

Brian Gordon<sup>\* 1</sup>, Yonatan Bitton<sup>\* 2</sup>, Yonatan Shafir<sup>1,2</sup>, Roopal Garg<sup>2</sup>, Xi Chen<sup>2</sup>, Dani Lischinski<sup>2,3</sup>, Daniel Cohen-Or<sup>1,2</sup>, Idan Szpektor<sup>2</sup>

<sup>1</sup> Tel Aviv University <sup>2</sup> Google Research <sup>3</sup> The Hebrew University of Jerusalem

## 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 method code and human curated test set are available at: <https://mismatch-quest.github.io/>

## 1. Introduction

Recently, text/image generative models [5, 13, 14, 27, 60, 66, 68, 81] 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 [42, 46]. They especially struggle with complex, nuanced, or out-of-distribution descriptions and fail to generate images which follow the prompt precisely [8, 61]. 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 [24, 74, 78]. 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



Figure 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.

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.

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 **Textual and Visual (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 [39], Flickr30K [53], PickaPic [33], ImageReward [74], ADE20K [84, 85], and OpenImages [34]. 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. Hu-

<sup>\*</sup>Equal contribution.

| Source                              | SD XL  | SD 2.1   | Adobe Firefly   | Composable Diffusion  |
|-------------------------------------|--|--|---|---|
| Input Prompt                        | "Two colleagues, one with a blue umbrella and the other without an umbrella, walking in the snow."   | "A young couple sharing pizza in a park, the man holds a slice in his hand"  | "A blue cat is sitting next to a green dog"   | "A red bench and a yellow clock"  |
| Generated Image                     |   |   |   |    |
| Predicted Textual & Visual Feedback | <br>One of the colleagues is holding an umbrella, not without an umbrella | <br>The man is holding a whole pizza, not a slice | <br>The cat is sitting next to a green cat, not a green dog | <br>The clock is black and white, not yellow |

Figure 2. Qualitative analysis of out-of-distribution results: Showcasing image-text pairs generated by Stable-Diffusion XL [54], Stable-Diffusion 2.1 [66], Adobe Firefly [1] and Composable Diffusion [43] (credits to Chefer et al. [8]) text-to-image models alongside the corresponding textual and visual feedback as predicted by the PaLI 55B model finetuned on TV-Feedback.

man 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 [18], 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) v1.0 and v2.1 [66], SD XL [54], Composable Diffusion [43], and Adobe Firefly [1]. Finally, our ablation studies verify the advantage of our multitask training, a single model generating both misalignment labels and feedback for different prompts, compared to training individual models for each task, as well as the effectiveness of our training set filtering strategy.

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 [23, 45, 63, 64, 75] to visual transformers and diffusion models, like DALL-E [59, 60], Parti [81],Imagen [68] and Stable Diffusion [54, 66]. While these models showcase improved capabilities in image generation from textual prompts, they still grapple with challenges in accurately reflecting intricate T2I correspondences [15, 52, 61].

**Vision-Language Models.** LLMs like the GPT series [7, 48, 56, 57] have revolutionized various fields but primarily focus on text, limiting their efficacy in vision-language tasks. Recent advancements [10–13, 21, 36, 37, 41, 72, 76, 77, 79, 82] 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.

| -- Synthetic Images --              | Positive Caption  | Misalignment Type    | Contradiction  | Feedback   | Textual Misalignment | Visual Misalignment | (long beard) |
|-------------------------------------|---|----------------------|--|--|----------------------|---------------------|--------------|
|                                     | An old wizard is depicted in black and white with his <b>long</b> beard and hat | Attribute (adj=long) | An old wizard is depicted in black and white with his <b>short</b> beard and hat | The old wizard has a long beard, not a short one | short beard          | Long beard          |              |
|                                     | A table with a blender, <b>oil</b> and several seasonings.                      | Object (noun=oil)    | A table with a blender, <b>beer</b> and several seasonings.                      | The table has a bottle of oil, not beer          | Beer                 | Oil                 |              |
| ----- LLM Feedback Generation ----- |   |                      |  |  |                      |                     |              |
| -- Visual Grounding --              |   |                      |  |  |                      |                     |              |

Figure 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 bounding box generation.

**T2I Automatic Evaluation.** Traditional T2I evaluation methods utilize metrics like Fréchet Inception Distance (FID) [25] and Inception Score [69]. Alignment classification uses methods such as CLIP [58], CLIPScore [24], and CLIP-R [51], or via image-captioning model comparison [2, 49, 71]. Methods such as [33, 74] learn image quality reward models based on datasets with side-by-side human preferences and general ratings. In contrast, [78] focuses on image-text alignment, producing alignment scores without detailed feedback on what is wrong with the generated image. Some studies [15, 22, 26] 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 [31] and  $VQ^2$  [78], 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 [16] 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 [17] generates a visual program using ChatGPT [47] 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.

### 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* [73], the model determines if an image corresponds to a given text description, outputs an alignment score to represent the likelihood of a “yes” answer<sup>1</sup>. 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,

<sup>1</sup>For direct comparison with other vision-language models, we present these outcomes as binary “Yes/No” responses instead of numerical scores.

| 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 <b>joker</b> with a green coat and a blue tie against a gray background. | A close up of a glass of <b>blue</b> liquid on a table with a gray wall behind it . | A kitchen with cabinets, a stove, microwave and <b>refrigerator</b> .             | Two men in Germany jumping <b>over</b> a rail at the same time without shirts.     | A duck with a yellow beak is <b>swimming</b> in water.                              | A bed and a table with a lamp on it are in a room with a window and a view of <b>trees</b> .  |
| Negative Caption              | A cartoon of a person dressed as a <b>clown</b> with a green coat and a blue tie against a gray background  | A close up of a glass of <b>red</b> liquid on a table with a gray wall behind it.   | A kitchen with cabinets, a stove, microwave and a <b>toaster</b> .                | Two men in Germany jumping <b>under</b> a rail at the same time without shirts.    | A duck with a yellow beak is <b>flying</b> 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 <b>lake</b> . |
| 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  |

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

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 Pick-aPic [33] and ImageReward [74]. For these images, we employ the PaLI 2 [13] model to predict captions that are aligned with the image.

We also include natural images sourced from two well-established datasets, COCO [39] and Flickr30k [53]. 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 [55], captions offering a detailed point-of-view from the annotators) from ADE20k [84, 85] and OpenImages [34] and transform them into more conventional positive captions. To this end, we apply PaLM 2 [3] with a few-shot prompt (see Appendix B) 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) Identify Misalignment Candidates.** For each aligned image/caption pair, we tag the caption for part of speech tags with spaCy [28]. 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 [3] 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 Appendix B.1.

**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 [19] 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



Figure 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.

caption (premise), as illustrated in Appendix B.2. 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 [44], 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 **Textual and Visual (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 [78], featuring aligned and misaligned image-text pairs. Each misaligned pair includes three human-generated descriptions detailing the misalignment. Similar to our 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 [35]<sup>2</sup>, we measure feedback quality by treating ground truth as the ‘premise’ and model predictions as the ‘hypothesis’, extracting an entailment score (0-100) as semantic alignment.
- **Misalignment in Text:** Using BART NLI, we evaluate shorter textual misalignments, determining the model’s precision in identifying text segments incongruent with the image.
- **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 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.

## 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 [40]), InstructBLIP (FlanT5<sub>XL</sub>) [20], mPLUG-Owl (LLaMa-7B-ft) [80], PaLI Series (5B, 17B, 55B) [12, 13]<sup>3</sup>.

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>?*”).

<sup>2</sup>[huggingface.co/facebook/bart-large-mnli](https://huggingface.co/facebook/bart-large-mnli)

<sup>3</sup>7B, 5B, 17B, 55B stands for the numbers of model parameters.

| Model / Split                             | Feedback NLI |              | Textual Misalignment NLI |              | Visual Misalignment F1-Score@0.75 |              | Binary Class. Acc. |
|---|--------------|--------------|--------------------------|--------------|-----------------------------------|--------------|--------------------|
|   | Test         | Val          | Test                     | Val          | Test                              | Val          | Test               |
| PaLI 5B                                   | 0.179        | 0.216        | 0.226                    | 0.455        | 0.471                             | 0.478        | 0.514              |
| InstructBLIP (FlanT5 <sub>XL</sub> ) [21] | 0.406        | 0.387        | 0.555                    | 0.502        | 0.481                             | 0.392        | 0.736              |
| mPLUG-Owl (LLaMa-7B-ft) [79]              | 0.627        | 0.580        | 0.297                    | 0.347        | 0.435                             | 0.485        | 0.504              |
| MiniGPT-v2 (7B-ft) [10]                   | 0.463        | 0.374        | 0.560                    | 0.583        | 0.440                             | 0.426        | 0.676              |
| LLaVa-1.5 (Vicuna-7b) [41]                | 0.571        | 0.483        | 0.173                    | 0.206        | 0.434                             | 0.484        | 0.717              |
| PaLI 5B ft. Multitask                     | 0.718        | <b>0.882</b> | 0.765                    | <b>0.925</b> | 0.488                             | <b>0.575</b> | 0.752              |
| PaLI 17B ft. Multitask                    | <b>0.749</b> | 0.867        | <b>0.785</b>             | 0.921        | 0.353                             | 0.395        | 0.770              |
| PaLI 55B ft. Multitask                    | 0.743        | 0.871        | 0.762                    | 0.898        | <b>0.490</b>                      | 0.550        | <b>0.787</b>       |

Table 2. Comparative performance of image/text alignment models on the SeeTRUE-Feedback Benchmark. “ft.” stands for fine-tuned on TV-Feedback.

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.

For the supervised experiments, we fine-tuned PaLI models with the visual question answering task using specific questions (additional fine-tuning details are in Appendix D). 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.

## 6.2. Main Results

Table 2 presents our main results on the SeeTRUE-Feedback benchmark, and Figure 5 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, 55B PaLI 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 PaLI 5B 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.

## 6.3. Human Ratings and Auto-Metrics Correlations

| Model        | Feedback    | T. Misalignment | V. Misalignment |
|--------------|-------------|-----------------|-----------------|
| PaLI 55B ft. | <b>75.7</b> | <b>80.1</b>     | <b>63.5</b>     |
| PaLI 5B 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            |

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.

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 accu-



Figure 5. Qualitative comparison of model outputs on two examples from SeeTRUE-Feedback. The PaLI 55B 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.

racy of these models in identifying and describing image-text misalignments, with each of the 1,500 instances being rated by three human raters.

Table 3 displays results, highlighting PaLI 55B with top scores in feedback accuracy (75.7%), textual misalignment (80.1%), and visual misalignment detection (63.5%), showcasing superior alignment with human judgments

To evaluate our auto-evaluation metrics, we compared our automated evaluation metrics against 1,750 human ratings. Textual metrics included BART NLI [35], BLEU-4 [50], ROUGE-L [38], METEOR [4], CIDEr [70], BERTScore [83], and TRUE NLI [30], while visual metrics comprised AP, IoU, Precision, Recall, and F1-Score at 0.75 threshold. Figure 6 presents these correlations, correlating human ratings with mean auto-metric scores, identifying BART NLI and F1-Score@0.75 as the most correlated textual and visual metrics, respectively. The analysis highlights a robust correlation between our automated metrics and human ratings, confirming the relevance and reliability of our automatic evaluation measures for our task.

#### 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 [8, 62, 67] and Reddit, created using models like Adobe Firefly [1], Composable Diffusion [43], and Stable Diffusion versions 1.0 and 2.1. Figure 2 shows a selection of these results, with more available Appendix A.

We employed the fine-tuned PaLI 55B 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.

### 7.1. Methodological Ablations Studies

We conduct an ablation study to evaluate the effectiveness of our methodologies. Initially, we compare our multi-task training approach, demonstrating superior performance with 75% entailment accuracy and a 0.72 BART-NLI [35] score in feedback, emphasizing the efficiency of our multi-task approach. Additionally, fine-tuning on our filtered dataset (77% of total data) improves performance in feedback and entailment tasks but degrades performance in others, highlighting the positive impact of our NLI model-based filtering. In a 2-step experiment simulating baselines, using GroundingDino [44] for grounding with predicted visual misalignment text labels improves bounding-box pre-

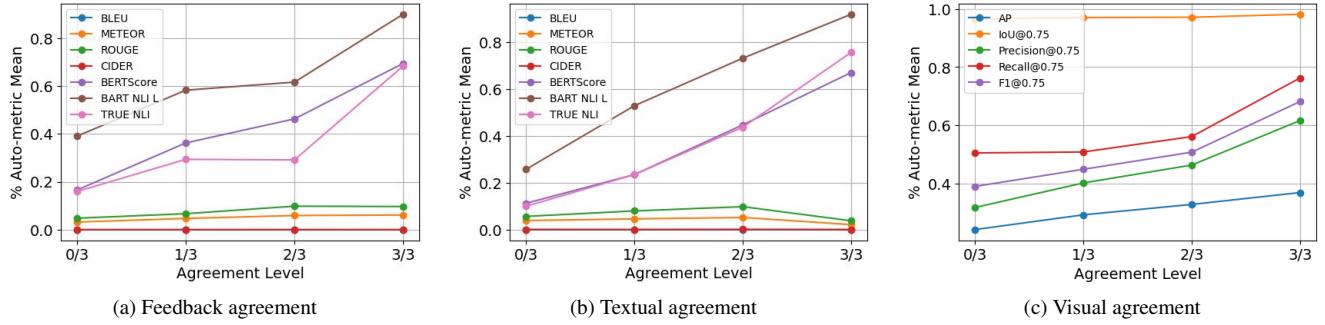


Figure 6. Correlation analysis between human ratings and automated metrics for feedback evaluation. Subfigures (a) and (b) show textual feedback correlations using metrics like BART NLI and BERTScore, while subfigure (c) displays visual feedback correlations with metrics such as IoU and F1-Score. The X-axis represents annotator agreement, and the Y-axis shows the mean metric scores, identifying the most correlated metrics.

| 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             | <b>0.61</b>         | -           |
| Multitask (Unf.) | 0.69         | <b>0.80</b>      | 0.51                | 0.74        |
| Multitask        | <b>0.72</b>  | 0.77             | 0.49                | <b>0.75</b> |

Table 4. Comparing PaLI 5B models across configurations: baseline (untrained), fine-tuned for entailment, feedback, multitask with unfiltered data, and multitask (entailment+feedback). +GD remarks that visual misalignment bounding-box prediction was performed in a 2-step way using the GroundingDino model. The study underscores the benefits of multitask training and the effectiveness of dataset filtering in enhancing performance.

cision by 0.11 in F1-Score, showcasing its efficacy over our model.

## 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:** cases where no visual feedback is expected, as shown in Figure 7a. Our model is trained to always predict visual feedback and incorrectly generates it. To address this, we plan to enrich TV-Feedback with such scenarios, such as “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 importantly, without generating a bounding box.
- **Multiple Misalignments:** instances requiring identification of multiple misalignments, as shown in Figure 7b. Our model often detects only one issue where several exist. For enhancement, we will integrate such cases into TV-Feedback, like transforming “a white dog and a black cat” into “a white cat and a black dog”, with feedback addressing both the color and species misalignments and

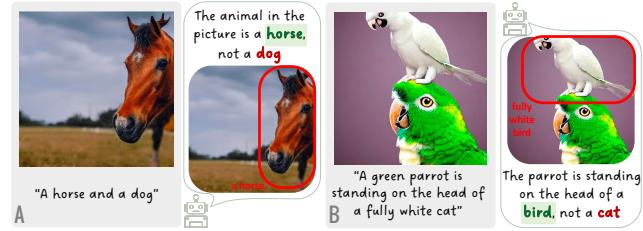


Figure 7. Model limitations: (a) Misalignment due to a *missing object*, where no bounding box is expected, yet the model produces one over the only present object, a horse; (b) A scenario with multiple misalignments where the model identifies only one - here, the top parrot should be green (not white), and the bottom should be a white cat (not a green parrot). This limitation is evident when the model addresses only one error at a time, necessitating re-generation for complete correction.

bounding boxes highlighting each.

- **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 bounding 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.

## 8. Conclusion

Our work introduces an end-to-end approach that provides visual and textual feedback in text-to-image models, identifying and elucidating alignment discrepancies with visual annotations for targeted model refinement. We created TV-Feedback, a dedicated dataset for fine-tuning feedback models, leading to the development and evaluation of several robust models. We thoroughly tested on SeeTRUE-Feedback and out-of-distribution scenarios. While our

current work focuses on improving feedback mechanisms within text-to-image models, we hope it will significantly contribute to the refinement of the generative accuracy of these models.

## Acknowledgments

We express our gratitude to Amir Globerson, Jay Tenenbaum, and Shlomi Fruchter for their valuable feedback on earlier versions of the manuscript. We thank Jialin Wu and Sebastian Goodman for their help. This research was supported in part by the Israel Science Foundation (grants no. 2492/20, 3441/21 and 3611/21), Len Blavatnik and the Blavatnik family foundation.

## References

- [1] Adobe. Adobe firefly (<https://www.adobe.com/sensei/generative-ai/firefly.html>). 2, 7, 1, 3
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation, 2016. 3
- [3] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Ke-fan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junghan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxi-ayou Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023. 4
- [4] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on Evaluation Measures for MT and Summarization*, 2005. 7
- [5] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions, 2023. 1
- [6] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. Jax: composable transformations of python+ numpy programs. 2018. 4
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 2, 1
- [8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4), 2023. 1, 2, 7
- [9] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 5
- [10] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2, 6
- [11] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel M. Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, A. J. Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim M. Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Palix: On scaling up a multilingual vision and language model. *ArXiv*, abs/2305.18565, 2023.

- [12] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. Pali-3 vision language models: Smaller, faster, stronger, 2023. 5
- [13] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Alexander Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. 2023. 1, 2, 4, 5
- [14] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-lxmert: Paint, caption and answer questions with multi-modal transformers. In *EMNLP*, 2020. 1
- [15] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. 2022. 2, 3
- [16] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian Scene Graph: Improving Reliability in Fine-Grained Evaluation for Text-to-Image Generation. In *arXiv:2310.18235*, 2023. 3
- [17] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. In *NeurIPS*, 2023. 3
- [18] Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 2
- [19] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches—erratum. *Natural Language Engineering*, 16(1): 105–105, 2010. 4, 1
- [20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 5
- [21] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2, 6
- [22] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 3
- [23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 2
- [24] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 1, 3
- [25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3
- [26] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *arXiv preprint arXiv:1910.13321*, 2019. 3
- [27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1
- [28] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. 4
- [29] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend.  $Q^2$ : Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 4, 1
- [30] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States, 2022. Association for Computational Linguistics. 7
- [31] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 3
- [32] Norman P Jouppi, Doe Hyun Yoon, George Kurian, Sheng Li, Nishant Patil, James Laudon, Cliff Young, and David Patterson. A domain-specific supercomputer for training deep neural networks. *Communications of the ACM*, 63(7):67–78, 2020. 4
- [33] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. 1, 3, 4, 7
- [34] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallochi, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 1, 4, 8
- [35] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and

- Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. 5, 7
- [36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2
- [37] Liunian Harold Li\*, Pengchuan Zhang\*, Haotian Zhang\*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2
- [38] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 7
- [39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 1, 4, 6
- [40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 5
- [41] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 6
- [42] Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, and Antonio Torralba. Learning to compose visual relations. In *Advances in Neural Information Processing Systems*, pages 23166–23178. Curran Associates, Inc., 2021. 1
- [43] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 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*, pages 423–439. Springer, 2022. 2, 7
- [44] Shilong Liu, ZhaoYang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5, 7, 1
- [45] Elman Mansimov, Emilio Parisotto, Jimmy Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *ICLR*, 2016. 2
- [46] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2, 2022. 1
- [47] OpenAI. Chatgpt, 2022. 3
- [48] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2
- [49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 3
- [50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 7
- [51] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 3
- [52] Vitali Petsiuk, Alexander E. Siemann, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A. Plummer, Ori Kerret, Tonio Buonasassi, Kate Saenko, Armando Solar-Lezama, and Iddo Drori. Human evaluation of text-to-image models on a multi-task benchmark, 2022. 2
- [53] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74 – 93, 2015. 1, 4
- [54] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 2, 1, 3
- [55] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 4, 8
- [56] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 2
- [57] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2
- [58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3
- [59] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2
- [60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1, 2
- [61] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models, 2022. 1, 2
- [62] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *arXiv preprint arXiv:2306.08877*, 2023. 7
- [63] Scott Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, 2016. 2
- [64] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The*

- 33rd International Conference on Machine Learning*, pages 1060–1069, New York, New York, USA, 2016. PMLR. 2
- [65] Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. Scaling up models and data with t5x and seqio, 2022. URL <https://arxiv.org/abs/2203.17189>. 4
- [66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 3
- [67] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 7
- [68] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 1, 2
- [69] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 3
- [70] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 7
- [71] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society, 2015. 3
- [72] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models, 2023. 2
- [73] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*, 2018. 3
- [74] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imageward: Learning and evaluating human preferences for text-to-image generation, 2023. 1, 3, 4
- [75] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. 2018. 2
- [76] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yu-mao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, 2022. 2
- [77] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. 2023. 2
- [78] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roei Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*, 2023. 1, 3, 5, 4
- [79] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023. 2, 6
- [80] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 5
- [81] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunnar Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 1, 2
- [82] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, 2019. 2
- [83] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020. 7
- [84] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 4, 8, 9
- [85] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 1, 4

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

## Supplementary Material

### A. Additional Results

In Figure 8, we present additional examples comparing our model-predicted Textual-Visual feedback against baseline competitors. We would like to emphasize that our PaLI fine-tuned models produce the full feedback end-to-end, including the bounding box. In contrast, for the other models, we are required to use the GroundingDino [44] model to complete a full feedback for them, as they don’t have the ability to output bounding boxes.

Further qualitative results on our SeeTRUE-Feedback are available at Figure 9.

Figure 10 presents another set of examples showcasing our model’s performance in an ‘in-the-wild’ setting using images generated by state-of-the-art text-to-image models: Stable-Diffusion XL [54], StableDiffusion 2.1 [66], and Adobe Firefly [1].

### B. ConGen-Feedback Details

As detailed in the paper, to generate our training data using ConGen-Feedback, we collected aligned text-image pairs from six different datasets: PickaPic [33], ImageReward [74], COCO [39], Flickr30k [53], Open Images [34], and ADE20k [84]. It’s important to note that we selected only the positive pairs from these datasets, totaling 1,302,201 text-image pairs. From these pairs, we generated negative examples with misalignments accompanied by textual and visual feedback. To validate the generation and maintain a high quality of generated data, we filtered out samples based on textual entailment tests, as detailed in Appendix B.2, resulting in 1,087,912 text-image pairs for our filtered ConGen-Feedback version.

#### B.1. LLM Prompts

As presented in Brown et al. [7], we also employ the few-shot Language Model (LLM) prompting technique. For each aligned image-caption pair, we generate a contradiction caption, a detailed explanation (feedback) of the misalignment between the original caption and its generated contradiction, a misalignment cue pointing to the contradiction source in the caption, and a textual label for the visual bounding box for the image.

Our ConGen-Feedback training dataset is compiled from six different datasets. To ensure the highest quality of generated data, we use a few-shot prompt for each misalignment type we want to generate, as well as per dataset, considering the varied caption styles across datasets.

For images in COCO and Flickr30K with human-annotated captions, we choose the longest caption from

multiple options, as it is usually the most descriptive. An example prompt for creating a *Relation* type misalignment in the Coco dataset is in Figure 13

The captions for PickaPic and ImageReward were predicted by the image captioning PaLI-2 model. For each image, five captions were generated, and we selected the caption with the highest score produced by the model. In Figure 14, we present the prompt used for the PickaPic dataset to generate an *Action* misalignment.

ADE20k and OpenImages captions are taken from the Localized Narratives dataset. The dataset creators asked annotators to describe an image with their voice while simultaneously hovering their mouse over the region they are describing. This type of caption results in a small paragraph per image describing what can be seen in it, but not a short and concise caption summarizing the main idea in the image. Therefore, for these two datasets, we added another step before generating our labels. We generated a summarized caption from the image description and then applied the same procedure detailed before to this summarized caption. The prompt used for description summarizing is shown at Appendix D, and the prompt used for *Attribute* misalignment type for the ADE20k dataset is available at Figure 16.

All our data is generated by PaLM-2 LLM API, the chosen parameters are: model=“chat-bison@001”, temperature=0.4, max-tokens=700, top-p=0.95 and top-k=30. The entire prompts and data-generation pipeline will be released and available on our project page.

#### B.2. Generation Validation

Although we strive to generate our data with the highest possible quality, there is still a potential for noise generation caused by LLM hallucination behavior or insufficient sample quality. To assess the quality of generated samples, we evaluate two crucial aspects: Contradiction Correctness and Feedback Correctness.

The Contradiction Correctness evaluation aims to assess whether the generated contradiction caption adequately contradicts the original one. Meanwhile, the Feedback Correctness evaluation checks whether the generated feedback accurately explains the misalignment between the original (aligned) image caption and the generated contradiction caption.

Textual Entailment [19] involves classifying whether a *hypothesis* text is *entailed* by a premise text. We employ an entailment model by Honovich et al. [29] to evaluate and filter our generated data. For the Contradiction Correctness evaluation, the input for our model includes the origi-

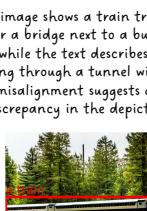
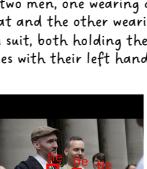
|                                     | mPlug-Owl 7B  | LLaVA 1.5 7B   | MiniGPT V2   | InstructBLIP   | PaLI 55B ft.  |   |
|-------------------------------------|---|--|--|--|---|---|
| Input                               |    |   |   |    |    |    |
| Predicted Textual & Visual Feedback | "The misalignments between the image and the text are that the green cup is on the left side, while the text is on the right side"  | "The green cup is not on the phone"  | "There is a green cup on the left, and a blue cell phone on the right"   | "Phone and cup are misaligned"   | "The cup is blue, not green"  |   |
| Input                               |    |   |   |    |    |    |
| Predicted Textual & Visual Feedback | "The image shows a train traveling under a bridge next to a bunch of trees, while the text describes a train traveling through a tunnel with trees. This misalignment suggests a slight discrepancy in the depiction" | "The train is not traveling under a bridge"  | "The image is of a train traveling underneath a bridge with trees all around it, while the text describes a bunch of trees"                  | "train"  | "The train is traveling over the bridge, not under it"                                |   |
| Input                               |   |    |    |   |   |   |
| Predicted Textual & Visual Feedback | "The man in the image is holding his tie with his left hand, but the text mentions him holding his tie with his right hand."  | "The man is wearing a suit and tie"  | "The image is a picture of two men, one wearing a hat and the other wearing a suit, both holding their ties with their left hands"           | "Left"   | "The man is holding his tie with his right hand, not his left hand"                   |   |
| Input                               |    |   |   |  |  |  |
| Predicted Textual & Visual Feedback | "The supercars are parked in front of a new building"   | "The misalignment is that the supercars, which include a Ferrari and a Lamborghini, are parked in front of a brick building instead of in front of a new building" | "The image shows three supercars parked in front of a building, but the text says they are parked on a street"                               | "Building"   | "The supercars are parked in front of an old building, not a new building"            |   |
| Input                               |    |   |   |  |  |  |
| Predicted Textual & Visual Feedback | "An arrangement of tulips is set in an urn style vase"  | "The image shows a vase of flowers, but the text says tulips"  | "The text describes an arrangement of tulips in an urn-style vase, while the image shows an arrangement of flowers in a vase with a red bow" | "vase"   | "no"  | "The arrangement is of roses, not tulips"   |

Figure 8. Additional qualitative comparison examples from SeeTRUE-Feedback. The PaLI 55B model, fine-tuned on TV-Feedback, produces the most aligned and concise feedback compared to the rest of the models.



Figure 9. Additional qualitative results on SeeTRUE-Feedback. Please note the precision of our produced Textual-Visual feedback, which includes: a concise explanation of the misalignment, a misalignment cue that pinpoints the contradictory source in the caption, and a labeled bounding box.

| Source   | SD XL  | SD 2.1   | SD XL   | Adobe Firefly   |
|--|--|--|---|---|
| <b>Input Prompt</b>                            | "A blue parrot is perched on the head of a black dog"                | "A photo of an astronaut holding balloons in his hand and wearing a cowboy hat." | "A group of friends holding deflated balloons, their faces painted with disappointment" | "A horse riding on top of an astronaut"                                 |
| <b>Generated Image</b>                         |  |  |   |   |
| <b>Predicted Textual &amp; Visual Feedback</b> | <br>The parrot is perched on a table, not on the head of a black dog | <br>The astronaut is not wearing a cowboy hat, instead is wearing a space helmet | <br>The balloons are not deflated, instead are inflated                                 | <br>The astronaut is riding on top of a horse, not the other way around |

Figure 10. Additional qualitative of out-of-distribution results using synthetic images generated using Stable-Diffusion XL [54], StableDiffusion 2.1 [66] and Adobe Firefly [1].

nal (positive) caption as the premise and the generated contradiction as the hypothesis. We anticipate receiving a *low* entailment score (NLI) in this evaluation, emphasizing that the negative caption truly differs from the original caption.

Feedback Correctness is evaluated by feeding the entailment model with the original image caption with the generated contradiction (premise) and the generated feedback (hypothesis). In this test, our goal is to receive a *high* entailment score, indicating that the feedback remarks the misalignment between the original caption and the generated contradiction. A template for our evaluation tasks and examples of generated data with their evaluated entailment scores are shown in Figure 11.

Finally, we filter out from our initial dataset generated samples based on their evaluation. Samples with a Contradiction Correctness score higher than 0.25 and samples with a Feedback Correctness score lower than 0.75 are filtered out. In Figure 12, we present a heatmap of the remaining percentage of our generated dataset filtered by a range of thresholds. The selected thresholds for our released ConGen-Feedback filtered version are  $< 0.25$  for Contradiction Correctness score and  $> 0.75$  for Feedback Correctness score.

## C. SeeTRUE-Feedback Details

Our SeeTRUE-Feedback benchmark is constructed based on the SeeTRUE dataset [78], which contains aligned and misaligned text-image pairs. Each misaligned pair is annotated with three human-generated misalignment feedbacks.

Similar to the approach detailed in B.1, we utilize PaLM 2 to generate the required labels. In this case, there is no need to generate a contradiction and feedback for a sample. Instead, we need to summarize the three human feedbacks into a single and concise feedback and generate the remaining labels, such as the misalignment cue source in the caption and the textual label for the visual bounding box.

The few-shot prompt used for SeeTRUE-Feedback generation is presented in Figure 17.

## D. Training details

We provide details on the fine-tuning process for PaLI models. We fine-tune PaLI using T5X [65] on JAX [6]. The fine-tuning process involves utilizing the Adam optimizer with a learning rate of 2e-5. The model undergoes two epochs of training, and 10% of the training set is reserved as a validation set for checkpoint selection.

For the supervised experiments, we fine-tuned the PaLI models with the visual question answering (VQA) task. This fine-tuning process aims to enhance the model’s performance in specific tasks related to VQA, such as image-text alignment and providing textual and visual feedback.

The fine-tuning tasks include:

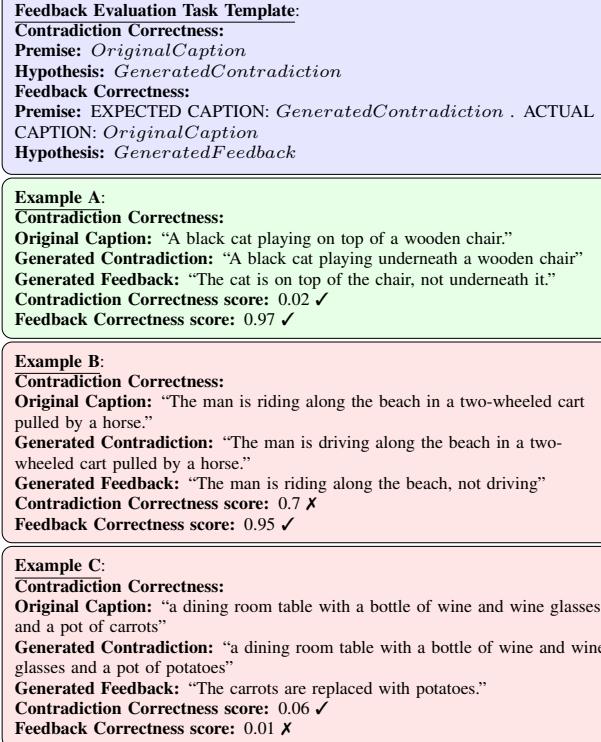


Figure 11. Automatic feedback evaluation. The blue sub-figure presents our evaluation template using the Honovich et al. [29] entailment model to calculate Contradiction Correctness and Feedback Correctness scores. The next three sub-figures are examples of generated data samples with their calculated scores. While example A (green sub-figures) passed our selected thresholds as good quality generated samples, examples B and C (red sub-figures) are filtered out. B is filtered out because its Contradiction Correctness score is higher than 0.25, while C is filtered out because the received Feedback Correctness score is lower than 0.75.

- 1. Image-Text Alignment:** In this task, we utilize the same query as in the zero-shot setup, “*Does this image entail the description <text>?*”. The model is expected to provide a binary response, either ‘yes’ or ‘no’, indicating whether the image aligns with the given text description.
- 2. Textual and Visual Feedback:** This task involves introducing a query for combined feedback: “*What are the misalignments between this image and the text <text>?*”. The expected response format comprises three components: feedback, misalignment in text, and misalignment in the image (bounding-box). The aim is to extract detailed feedback and specific misalignment indicators in a single model interaction.

**Hardware Requirements:** The computational resources required for fine-tuning vary depending on the specific PaLI model. PaLI 5B and 17B fine-tuning require four v4 chips [32], while fine-tuning for PaLI 55B necessitates sixteen v4 chips.

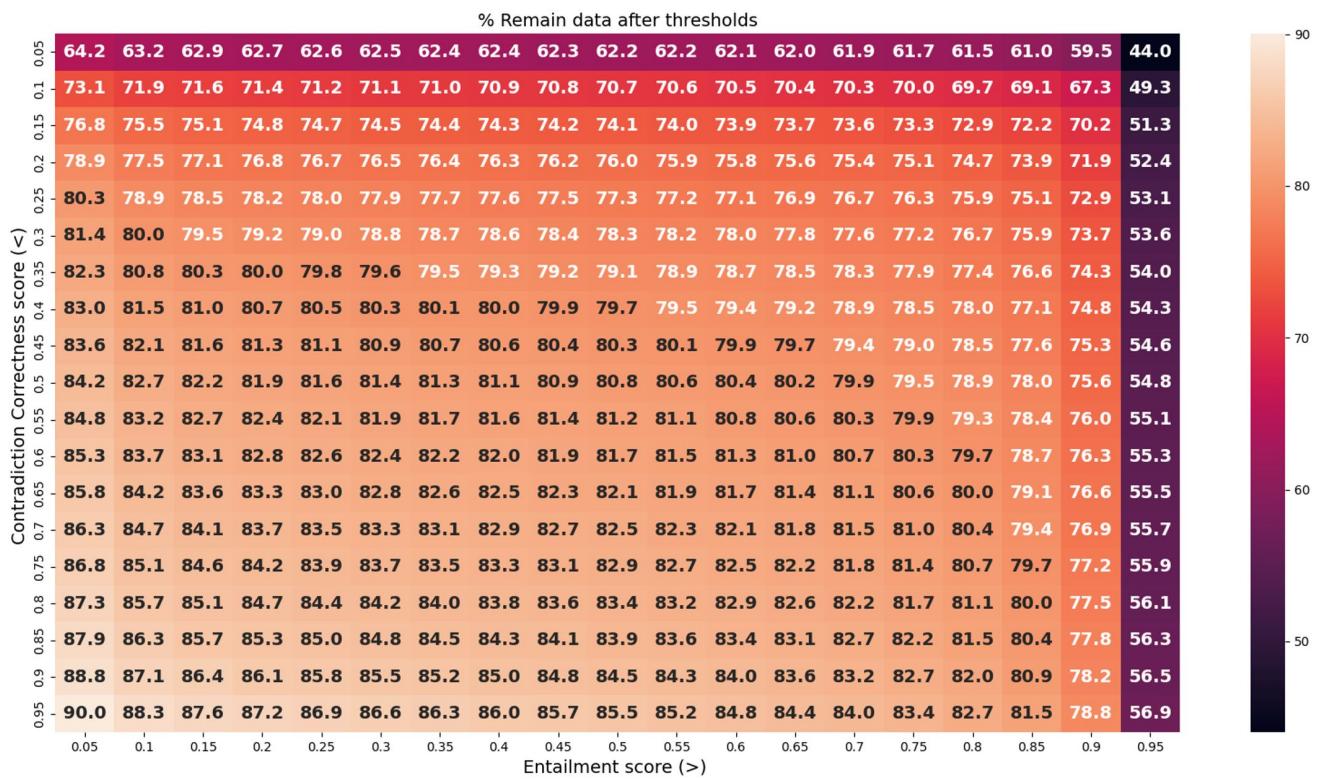


Figure 12. A heatmap overlay presenting the percentage of remaining data from the entire set generated after applying a range of thresholds for Contradiction Correctness and Feedback Correctness scores.

---

### Context

Your primary task is to generate subtly contradictory captions based on an original input caption. This test is intended to observe the capacity of models to recognize nuanced semantic changes which might be easily detected by humans but challenging for models.

#### Guidelines:

Caption Modification: Introduce a delicate semantic change to the original caption. The alteration should subtly change the meaning while remaining potentially overlooked. It should neither be drastic nor too minor.

MISALIGNMENT: After constructing your contradictory caption, articulate the disparity between the original CAPTION and the CONTRADICTION in a crisp manner. The elucidation should:

Spotlight the deviation from the angle of the CONTRADICTION, pinpointing what the CONTRADICTION suggests but is countered by the CAPTION.

For the CAPTION description, ensure it thoroughly represents the true content of the CAPTION (e.g., “CAPTION: a blue ball on a shelf”).

For the CONTRADICTION description, keep it brief and directly replicate the relevant segment (e.g., “CONTRADICTION: green ball”).

MISALIGNMENT TYPE: Categorize the nature of the misalignment into one of the following categories: Object/Noun, Attribute/Adjective, Action/Verb, or Relation

---

### Few-Shot

#### Guidelines:

- Relation Misalignment: Change the relationship or position between objects. This entails tweaking prepositions, positioning, or context which defines the relationship between the mentioned items. In the misalignment description, specifically indicate the changed relation and the affected subjects.- Mention at the SOURCE fields the changed preposition and related subject

#### Examples:

CAPTION: “A laptop next to a stack of books on a wooden desk”

CONTRADICTION: A laptop under a stack of books on a wooden desk.

MISALIGNMENT: The laptop is next to the books (CAPTION: laptop next to a stack of books), not under them (CONTRADICTION: laptop under a stack of books).

MISALIGNMENT TYPE: Relation

CAPTION: Two cats are sitting below a branch of a tree in front of a building with windows and trees in the background .

CONTRADICTION: Two cats are sitting above a branch of a tree in front of a building with windows and trees in the background .

MISALIGNMENT: The cats are in front of the building sitting below a branch (CAPTION: Two cats sitting below a branch), not above it (CONTRADICTION: Two cats sitting above a branch)

MISALIGNMENT TYPE: Relation

CAPTION: portrait of a redhead girl with blue eyes in a violet color t shirt sitting on a yellow color sofa near a gray wall

CONTRADICTION: portrait of a redhead girl with blue eyes in a violet color t shirt sitting on a yellow color sofa apart from a gray wall

MISALIGNMENT: The sofa is near the wall (CAPTION: a yellow sofa near a gray wall), not apart from it (CONTRADICTION: a yellow sofa apart from a gray wall)

MISALIGNMENT TYPE: Relation

CAPTION: Old fashioned chests are stacked atop each other on a city block.

CONTRADICTION: Old fashioned chests are stacked next to each other on a city block.

MISALIGNMENT: The chests are stacked atop each other (CAPTION: chests stacked atop each other), not next to each other (CONTRADICTION: chests stacked next to each other).

MISALIGNMENT TYPE: Relation

CAPTION: Boy holding large stuffed toy bear outside of building.

CONTRADICTION: Boy holding large stuffed toy bear inside of building.

MISALIGNMENT: The boy is outside of a building (CAPTION: Boy holding large stuffed toy bear outside of building), not inside of it (CONTRADICTION: boy holding large stuffed toy bear inside of building) .

MISALIGNMENT TYPE: Relation

#### Create a MISALIGNMENT of type: Relation

Remember when writing the MISALIGNMENT that the expectation is that CONTRADICTION should happen, but instead we got the scene of CAPTION.

We expect for CONTRADICTION but CAPTION counters it.

CAPTION: A crystal bowl filled with oranges on top of a table.

---

### Generated

CONTRADICTION: A crystal bowl filled with oranges beneath a table.

MISALIGNMENT: The bowl is on top of the table (CAPTION: bowl on top of a table), not beneath it (CONTRADICTION: bowl beneath a table).

MISALIGNMENT TYPE: Relation

---

Figure 13. Few-shot prompt example for generating a *Relation* misalignment in the COCO [39] dataset.

---

## Context

Your primary task is to generate subtly contradictory captions based on an original input caption. This test is intended to observe the capacity of models to recognize nuanced semantic changes which might be easily detected by humans but challenging for models.

Guidelines:

Caption Modification: Introduce a delicate semantic change to the original caption. The alteration should subtly change the meaning while remaining potentially overlooked. It should neither be drastic nor too minor.

MISALIGNMENT: After constructing your contradictory caption, articulate the disparity between the original CAPTION and the CONTRADICTION in a crisp manner.

The elucidation should:

Spotlight the deviation from the angle of the CONTRADICTION, pinpointing what the CONTRADICTION suggests but is countered by the CAPTION.

For the CAPTION description, ensure it thoroughly represents the true content of the CAPTION (e.g., "CAPTION: a blue ball on a shelf").

For the CONTRADICTION description, keep it brief and directly replicate the relevant segment (e.g., "CONTRADICTION: green ball").

MISALIGNMENT TYPE: Categorize the nature of the misalignment into one of the following categories: Object/Noun, Attribute/Adjective, Action/Verb, or Relation

---

## Few-Shot

Guidelines:

- Action/Verb Misalignment: Modify a central action or verb in the original caption. This change should introduce a contradiction in terms of the activity or behavior described. The changed verb should make sense within the context and should neither be too drastic nor too obscure. Ensure the difference between the original caption's action and the contradictory action is clearly and succinctly articulated in the misalignment description.

Examples:

CAPTION: "A person dressed as Captain America is holding a shield in front of a wall."

CONTRADICTION: A person dressed as Captain America is throwing a shield and standing in front of a wall.

MISALIGNMENT: The person is not throwing a shield (CONTRADICTION: person throwing a shield), instead is holding it (CAPTION: person holding a shield)

MISALIGNMENT TYPE: Action/Verb

CAPTION: "a close up of a lion with a long mane sitting on the grass with trees in the background"

CONTRADICTION: a close up of a lion with a long mane drinking water on the grass in the forest with trees in the background

MISALIGNMENT: The lion is sitting (CAPTION: a lion sitting), not jumping on the grass (CONTRADICTION: lion drinking water)

MISALIGNMENT TYPE: Action/Verb

CAPTION: "A painting of a man standing in front of a fire in a field with trees and sky in the background ."

CONTRADICTION: A painting of a man running in front of a fire in a field with trees and a cloudy sky .

MISALIGNMENT: The main the painting is standing (CAPTION: man standing in front of a fire), not running (CONTRADICTION: man running in front of a fire)

MISALIGNMENT TYPE: Action/Verb

CAPTION: "A tattooed woman in a purple brassiere and purple underpants is standing in front of a table with a bottle on it."

CONTRADICTION: A tattooed woman in a purple brassiere and purple underpants is dancing in front of a table with a bottle on it.

MISALIGNMENT: The woman is not dancing (CONTRADICTION: tattooed woman dancing), instead is sitting (CAPTION: A tattooed woman is standing)

MISALIGNMENT TYPE: Action/Verb

CAPTION: "portrait of a smiling man and woman hugging each other on the street with trees and buildings in the background"

CONTRADICTION: portrait of a smiling man and woman fighting on the street with trees and buildings in the background

MISALIGNMENT: The man and woman are not fighting each other (CONTRADICTION: man and woman fighting), instead are hugging (CAPTION: a smiling man and woman hugging each other)

MISALIGNMENT TYPE: Action/Verb

CAPTION: A black background with a picture of a guy in a red hat jumping with a skate board.

CONTRADICTION: A black background with a picture of a guy in a red hat drinking water with a skate board.

MISALIGNMENT: The guy is not drinking water (CONTRADICTION: guy drinking water), instead is sitting (CAPTION: guy sitting)

MISALIGNMENT TYPE: Action/Verb

Create a MISALIGNMENT of type: Action/Verb

Remember when writing the MISALIGNMENT that the expectation is that CONTRADICTION should happen, but instead we got the scene of CAPTION. We expect for CONTRADICTION but CAPTION counters it.

CAPTION: A white cat is sitting on the grass in front of a house with trees and a blue sky in the background

---

## Generated

CAPTION: A white cat is sitting on the grass in front of a house with trees and a blue sky in the background .

CONTRADICTION: A white cat is running on the grass in front of a house with trees and a blue sky in the background.

MISALIGNMENT: The cat is not running (CONTRADICTION: cat running), instead is sitting on the grass (CAPTION: cat sitting)

MISALIGNMENT TYPE: Action/Verb

---

Figure 14. Few-shot prompt example for generating an *Action* misalignment in the PickaPic [33] dataset.

---

## Context

---

Task Definition: Extracting Concise Image Captions

Input:

DESCRIPTION: A long description of an image containing various objects, people, and scene details.

Output:

CAPTION: A single-sentence caption that preserve all the information from the description.

Guidelines:

- The goal is to create a concise caption that effectively communicates the image's information written as a caption not a description.
  - Include what people are doing and wearing from the description.
  - Describe where objects are and how they look from the description.
- 

## Few-Shot

---

Examples:

DESCRIPTION: Front we can see ball. Background it is blur. We can see trees, people and sky.

CAPTION: A ball with a blur background that includes trees, people and sky

DESCRIPTION: There is a poster in which, there is a image. In the image, there are two vehicles parked on the grass on the ground. In the background, there are trees. Below this image, there are texts and watermark on the white page.

CAPTION: An image of two vehicles parked on the grass, below the image there is a text.

DESCRIPTION: In this picture I can see food items, fork and a bowl with a liquid in it, on the plate, on an object.

CAPTION: Food items, fork and a bowl with liquid in it are on a plate.

DESCRIPTION: This image is taken outdoors. At the top of the image there is the sky. At the bottom of the image there is a runway. In the background there are a few buildings and there is a ground with grass on it. In the middle of the image there are two airplanes on the runway and there is a pole. There are two objects on the runway.

CAPTION: Two airplanes on an outdoors runway with buildings and sky on the background

DESCRIPTION: Here in this picture we can see an air plane present in an airport over there and we can also see a couple of trucks present beside it over there and we can see people standing on the ground over there and in the far we can see trees and plants present over there.

CAPTION: An airplane in an airport with trucks beside it, there are people standing on the ground.

DESCRIPTION: In this picture there is a view of the airport. In the front there is a white color tower and some flights parked on the ground. In the background there are some buildings. On the top we can see the plane flying in the sky.

CAPTION: An airport view, at the foreground a white color tower and planes parked on the ground. Building are present in the background and a plane is flying in the sky

DESCRIPTION: This picture is a black and white image. In this image we can see one woman with black sunglasses walking, some plants near the building, one big pole, some text on the bottom left side corner of the image, some grass on the ground, one object and one wire attached to the pole.

CAPTION: A black and white image of a woman walking with black sunglasses, a building is present also with plants near to it

DESCRIPTION: In this picture I can observe a man and woman. Woman is holding a paper in her hand. In the background I can observe a building and plants.

CAPTION: A man with a woman who is holding a paper in her hand, there is a building and plants at the background.

DESCRIPTION: In this image I can see few women sitting on the concrete wall and I can see a dog which is brown, cream and black in color is sitting on the ground. In the background I can see a person standing, the white colored cloth, few plants and a table, and on the table I can see a white colored cloth.

CAPTION: Women sitting on the concrete wall outside a building with a brown, cream and black dog. At the background there is plants, a table and a person wearing white colored cloth

DESCRIPTION: In this picture we can see a flower vase and a name board on the platform and here we can see four people are standing on the floor. In the background we can see the name on the wall and we can see plants, roof and lights.

CAPTION:

---

Generated

CAPTION: People standing on the floor near a flower vase and a name board.

---

Figure 15. Few-shot prompt example for generating a short and concise caption from a lengthy image description annotation for the Localized Narratives [55] datasets: ADE20k [84] and OpenImages [34].

---

## Context

Your primary task is to generate subtly contradictory captions based on an original input caption. This test is intended to observe the capacity of models to recognize nuanced semantic changes which might be easily detected by humans but challenging for models.

### Guidelines:

Caption Modification: Introduce a delicate semantic change to the original caption. The alteration should subtly change the meaning while remaining potentially overlooked. It should neither be drastic nor too minor.

MISALIGNMENT: After constructing your contradictory caption, articulate the disparity between the original CAPTION and the CONTRADICTION in a crisp manner. The elucidation should:

Spotlight the deviation from the angle of the CONTRADICTION, pinpointing what the CONTRADICTION suggests but is countered by the CAPTION.

For the CAPTION description, ensure it thoroughly represents the true content of the CAPTION (e.g., "CAPTION: a blue ball on a shelf").

For the CONTRADICTION description, keep it brief and directly replicate the relevant segment (e.g., "CONTRADICTION: green ball").

MISALIGNMENT TYPE: Categorize the nature of the misalignment into one of the following categories: Object/Noun, Attribute/Adjective, Action/Verb, or Relation

---

## Few-Shot

### Guidelines:

- Attribute/Adjective Misalignment: Alter an attribute or adjective that describes an object or scene in the original caption. This alteration should create a contradiction in the description. Ensure that the changed attribute/adjective remains plausible within the context of the caption and doesn't alter the core meaning. The discrepancy between the original caption's attribute and the contradictory attribute should be highlighted in the misalignment description.

### Examples:

CAPTION: "a waterfall in a tropical forest with green plants and trees"

CONTRADICTION: a waterfall in a snowy forest with green plants and trees

MISALIGNMENT: The forest is not snowy (CONTRADICTION: snowy forest), instead is a tropical one (CAPTION: tropical forest)

MISALIGNMENT TYPE: Attribute/Adjective

CAPTION: "a cartoon of a red haired mermaid holding a fish in the sea"

CONTRADICTION: a cartoon of a mermaid with black hair in the water

MISALIGNMENT: The hair of the cartoon mermaid is red (CAPTION: red haired mermaid), not black (CONTRADICTION: mermaid with black hair)

MISALIGNMENT TYPE: Attribute/Adjective

CAPTION: portrait of a naked woman with long hair on a grey background

CONTRADICTION: A portrait of a naked woman with short hair on a grey background

MISALIGNMENT: The woman has long hair (CAPTION: long hair), not short hair (CONTRADICTION: short hair)

MISALIGNMENT TYPE: Attribute/Adjective

CAPTION: The colorful flowers are in the vase next to the window.

CONTRADICTION: The white and black flowers are in the vase next to the window.

MISALIGNMENT: The flowers are colorful (CAPTION: colorful flowers), not white and black (CONTRADICTION: white and black flowers)

MISALIGNMENT TYPE: Attribute/Adjective

CAPTION: Vase sitting next to the window with droopy flowers in it.

CONTRADICTION: Vase sitting next to the window with fresh flowers in it.

MISALIGNMENT: The vase has droopy flowers (CAPTION: droopy flowers), not fresh flowers (CAPTION: fresh flowers)

MISALIGNMENT TYPE: Attribute/Adjective

CAPTION: A bicycle chained to a red pole and a green pole on the sidewalk in front of a building.

CONTRADICTION: A bicycle chained to a red pole and a blue pole on the sidewalk in front of a building.

MISALIGNMENT: One of the poles is red (CAPTION: green pole), not blue (CONTRADICTION: blue pole)

MISALIGNMENT TYPE: Attribute/Adjective

Create a MISALIGNMENT of type: Attribute/Adjective

Remember when writing the MISALIGNMENT that the expectation is that CONTRADICTION should happen, but instead we got the scene of CAPTION.

We expect for CONTRADICTION but CAPTION counters it.

CAPTION: A wall with light and dust particles under a wooden beam.

---

## Generated

CONTRADICTION: A wall with light and dust particles under a metal beam.

MISALIGNMENT: The beam is wood (CAPTION: wooden beam), not metal (CONTRADICTION: metal beam)

MISALIGNMENT TYPE: Attribute/Adjective

---

Figure 16. Few-shot prompt example for generating an *Attribute* misalignment in the ADE20k [84] dataset.

---

## Context

Your primary task is to generate a single summarized feedback based on visual feedback provided in the input, and identify the source of misalignment between the caption and the visual content. The goal is to extract meaningful insights from the feedback and link them to the corresponding elements in the caption.

### Guidelines:

**SUMMARIZED FEEDBACK:** Given a caption and a set of visual feedbacks, your objective is to create a concise and informative feedback statement that captures the most significant misalignment between the caption and the visual content. No need to mention "at the caption" or "at the image".

**Misalignment Identification:** After constructing the summarized feedback, you should indicate the source of misalignment for both the caption and the visual content. In parentheses, specify the origin of the misalignment as follows:

**CAPTION:** Describe the source of misalignment that originates from the text (e.g., a description or detail in the caption that does not match the image). Mention also the related subject.

**CONTRADICTION:** Based on the feedback, pinpoint the source of misalignment in the visual content that contradicts the caption. Make sure the CONTRADICTION refers only to the relative part in FEEDBACKS. Mention also the related subject.

---

## Few-Shot

### Examples:

**CAPTION:** "A woman with a blue shirt and yellow flower headband, sitting on a wooden bench outside on the grass."

**FEEDBACKS:** ["The woman's shirt is yellow, not blue.", "The woman sitting on the wooden bench is wearing a white shirt, not a blue shirt.", "The woman has a yellow shirt on, not blue."]

**MISALIGNMENT:** The sitting woman is wearing a yellow shirt (CONTRADICTION: yellow shirt), not a blue shirt (CAPTION: blue shirt)

**CAPTION:** A couch on the left of a chair.

**FEEDBACKS:** ["A couch is to the left of a coffee table, there is no chair present.", "A couch and two tables under a window, there is no chair.", NaN]

**MISALIGNMENT:** The couch is to the left of a table (CONTRADICTION: coffee table), not left of a chair (CAPTION: chair).

**CAPTION:** A sheep to the right of a wine glass.

**FEEDBACKS:** ["A sheep is facing the camera, not to the right of a wine glass, which is not present.", "A sheep looking at the camera, there is no wine glass.", "There is only a sheep alone in a field, it is not to the right of a wine glass"]

**MISALIGNMENT:** There is no wine glass (CAPTION: wine glass) (CONTRADICTION: no wine glass)

**CAPTION:** A tomato has been put on top of a pumpkin on a kitchen stool. There is a fork sticking into the pumpkin. The scene is viewed from above.

**FEEDBACKS:** [NaN, "A pumpkin has been put on top of a kitchen stool. There is a fork sticking into the pumpkin. The scene is viewed from above. There is no tomato on top of the pumpkin.", "There is only a pumpkin, and no tomato"]

**MISALIGNMENT:** There is no tomato on top of a pumpkin (CAPTION: tomato on top of a pumpkin) (CONTRADICTION: no tomato)

**CAPTION:** a red glass ball with a reflection on top of a gray surface

**FEEDBACKS:** ["The glass ball is blue, not red.", "The ball with the reflection should be blue instead of red.", "The ball is blue, not red"]

**MISALIGNMENT:** The color of the ball is blue (CONTRADICTION: blue ball), not red (CAPTION: red ball)

**CAPTION:** A lunch salad in a yellow bowl made out of fruit, vegetables, and meats with chopsticks.

**FEEDBACKS:** ["There is no meat present, nor are there any chopsticks, simply the bowl with fruit and vegetable.", "There is a bowl of food, but no chopsticks.", "There is a salad with fruit and vegetables in a yellow bowl but there does not appear to be meat in it and there are no chopsticks."]

**MISALIGNMENT:** There is a yellow bowl with fruit and vegetables (CONTRADICTION: salad with fruit and vegetables in a yellow bowl) but without meat, also no chopsticks present (CAPTION: salad made out of fruit, vegetables, and meats with chopsticks)

**CAPTION:** A cup filled with umbrellas and canes next to a white wall.

**FEEDBACKS:** ["The wall is gray, not white.", "A bucket has two canes in it, not umbrellas and canes and it is front of a gray wall, not a white wall.", "A cup has canes in it, but no umbrellas"]

**MISALIGNMENT:** The cup has canes in it (CAPTION: umbrellas and canes), but no umbrellas (CONTRADICTION: no umbrellas). The wall is not white (CAPTION: white tall), instead is gray (CONTRADICTION: gray wall).

**CAPTION:** Small white toilet with spare rolls of toilet paper beside it.

**FEEDBACKS:** ["A small white toilet is here without any rolls of toilet paper, not any spares.", "There is a toilet but no toilet paper.", "The toilet only has curtains/walls next to it, not spare rolls of toilet paper."]

**MISALIGNMENT:** The toilet does not have any spare rolls of toilet paper beside it (CONTRADICTION: NO toilet paper), as described in the caption (CAPTION: spare rolls of toilet paper)

**CAPTION:** A cat is holding a frisbee in its mouth

**FEEDBACKS:** ["A dog is holding a frisbee in its mouth", "A dog is holding a frisbee in its mouth.", "A dog is holding a frisbee, not a cat"]

**MISALIGNMENT:**

---

## Generation

**CAPTION:** A cat is holding a frisbee in its mouth

**FEEDBACKS:** ["A dog is holding a frisbee in its mouth", "A dog is holding a frisbee in its mouth.", "A dog is holding a frisbee, not a cat"]

**MISALIGNMENT:** The animal holding the frisbee is a dog (CONTRADICTION: dog holding a frisbee), not a cat (CAPTION: cat holding a frisbee)

---

Figure 17. Few-shot prompt example for generating the required labels for our SeeTRUE-Feedback test dataset from human-annotated feedbacks.