feedback, containing only misaligned pairs. Examples showing our metric calculations can be seen at Figure 5.

Fig. 5: Metric results on the SeeTRUE-Feedback, showcasing calculations given the input, ground truth, and PaLI ft. model predictions, with NLI entailment scores calculated with BART NLI. The first row shows a high-scoring success example, while the second highlights a low-scoring failure with incorrect feedback and predictions.

6 Experiments

This section describes our experiments, encompassing model selection, fine-tuning methods on TV-Feedback, and thorough evaluation via the SeeTRUE-Feedback benchmark. We also validate automated metric reliability through human annotation and assess model robustness with 'out-of-distribution' examples from diverse sources.

6.1 Models and Baselines

Our experiments span multiple leading vision-language models, examined in both zero-shot and fine-tuned scenarios: MiniGPT-v2 (7B-ft) [9], LLaVa-1.5 (Vicuna-7b [39]), InstructBLIP (FlanT5$_{\text{XL}}$) [20], mPLUG-Owl (LLaMa-7B-ft) [77], PaLI Series [10–12]. Our methodology introduces a feedback task that, while new, aligns with the capabilities expected of leading VLMs, renowned for their instruction following capabilities. The task's design mirrors scenarios these models encounter during training, ensuring they're well-equipped to handle it.

For the zero-shot experiments, we queried the models with specific questions to assess their inherent capabilities:

1. **Image-Text Entailment:** Assessing if an image semantically aligns with a given description (*"Does this image entail* `<text>`*?"*).
2. **Textual Misalignment Detection:** Identifying misaligned text elements (*"Which part of* `<text>` *doesn't align with the image?"*).
3. **Visual Misalignment Identification:** *"What part of the following image is not aligned with the text:* `<text>`*?"* – aimed at pinpointing visual discrepancies in the image relative to the text.

Our work uniquely offers an end-to-end assessment of both textual and visual misalignment. To evaluate baseline models for visual misalignment, we adopt a two-step approach. First, we ask for a textual misalignment description. Then, we employ the GroundingDINO grounding model to extract bounding-box information, since the baseline models do not output a bounding-box. In addition, our fine-tuned model is capable of predicting the feedback along with both textual and visual misalignments in a single inference. To accurately assess our model's performance alongside the baselines, we report the visual misalignment performance using both the GroundingDINO output and our model's predictions

For the supervised experiments, we fine-tuned PaLI models with the visual question answering task using specific questions (additional fine-tuning details are in at the appendix). The fine-tuning tasks encompass:

1. **Image-Text Alignment:** Using the same query as in the zero-shot setup, *"Does this image entail the description* `<text>`*?"*, we expected a binary 'yes'/'no' response.
2. **Textual and Visual Feedback:** We use a query for combined feedback: *"Describe the misalignments between the image and the text:* `<text>`*"*. The expected response format is *'<feedback> | <misalignment in text> | <misalignment in image (bounding-box)>'* , aiming to extract detailed feedback and specific misalignment indicators in a single model interaction.

Table 2: Comparative performance of image/text alignment models on the SeeTRUE-Feedback Benchmark. "ft." stands for fine-tuned on TV-Feedback. Legend: (∗) marks the performance using PaLI bounding-box detector instead of GroundingDINO used for the baseline models.

	Feedback NLI		Textual Misalignment NLI		Visual Misalignment F1-Score@0.75		Binary Class. Acc.
Model / Split	Test	Val	Test	Val	Test	Val	Test
PaLI-3 [11]	0.18	0.22	0.23	0.46	0.47/0.47*	0.35/0.48*	0.51
InstructBLIP (FlanT5$_{XL}$) [19]	0.41	0.39	0.56	0.50	0.48	0.39	0.74
mPLUG-Owl (LLaMa-7B-ft) [76]	0.63	0.58	0.30	0.35	0.43	0.48	0.50
MiniGPT-v2 (7B-ft) [8]	0.46	0.37	0.56	0.58	0.44	0.43	0.68
LLaVa-1.5 (Vicuna-7b) [38]	0.57	0.48	0.17	0.21	0.43	0.48	0.72
PaLI-3 ft. Multitask [11]	0.72	**0.88**	0.76	**0.92**	0.61/0.49*	0.83/0.57*	0.75
PaLI ft. Multitask [12]	**0.75**	0.87	**0.78**	0.92	**0.65**/0.35*	**0.84**/0.39*	0.77
PaLI-X ft. Multitask [10]	0.74	0.87	0.76	0.90	0.61/0.49*	0.84/0.55*	**0.79**

Fig. 6: Qualitative comparison of model outputs on two examples from SeeTRUE-Feedback. The PaLI-X model, fine-tuned on TV-Feedback, effectively identifies a distinct misalignment related to the tennis player's action and the relative position between the teddy bear and laptop, demonstrating its refined feedback ability.

6.2 Main Results

Table 2 presents our main results on the SeeTRUE-Feedback benchmark, and Figure 6 provides qualitative examples. *Val* results refer to "in-distribution" auto-generated data, while *Test* results refer to "out-of-distribution" human created examples.

Overall, the PaLI models fine tuned on TV-Feedback outperform the baselines on all metrics. For example, Non-PaLI models achieved Feedback NLI scores from 0.406 to 0.627, while PaLI models reached 0.718 to 0.749. The largest, PaLI-X [10] model achieved the highest performance on the binary alignment classification task. Surprisingly, it underperformed the smaller PaLI models on most feedback generation tasks. Specifically, the smaller but most recent PaLI-3 model, is best performing on the in-distribution testset, but less so on the out-of-distribution examples. The PaLI models gap over the baselines is very large on the textual feedback tasks, but less so on the bounding box task. In future work, we plan to improve the multitasking efficiency of the fine-tuned models. Figure 5 shows metrics results calculated on SeeTRUE-Feedback examples to give a more clear overview. More details about our metrics and evaluation process are available at the appendix chapter.

6.3 Human Ratings and Auto-Metrics Correlations

For unbiased model evaluations and automatic metric validation, we conducted an Amazon Mechanical Turk study involving 1,500 instances. These instances included 250 samples from each of the six models used in our experiments. Annotators were assigned to evaluate the accuracy of these models in identifying and describing image-text misalignments, with each of the 1,500 instances being rated by three human raters.

At the appendix chapter we present results, highlighting PaLI-X with top scores in feedback accuracy (75.7%), textual misalignment (80.1%), and visual misalignment detection (63.5%), showcasing superior alignment with human judgments. We present the annotators' agreement chart for each model's predictions as well.

We evaluated our auto-evaluation metrics against 1,750 human ratings. Textual metrics included BART NLI [33], BLEU-4 [48], ROUGE-L [36], METEOR [5], CIDEr [67], BERTScore [81], and TRUE NLI [28]. Visual metrics comprised AP, IoU, Precision, Recall, and F1-Score at 0.75 threshold. Figure 7 shows the correlations, identifying BART NLI and F1-Score@0.75 as the most correlated textual and visual metrics, respectively. This analysis confirms the relevance and reliability of our automatic evaluation measures.

Table 3: Human annotation results comparing model performances in feedback accuracy and misalignment identification. The values represent the mean percentage of "yes" responses from annotators. T. Misalignment stands for textual misalignments, and V. Misalignment for visual misalignments.

Model	Feedback	T. Misalignment	V. Misalignment
PaLI-X ft.	**75.7**	**80.1**	**63.5**
PaLI-3 ft.	68.1	72.4	61.6
LLaVA 1.5 7B	29.9	5.1	16.2
mPlug-Owl 7B	14.22	5.5	5.9
MiniGPT V2	11.6	39.1	21.7
InstructBLIP	1.3	32.6	29.9

6.4 Out-of-distribution Generalization

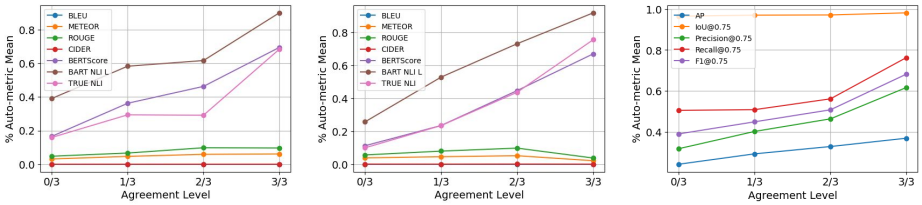
We evaluate our model's generalization capabilities on 100 'in-the-wild' Text-to-Image (T2I) generations from academic papers [7, 59, 64] and Reddit, created using models like Adobe Firefly [2], Composable Diffusion [41], and Stable Diffusion versions 1.0 and 2.1. Figure 2 shows a selection of these results, with more available at the appendix chapter.

We employed the fine-tuned PaLI-X model on TV-Feedback to predict textual and visual feedback, and these results were rated by three human annotators following our benchmark protocol (Section 5 and Figure 4).

Results indicated a feedback accuracy of 71%, textual misalignment detection accuracy of 80%, and visual misalignment accuracy of 60%, showcasing the model's broad generalization to various out-of-distribution prompts and models. These findings also highlighted areas for potential model enhancement.

7 Analysis and Limitations

In this section, we analyze methodological ablation studies and discuss the limitations along with future directions for enhancing our model.



(a) Feedback agreement **(b)** Textual agreement **(c)** Visual agreement

Fig. 7: Correlation analysis between human ratings and automated metrics for feedback evaluation. Subfigures (a) and (b) explore textual feedback correlations with metrics like BART NLI and BERTScore, while subfigure (c) illustrates visual feedback correlations with metrics like IoU and F1-Score. The X-axis denotes annotator agreement; the Y-axis shows mean metric scores, identifying most correlated metrics.

7.1 Methodological Ablations Studies

We conduct an ablation study to evaluate our methodologies. Our multi-task training approach achieves superior performance, with 75% entailment accuracy and a 0.72 BART-NLI [33] score in feedback, highlighting its efficiency. Fine-tuning on our filtered dataset (77% of total data) improves feedback and entailment tasks but degrades others, underscoring the positive impact of NLI model-based filtering. In a 2-step experiment simulating baselines, using GroundingDino [42] for grounding with predicted visual misalignment text labels improves bounding-box precision by 0.11 in F1-Score, showcasing its efficacy over our model.

Table 4: Comparing PaLI-3 [11] models: baseline, fine-tuned (entailment, feedback), multitask (unfiltered data, entailment+feedback). $+GD$ denotes two-step visual misalignment (b-box) prediction via GroundingDino. The study underscores the benefits of multitask training and the effectiveness of dataset filtering in enhancing performance.

Model	Feedback NLI	Textual Mis. NLI	Visual Mis. F1@0.75	Binary Acc.
Baseline	0.18	0.23	0.47	0.51
Entailment	-	-	-	0.74
Feedback	0.70	0.76	0.50	-
Feedback+GD	0.72	0.77	**0.61**	-
Multitask (Unf.)	0.69	**0.80**	0.51	0.74
Multitask	**0.72**	0.77	0.49	**0.75**

7.2 Limitations and Future Work

In our evaluation across various datasets, our model showed proficiency but also revealed key improvement areas:

– **No Visual Feedback:** In cases where no visual feedback is expected (Figure 8a), our model incorrectly predicts it. To address this, we plan to enrich TV-Feedback with scenarios like “an image of a horse” becoming “an image

of a horse *and a dog*," with feedback like "there is only a horse, not a dog and a horse," and without generating a bounding box.
- **Multiple Misalignments:** Instances requiring identification of multiple misalignments (Figure 8b). Our model often detects only one issue where several exist. We will enhance TV-Feedback with cases like transforming "a *white dog* and a *black cat*" into "a *white cat* and a *black dog*," with feedback addressing both color and species misalignments and bounding boxes highlighting each. In the appendix chapter , we show how a finetuned model detects multiple misalignments sequentially using a MagicBrush [80] dataset example. The model identifies one misalignment at a time, and feedback signals guide an instruction editing model to iteratively correct them.
- **Loose Bounding Boxes:** As observed in Fig. 2 for the SD2.1 example, our model occasionally generates loose bounding boxes. For instance, rather than confining the b-box to the pizza, it may encompass the entire person.

These enhancements to the TV-Feedback are aimed to improve the model's ability to address various misalignment types, making it more effective and applicable in real-world situations.

Fig. 8: Model limitations: (a) Misalignment due to a *missing object*, where the model incorrectly adds a bounding box over a horse; (b) Multiple misalignments, with the model only identifying one - the top parrot should be green and the bottom a white cat. The model requires multiple iterations for full correction.

8 Conclusion

Our research develops an end-to-end strategy providing visual and textual feedback for text-to-image models, targeting and clarifying alignment issues for refinement. We introduced TV-Feedback, a specialized dataset for fine-tuning feedback in these models, leading to several robust developments. This dataset and methodology demonstrate broad potential, notably in enhancing text-to-image generation, improving dataset annotation accuracy, and refining image captioning through detailed feedback. Our comprehensive testing on SeeTRUE-Feedback and various scenarios validates our approach's effectiveness. While primarily aimed at text-to-image feedback enhancement, we anticipate our work will significantly improve generative model accuracy across different domains.

Bibliography

[1] Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] Adobe: Adobe firefly (https://www.adobe.com/sensei/generative-ai/firefly.html), https://www.adobe.com/sensei/generative-ai/firefly.html

[3] Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation (2016)

[4] Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J.H., Shafey, L.E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G.H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C.A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A.C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D.R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., Wu, Y.: Palm 2 technical report (2023)

[5] Banerjee, S., Lavie, A.: METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. In: ACL workshop on Evaluation Measures for MT and Summarization (2005)

[6] Betker, J., Goh, G., Jing, L., TimBrooks, Wang, J., Li, L., LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, Ramesh, A.: Improving image generation with better captions (2023), https://cdn.openai.com/papers/dall-e-3.pdf

[7] Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Trans. Graph. **42**(4) (jul 2023). https://doi.org/10.1145/3592116, https://doi.org/10.1145/3592116

[8] 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)

[9] 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)

[10] Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., Shakeri, S., Dehghani, M., Salz, D.M., Lucic, M., Tschannen, M., Nagrani, A., Hu, H., Joshi, M., Pang, B., Montgomery, C., Pietrzyk, P., Ritter, M., Piergiovanni, A.J., Minderer, M., Pavetic, F., Waters, A., Li, G., Alabdulmohsin, I.M., Beyer, L., Amelot, J., Lee, K., Steiner, A., Li, Y., Keysers, D., Arnab, A., Xu, Y., Rong, K., Kolesnikov, A., Seyedhosseini, M., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: Pali-x: On scaling up a multilingual vision and language model. ArXiv **abs/2305.18565** (2023), https://api.semanticscholar.org/CorpusID:258967670

[11] Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., Mustafa, B., Goodman, S., Alabdulmohsin, I., Padlewski, P., Salz, D., Xiong, X., Vlasic, D., Pavetic, F., Rong, K., Yu, T., Keysers, D., Zhai, X., Soricut, R.: Pali-3 vision language models: Smaller, faster, stronger (2023)

[12] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S.A., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: Pali: A jointly-scaled multilingual language-image model (2023), https://arxiv.org/abs/2209.06794

[13] Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian Scene Graph: Improving Reliability in Fine-Grained Evaluation for Text-to-Image Generation. In: arXiv:2310.18235 (2023)

[14] Cho, J., Lu, J., Schwenk, D., Hajishirzi, H., Kembhavi, A.: X-lxmert: Paint, caption and answer questions with multi-modal transformers. In: EMNLP (2020)

[15] Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers (2022)

[16] Cho, J., Zala, A., Bansal, M.: Visual programming for text-to-image generation and evaluation. In: NeurIPS (2023)

[17] Crowston, K.: Amazon mechanical turk: A research tool for organizations and information systems scholars. In: Bhattacherjee, A., Fitzgerald, B. (eds.) Shaping the Future of ICT Research. Methods and Approaches. pp. 210–221. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)

[18] Dagan, I., Dolan, B., Magnini, B., Roth, D.: Recognizing textual entailment: Rational, evaluation and approaches–erratum. Natural Language Engineering **16**(1), 105–105 (2010)

[19] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)

[20] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)

[21] Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., Yang, Y.: Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022)

[22] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[23] Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021)

[24] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[25] Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. arXiv preprint arXiv:1910.13321 (2019)

[26] Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI

[27] Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spacy: Industrial-strength natural language processing in python (2020). https://doi.org/10.5281/zenodo.1212303, https://github.com/explosion/spaCy/tree/master

[28] Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., Matias, Y.: TRUE: Re-evaluating factual consistency evaluation. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3905–3920. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.287, https://aclanthology.org/2022.naacl-main.287

[29] Honovich, O., Choshen, L., Aharoni, R., Neeman, E., Szpektor, I., Abend, O.: Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7856–7870. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.619, https://aclanthology.org/2021.emnlp-main.619

[30] Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897 (2023)

[31] Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation (2023)

[32] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020)

[33] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019)

[34] 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)

[35] Li*, L.H., Zhang*, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022)

[36] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)

[37] Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014), https://api.semanticscholar.org/CorpusID:14113767

[38] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)

[39] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)

[40] Liu, N., Li, S., Du, Y., Tenenbaum, J., Torralba, A.: Learning to compose visual relations. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 23166–23178. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/c3008b2c6f5370b744850a98a95b73ad-Paper.pdf

[41] Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: Computer Vision–ECCV

2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII. pp. 423–439. Springer (2022)

[42] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)

[43] Mansimov, E., Parisotto, E., Ba, J., Salakhutdinov, R.: Generating images from captions with attention. In: ICLR (2016)

[44] Marcus, G., Davis, E., Aaronson, S.: A very preliminary analysis of dall-e 2 (2022)

[45] OpenAI: Chatgpt (2022), https://openai.com/blog/chatgpt

[46] OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023), https://api.semanticscholar.org/CorpusID:257532815

[47] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040

[48] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)

[49] Park, D.H., Azadi, S., Liu, X., Darrell, T., Rohrbach, A.: Benchmark for compositional text-to-image synthesis. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021), https://openreview.net/forum?id=bKBhQhPeKaF

[50] Petsiuk, V., Siemenn, A.E., Surbehera, S., Chin, Z., Tyser, K., Hunter, G., Raghavan, A., Hicke, Y., Plummer, B.A., Kerret, O., Buonassisi, T., Saenko, K., Solar-Lezama, A., Drori, I.: Human evaluation of text-to-image models on a multi-task benchmark (2022)

[51] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. International Journal of Computer Vision **123**, 74 – 93 (2015), https://api.semanticscholar.org/CorpusID:6941275

[52] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)

[53] Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: ECCV (2020)

[54] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html

[55] Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018), https://api.semanticscholar.org/CorpusID:49313245
[56] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
[57] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
[58] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8821–8831. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/ramesh21a.html
[59] Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. arXiv preprint arXiv:2306.08877 (2023)
[60] Rassin, R., Ravfogel, S., Goldberg, Y.: Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models (2022)
[61] Reed, S., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: Advances in Neural Information Processing Systems (2016)
[62] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1060–1069. PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/reed16.html
[63] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
[64] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
[65] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. ArXiv **abs/2205.11487** (2022), https://api.semanticscholar.org/CorpusID:248986576
[66] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016), https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf

[67] Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
[68] Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR. pp. 4566–4575. IEEE Computer Society (2015)
[69] Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models (2023)
[70] Xie, N., Lai, F., Doran, D., Kadav, A.: Visual entailment task for visually-grounded language learning. arXiv preprint arXiv:1811.10582 (2018)
[71] Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation (2023)
[72] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks (2018)
[73] Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of gpt-3 for few-shot knowledge-based vqa. In: AAAI (2022)
[74] Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action (2023)
[75] 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)
[76] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Jiang, C., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., Huang, F.: mplug-owl: Modularization empowers large language models with multimodality (2023)
[77] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
[78] 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., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation (2022)
[79] Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6713–6724 (2019). https://doi.org/10.1109/CVPR.2019.00688
[80] Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In: Advances in Neural Information Processing Systems (2023)
[81] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. In: ICLR (2020)
[82] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)

[83] Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)