
VisMin: Visual Minimal-Change Understanding

Rabiul Awal* Saba Ahmadi* Le Zhang* Aishwarya Agrawal
Mila - Quebec AI Institute
Université de Montréal
{rabiul.awal,le.zhang,aishwarya.agrawal}@mila.quebec

Abstract

Fine-grained understanding of objects, attributes, and relationships between objects is crucial for visual-language models (VLMs). To evaluate VLMs’ fine-grained understanding, existing benchmarks primarily focus on evaluating VLMs’ capability to distinguish between two very similar *captions* given an image. In this paper, our focus is on evaluating VLMs’ capability to distinguish between two very similar *images* give a caption. To this end, we introduce a new, challenging benchmark termed **Visual Minimal-Change Understanding (VisMin)**, which requires models to predict the correct image-caption match given two images and two captions. Importantly, the image pair (as well as the caption pair) contains minimal-changes, i.e., between the two images (as well as between the two captions), only one aspect changes at a time from among the following possible types of changes: *object*, *attribute*, *count*, and *spatial relation*. These four types of minimal-changes are specifically designed to test the models’ understanding of objects, attributes of objects (such as color, material, shape), counts of objects and spatial relationship between objects. To curate our benchmark, we built an automatic framework using large language models and diffusion models, followed by a rigorous 4-step verification process by human annotators. Empirical experiments reveal that current VLMs exhibit notable deficiencies in understanding spatial relationships and counting abilities. Furthermore, leveraging the automated nature of our data creation process, we generate a large-scale training dataset, which we use to finetune CLIP (a foundational VLM) and Idefics2 (a multimodal large language model). Our findings show that both these models benefit significantly from fine-tuning on this data, as evident by marked improvements in fine-grained understanding across a wide range of benchmarks. Additionally, such fine-tuning improves CLIP’s general image-text alignment capabilities too. We release all resources including the benchmark, the training data and the finetuned model checkpoints at <https://vismin.net/>.

1 Introduction

Fine-grained understanding of objects, attributes, and their relationships is critical for Visual-Language Models (VLMs) to generalize effectively to new, unseen scenes and compositions. Previous studies such as ARO [43] and Sugarcrepe [8], highlighting the deficiencies of VLMs in this domain predominantly focus on understanding fine-grained differences between two very similar *captions* – a human-written caption and an automatically generated hard-negative² caption, where the hard-negative caption differs from the original caption only with respect to an *object*, or an *attribute* or a *relationship* between two objects. While such hard-negative examples for *captions* can be synthesized using rule-based approaches, synthesizing such hard-negative examples for images is very

*denotes equal contribution

²In the context of contrastive learning, a hard-negative is a specific type of negative example that is particularly challenging to distinguish from the positive example.



Figure 1: Overview of our VisMin benchmark. VisMin consists of four types of minimal-changes – object, attribute, count and spatial relation – between two image-captions pairs. The evaluation task requires a model to predict the correct image-caption match given: 1) two images and one caption, 2) two captions and one image.

challenging. Existing benchmarks presenting *visual* hard-negatives suffer from two main limitations: 1) **Limited Difficulty**: In benchmarks such as Winoground [36], MMVP[37], the original images and their hard-negative counterparts differ in multiple aspects (objects, attributes of objects, image background, etc.). This multiplicity limits the difficulty of the benchmark and makes it challenging to precisely evaluate the models’ fine-grained understanding of specific aspects. 2) **Limited Complexity**: Although benchmarks such as EQBEN [38], SPEC [28] have controlled hard-negatives, the visual domain is limited to graphic engines, a few video domains or reliance on purely synthetic images depicting simplistic scenes.

Motivated by these observations, we propose a new benchmark, **Visual Minimal-Change Understanding (VisMin)**, built on top of the images from the COCO [22] dataset that consists of complex everyday scene images. VisMin is designed to measure VLMs’ ability to comprehend minimal changes, i.e., changes with respect to only one aspect (see Fig. 1), from among the following aspects: *object*, *attribute*, *count*, and *spatial relation*, while keeping other aspects unchanged as much as possible.

The evaluation task for a model is to predict the correct image-caption match given: 1) two images and one caption, 2) two captions and one image. To curate VisMin, we built an automated pipeline using large language models and diffusion models. To ensure the quality of our benchmark, the synthetic data generated using the automated pipeline undergoes a rigorous 4-step verification process by human annotators, with data retained in the benchmark only if it passes all four steps. We meticulously designed the benchmark ensuring uniformity across various categories to the extent possible. We conduct a detailed analysis of our benchmark, which enables a more transparent assessment of the various strengths and weaknesses of the models.

We conducted empirical tests on eight open-source VLMs, including foundational models like CLIP [31] and Multimodal Large Language Models (MLLMs) such as Llava[24] and Idefics2[15]. We also evaluated two closed-source APIs, GPT-4 and Gemini. Our findings suggest that both foundational models and MLLMs perform relatively well in understanding minimal changes in objects and attributes. Surprisingly, MLLMs underperform foundational VLMs in object and attribute understanding! For spatial relation understanding, although MLLMs perform better than VLMs, both families of models perform below random chance! Similarly, both families of models show considerable room for improvement in counting capabilities. Our results underscore the need for an emphasis on spatial reasoning and counting understanding over attribute/object recognition in VLM evaluations. We anticipate that our benchmark will catalyze advancements in these critical areas within the community.

Lastly, owing to the automated nature of our synthetic data creation process, we generated a large-scale (64,392 samples) minimal-change image-text data for fine-tuning the VLMs to enhance their fine-grained understanding. Fine-tuning CLIP (a foundational VLM) and Idefics2 (a MLLM) on our minimal-change data, without any additional modifications to the model architecture or loss functions, results in significant improvements in fine-grained understanding across various benchmarks. Notably, such fine-tuning also enhances foundational VLMs’ general image-text alignment capabilities as evident by marked improvements in CLIP’s image-text retrieval performance on COCO. These observations suggest that our minimal-change dataset can serve as model-agnostic, general-purpose resource to enhance the capabilities of VLMs.

To summarize, our contributions are threefold: 1) **A controlled and challenging benchmark.** We introduce the VisMin benchmark, which challenges models to detect semantic differences between visually similar but semantically different images. Extensive testing on foundational VLMs and MLLMs reveals their difficulties with this task, highlighting areas for improvement. 2) **A pipeline for automated data creation and benchmark development.** We create an automated pipeline to generate visual minimal-change data at scale using large language models and diffusion models, with a rigorous four-step human verification system to ensure high data quality. 3) **Enhancement of VLMs’ fine-grained understanding with fine-tuning on minimal-change data.** We improve the fine-grained understanding of CLIP and Idefics2 by fine-tuning them on our large-scale minimal-change image-text data, demonstrating improved image-text alignment and overall performance.

2 Related work

Fine-grained understanding benchmarks: Most existing benchmarks focus on understanding fine-grained textual differences, such as VL-checklist [47], ARO [43], and Sugarcrepe [8]. Benchmarks presenting visual hard-negatives, such as EQBEN [38], Winoground [36], ImageCode [14], SPEC [28], either lack minimal changes or have limited visual complexity – graphic engines, a few video domains or purely synthetic images depicting simplistic scenes. Our benchmark addresses these gaps by utilizing the advances in LLMs [10] and diffusion models [30, 20, 21] to achieve minimal changes in complex COCO-like scenes without compromising the naturalness of the images, thus providing a more robust evaluation of fine-grained visual understanding in VLMs. Detailed comparisons of benchmarks are provided in section 4.

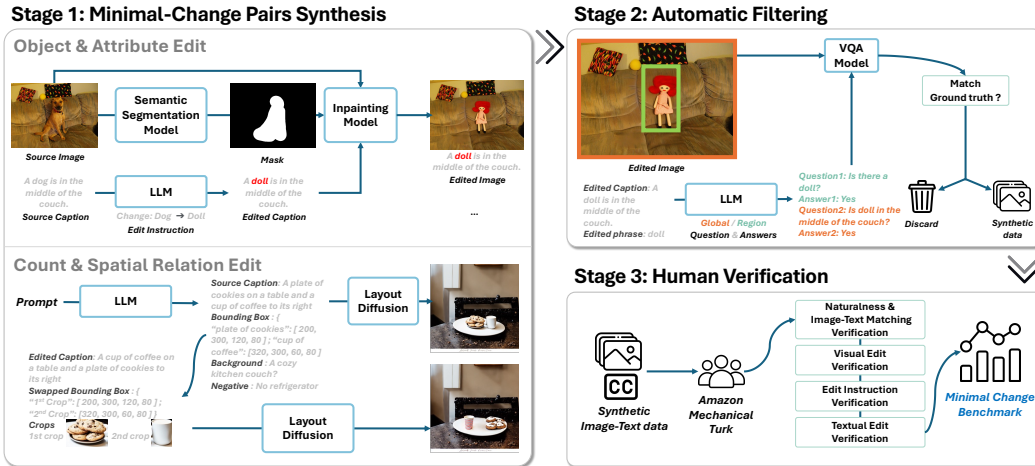
Automatic approach to generate visual hard negatives: Existing approaches to automatically generate visual hard negatives fall into three broad categories: (i) using nearby video frames with semantic changes [14, 38], (ii) using graphic engines [38], (iii) using diffusion models [28, 38, 17]. Our proposed framework falls in the third category. DEMON[17] is the closest to our work, creating training data using diffusion models to improve the learning of a given vision-language model. They use diffusion models to perform local editing on the images given the target object mask. However, this approach requires attention masks from the vision-language model being studied. SPEC [28] proposes a diffusion-based canvas-filling method for generating minimally-different image pairs limited to four types of minimal changes: size, position, count and existence. Compared to these existing methods, our automated pipeline to generate minimal-change data is more involved in order to achieve minimal-changes in complex scenes while maintaining the photo-realism of the scene and controlling changes across diverse categories. Our pipeline also has more a comprehensive automated filtering mechanism compared to previous pipelines that mainly rely on CLIP-based filtering.

Enhancing fine-grained understanding in VLMs with hard negatives: Most methods to enhance fine-grained understanding in VLMs like CLIP focus on fine-tuning with caption-based hard negatives and optimizing loss functions to better use these signals [42, 46, 33]. Common strategies for generating textual hard negatives include heuristic rules [43], language models [46, 5], scene-graph information [7, 33], and LLMs integrated with semantic segmentation [6]. In contrast, fewer works explore visual hard negatives; methods like NegCLIP [42] and General Scene Difference [16] rely on nearest-neighbor images, which often differ too much or too little in context, limiting fine-grained learning. Our approach is closest to SPEC [28] and CounterCurate [45], which also fine-tune VLMs using minimal-change visual hard negatives. Unlike SPEC, we and CounterCurate extend this to multimodal large language models, but our work evaluates performance on 10 out-of-distribution benchmarks (compared to 1 or 2 in SPEC and CounterCurate) and outperforms baseline models in most cases, demonstrating the strength of our approach (see Tables 3 and 4).

3 Minimal-Change Image-Text Dataset Creation

We devised a framework to synthesize large-scale minimal-change data and introduce the VisMin benchmark (see overview fig. 2). The pipeline includes three stages: **Minimal-Change Pairs Synthesis**, where we minimally edit image and text pairs; **Automatic Filtering**, which verifies the faithfulness of the texts and synthesized images; and **Human Verification**, a four-step process to ensure that only data meeting all quality criteria is included. We will discuss each stage in detail.

Figure 2: Our dataset creation pipeline includes three stages: (i) **Minimal-Change Pairs Synthesis**: We develop methods for synthesizing minimal-change image-caption pairs involving Objects & Attributes and Counting & Spatial Relations. (ii) **Automatic Filtering**: An LLM generates questions and answers based on captions, and a VQA model predicts answers from images. Synthetically generated minimal-change data are excluded if answers don't match. (iii) **Human Verification**: Synthetically generated minimal-change data undergoes a rigorous 4-steps human verification, and only examples passing all stages are included in the benchmark.



Stage 1: Minimal-Change Pairs Synthesis In the first stage of our pipeline, we focus on synthesizing minimal-change image-text pairs across four strategic categories: objects, attributes, counting, and spatial relations. These categories are specifically chosen to test various levels of visual-linguistic comprehension. We generate minimal-change text pairs using LLM and then generate minimal-change image pairs using diffusion models. Our synthesis process distinctly tailors the creation of image-text pairs to the specific needs of each category, depicted in Stage 1 in Figure 2 (Object & Attribute Edit and Count & Spatial Relation Edit blocks).

LLM-guided Edit Instructions Generation To generate minimal-change text pairs, we start with source captions and then prompt an LLM (Mistral 47B [10]) to generate both the edit instructions specific to each edit category and the corresponding edited caption (see Appendix A.2.1 for the prompt used). For **Object and Attribute** edits, we use human-written captions from COCO [22] and VSR [23] datasets as our source captions. The LLM processes these captions to suggest edits targeting specific objects or attributes. For example, given the *source caption* “A dog in the middle of the couch”, the LLM generates the *edit instruction* “change dog to doll” which contains both the *source phrase* (“dog”) and the *edited phrase* (“doll”). The LLM also generates the *edited caption* “A doll in the middle of the couch”. We generate five plausible (based on the criteria outlined for LLM prompting) edit instructions and edited captions per source caption. To ensure the edited captions are minimally changed w.r.t the source caption and contain visually plausible changes, we prompt the LLM again for filtering, removing 40% of the total LLM outputs that do not meet those criteria (see appendix A.2.2 for details on the criteria). For **Counting and Spatial Relation** edits, we generate the source captions synthetically due to the absence of a suitable human-written captions dataset containing descriptions of counts and spatial relations of objects. We prompt the LLM to create captions and outline object layouts and bounding boxes. For instance, the LLM might generate a *caption* like “A plate of cookies on a table and a cup of coffee to its right,” with the corresponding *bounding boxes*: {“plate of cookies”: [200, 300, 120, 80]; “cup of coffee”: [320, 300, 60, 80]}. The LLM generates a large pool of such synthetic captions. The edit instructions and the corresponding edited captions are generated using a rule-based method aimed at swapping the object positions for

spatial relation edits (e.g. *edited caption*: “A cup of coffee on a table and a plate of cookies to its right”, *swapped bounding boxes*: {“1st crop”: [200, 300, 120, 80]; “2nd crop”: [320, 300, 60, 80]}) or adjusting the object counts for counting edits (e.g. *edited caption*: “A cup of coffee on a table”), *removed bounding boxes*: {[200, 300, 120, 80]}), in this example removing the plate of cookies from the image.

Diffusion-guided Image Synthesis We modify images according to the edit instructions generated by the LLM in the previous step. For **Object and Attribute** edits, we first mask the object to be edited in the source image using the Grounding-DINO model [25]. We obtain the source images from the COCO dataset. The object to be edited is specified in the source phrase of the edit instruction (e.g., “a dog” in the edit instruction “change dog to doll”). We then apply the SDXL inpainting model [30], using the *input image, masked region, and edited phrase* (obtained from the edit instruction, e.g., “a doll” in the edit instruction “change dog to doll”) to alter the masked image region to match the desired outcome, e.g., changing “a dog” to “a doll.” For **Counting and Spatial Relation** edits, we create a synthetically generated source image dataset based on LLM-suggested layouts from the previous step, using the LLM-grounded Diffusion (LMD) model [21] for image synthesis. To create an edited image, for the spatial relation edits, we first reposition the source image’s bounding boxes using a rule-based method. We then obtain image crops from the source image corresponding to the objects which we need to reposition w.r.t each other. Lastly, we use the GLIGEN layout-diffusion model [20] to smoothly insert the obtained crops into the source image at the repositioned bounding box locations. For counting edits, we obtain the edited image by always removing one or multiple objects from the source image. The object to be removed is specified by masking and we use the Lama model [34] to carry out the object removal. We employ layout-based diffusion models [21, 20] instead of using end-to-end diffusion models like Stable diffusion [30] as the layout-based model facilitates precise control over the object positions and counts and thus ensures the changes are faithful to the edit instruction as well as minimal. Unfortunately, end-to-end models such as Stable Diffusion are not good at precisely editing object positions and counts.

Stage 2: Automatic Filtering To ensure consistency of synthesized hard-negative images, we use a VQA-based filtering system, which is more effective than object detection (see Stage 2 in Figure 2). Questions generated by an LLM [10] based on the edit instruction and caption (following TIFA [9]) verify that edits align with the caption and that the positive caption no longer applies to the negative image. We use LLaVa 7B [24] to answer these questions, with region-specific questions for object/attribute edits and global questions for background consistency. This process removes 75% of synthesized images, ensuring dataset quality.

Stage 3: Human Verification To ensure high-quality of the benchmark, on top of automated filtering, we conduct human verification using the Amazon Mechanical Turk platform. The images and captions undergo four steps of verification, requiring agreement from at least four out of five annotators at each step to pass the human verification. The steps are: **1) Naturalness and Image-Text Matching Verification:** Annotators assess if (a) the image looks natural, (b) the caption is sensible, and (c) the image matches the caption. Only 26% of synthetic images pass this step, mainly due to the criterion (a), where counting and spatial relation images often look unnatural. See Appendix Table 7 for detailed acceptance rates. **2) Visual Edit Verification:** Annotators assess if the images faithfully reflect the specified minimal edits without additional changes, with an acceptance rate of 80%. **3) Edit Instruction Verification:** Annotators assess if the LLM-generated edit instructions are minimal, i.e., the suggested edit modifies only one aspect of the sentence (out of object, attribute, counting, or spatial relation), with an 84% acceptance rate. **4) Textual Edit Verification:** Annotators assess if the edited sentence faithfully reflects the specified minimal edits without additional changes, with a 95% acceptance rate. Annotators also verify the LLM’s categorization of the types of edits (object, attribute, counting, or spatial relation). These steps ensure precise, minimal changes in images and captions, delivering a high-quality benchmark for fine-grained visual understanding. See Appendix A.4.1 for annotator instructions.

4 Training and Benchmark sets

In our study, we create training and benchmark sets to improve and assess fine-grained understanding in VLMs. The training data is generated through a scalable pipeline with automatic filtering, while the benchmark data undergoes additional rigorous human verification to ensure high quality (as explained above). For object and attribute edit types, that make use of natural images, the training data is sourced from VSR (images sourced from COCO) and the COCO 2017 training split

(118K images), while the benchmark data is sourced from the COCO 2017 validation split (5K images). This ensures benchmark images are unseen during training, maintaining evaluation reliability by community standards. The Training dataset has 64,392 samples (37,017 objects, 10,352 attributes, 10,050 counting, 6,973 relations), while the VisMin benchmark has 2,084 samples (579 objects, 294 attributes, 589 counting, 622 relations). We aimed for a balanced benchmark across categories. However, the number of attribute samples in VisMin is relatively low because the LLM suggested attribute edits for only 2000 samples in the COCO 5K validation set. Moreover, most of these suggested edits were color edits. So we further downsampled the color edit instances to balance the distribution of the types of attribute edits. Figure 3 shows subcategories of the changes in VisMin. For detailed training set subcategories, see Appendix 15. For qualitative samples, refer to Appendix 13 and 14.

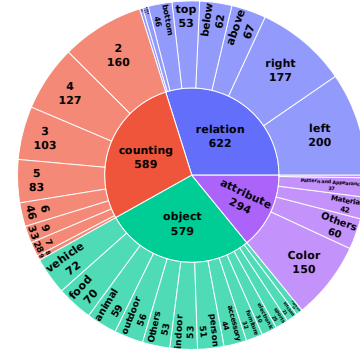


Figure 3: VisMin categories and subcategories.

In table 1, we compare VisMin with related benchmarks. **Visual Minimal HN:** This criterion evaluates if the visual hard negatives contain minimal changes. In Winoground and MMVP, hard negatives differ across multiple aspects (object, attribute, background, etc.). In contrast, VisMin’s hard negatives vary only in one aspect, keeping others unchanged as much as possible. This minimal-change property is also present in What’sUp, EQBEN, SPEC, and in a subset of images from ImageCoDe and CounterCurate. **Visual Complexity:** This criterion assesses the complexity of visual scenes. ImageCoDe and EQBEN mainly feature images from limited video domains and graphic engines, while What’sUp uses simple household and tabletop images. SPEC generates simplistic scenes using diffusion models. In contrast, Winoground uses expert-curated Getty Images, and MMVP uses ImageNet and LAIONAesthetics. VisMin and CounterCurate (concurrent work) stand out by using diverse, complex everyday scenes from COCO [22] and Flickr30K Entities [29], featuring common objects in natural contexts.

Textual Complexity: Benchmarks like ImageCoDe, Winoground, and MMVP use free-form human-written captions. In contrast, What’sUp (focused on spatial changes) and SPEC (focused on controlled changes) use template-based captions, which often lack diversity.

Table 1: Comparison of benchmarks offering visual hard negatives (HN) across: minimal HN, visual complexity, textual complexity, human-approved captions (!) and images (👤), and size. \checkmark : criterion holds for a subset of the benchmark.

| Benchmark | | | | | | |
|---------------------|--------------|--|----------------------------------|--------------|--------------|---------|
| ImageCoDe [14] | \checkmark | Limited video domains, Open Images | Free-form (Human) | \checkmark | \checkmark | 2,306 |
| What’sUp (A/B) [11] | \checkmark | Household, Tabletop | Template | \checkmark | \checkmark | 1,232 |
| Winoground [36] | \times | Expert curated using Getty Images API | Free-form (Human) | \checkmark | \checkmark | 400 |
| EQBEN [38] | \checkmark | Limited video domains, Graphic engine, Synthetic-diffusion | Free-form (Human), Template | \times | \checkmark | 250,612 |
| SPEC [28] | \checkmark | Synthetic-diffusion (limited objects) | Template | \times | \times | 3,000 |
| MMVP [37] | \times | ImageNet, LAIONAesthetics | Free-form (Human) | \checkmark | \checkmark | 150 |
| CounterCurate [45] | \checkmark | Flicker30K Entities, Synthetic-diffusion | Free-form (Human, LLM), Template | \checkmark | \checkmark | 45,400 |
| VisMin (Ours) | \checkmark | COCO, Synthetic-diffusion | Free-form (Human, LLM) | \checkmark | \checkmark | 2,084 |



EQBEN and CounterCurate mix free-form (human or LLM-generated) and template-based captions. VisMin combines human-written and LLM-generated free-form captions, yielding sufficient textual complexity to the benchmark. **Human Verification:** For benchmarks using synthetic images like EQBEN, SPEC, CounterCurate, and VisMin, human evaluation is essential to ensure images look natural. It’s also crucial for benchmarks with automatically generated hard negative captions, as these may be nonsensical unless well-defined templates like What’sUp are used. Nonsensical captions make it easier for VLMs to identify them as incorrect [8]. VisMin is notably the only benchmark with full human verification, ensuring both captions and images are high quality. CounterCurate also conducts human verification but only checks for image-caption consistency (on 300 examples), without verifying image naturalness or caption sensibility. **Size:** This criterion assesses the dataset’s size. VisMin excels by combining **controlled minimal changes with complex, natural scenes and captions**, providing an optimal balance for robust evaluation.

5 Benchmarking VLMs on VisMin Benchmark

Setup We have comprehensively benchmarked existing *state-of-the-art* VLMs on VisMin, encompassing both foundational VLMs—such as CLIP [31], SigLip [44], BLIP [18], and Coca [40] and generative MLLMs including Llava [24], Idefics2 [15] and InternVL1.5 [2]. Additionally, closed-source MLLMs such as GPT4-o [1] and Gemini1.0 Pro [35] are also evaluated.

For foundational models like CLIP, we conducted an image-text matching task using cosine similarity, following [36]. The tasks involved two settings: choosing the correct image from two captions and selecting the correct caption from two images. In VisMin examples (see Fig. 1) with pairs $\{(I_1, C_1), (I_2, C_2)\}$, the **text score** is 1 if $(s(C_0, I_0) > s(C_1, I_0)) \wedge (s(C_1, I_1) > s(C_0, I_1))$, and the **image score** is 1 if $(s(C_0, I_0) > s(C_0, I_1)) \wedge (s(C_1, I_1) > s(C_1, I_0))$; the **group score** is 1 when both scores are 1. For MLLMs, we adapted these tasks to a visual question answering format with binary questions about the matching relationship between images and captions $\{(I_1, C_1), (I_2, C_2)\}$. To calculate the **text score**, we presented the model with one image and two captions, using the prompt “Does this image depict: $\{C_1 \text{ or } C_2\}$?”.³ To calculate the **image score**, we presented the model with two images and one caption, using the prompt “Which image better aligns with the description: ‘ C ’? The first or the second image?”. The score is 1 if the predicted answer matches the ground truth. Once both scores are obtained, the **group score** is 1 if both individual scores are 1.

Table 2: Performance of foundational variants and MLLMs across categories on the VisMin Dataset. Columns ‘I’, ‘T’, and ‘G’ denote Image, Text, and Group scores from Winoground [36]. AVG denotes the average across columns. The best results are highlighted in **bold**.

| | Object | | | Attribute | | | S. Relation | | | Count | | | AVG |
|---|--------|-------|--------------|-----------|-------|--------------|-------------|-------|--------------|-------|-------|--------------|--------------|
| | T | I | G | T | I | G | T | I | G | T | I | G | |
| Random Chance | 25 | 25 | 16.67 | 25 | 25 | 16.67 | 25 | 25 | 16.67 | 25 | 25 | 16.67 | 22.22 |
| MTurk Human | 86.87 | 95.50 | 83.07 | 82.31 | 91.15 | 76.87 | 81.67 | 92.76 | 76.20 | 88.96 | 96.77 | 86.41 | 86.54 |
| CLIP (ViT-B/32) [31] | 79.62 | 77.89 | 67.53 | 72.11 | 65.99 | 55.1 | 8.2 | 4.34 | 0.48 | 31.24 | 20.71 | 10.53 | 41.15 |
| CLIP (ViT-B/16) | 86.53 | 79.1 | 71.68 | 70.75 | 65.31 | 52.38 | 8.84 | 3.22 | 0.8 | 34.3 | 22.58 | 13.58 | 42.42 |
| CLIP (ViT-L/14) | 87.56 | 83.59 | 78.07 | 74.49 | 69.73 | 57.82 | 9.16 | 4.66 | 1.45 | 37.01 | 30.56 | 18.17 | 46.42 |
| NegCLIP [42] | 87.74 | 87.05 | 80.66 | 81.63 | 80.27 | 71.77 | 10.13 | 4.66 | 1.13 | 55.01 | 57.72 | 42.28 | 48.96 |
| SigLip (ViT-B/16)[44] | 90.5 | 88.95 | 83.25 | 86.05 | 79.25 | 73.13 | 11.58 | 6.43 | 1.77 | 60.95 | 47.03 | 38.37 | 55.61 |
| SigLip (ViT-L/16) | 93.44 | 88.43 | 84.46 | 84.35 | 78.23 | 68.37 | 10.29 | 4.82 | 1.29 | 61.8 | 57.05 | 44.14 | 56.39 |
| Flava [32] | 81.69 | 75.12 | 66.66 | 74.49 | 60.54 | 52.04 | 6.75 | 5.79 | 0.96 | 35.99 | 27.5 | 15.45 | 41.91 |
| BLIP [18] | 92.4 | 92.57 | 87.05 | 88.44 | 86.73 | 78.57 | 11.25 | 4.98 | 2.09 | 52.97 | 46.01 | 33.28 | 56.36 |
| Coca [40] | 84.97 | 81.52 | 73.58 | 78.57 | 66.33 | 57.82 | 11.25 | 5.95 | 1.77 | 60.1 | 35.82 | 28.52 | 48.85 |
| LlaVa1.6 (7B) [24] | 93.0 | 32.8 | 32.2 | 92.2 | 34.4 | 33.3 | 91.8 | 7.8 | 7.4 | 73.6 | 25.0 | 20.2 | 38.28 |
| Idefics2 (8B) [15] | 95.4 | 69.4 | 67.6 | 89.1 | 71.4 | 67.0 | 18.6 | 18.8 | 4.8 | 72.2 | 50.6 | 47.0 | 55.99 |
| CogVLM (7B) [39] | 94.64 | 89.63 | 87.56 | 89.11 | 88.43 | 81.63 | 21.70 | 11.09 | 2.25 | 58.40 | 6.45 | 3.56 | 52.87 |
| InternVL1.5 (25B) [2] | 94.65 | 40.24 | 39.72 | 91.16 | 42.86 | 41.16 | 74.28 | 14.79 | 11.74 | 73.51 | 31.58 | 27.5 | 48.60 |
| Gemini1.0 Pro  | 94.99 | 79.97 | 78.76 | 91.84 | 74.83 | 72.45 | 52.57 | 15.43 | 9.81 | 67.74 | 44.14 | 37.52 | 49.63 |
| GPT4-o  | 95.51 | 96.2 | 93.44 | 92.18 | 90.48 | 87.07 | 89.07 | 50.48 | 46.78 | 77.42 | 78.27 | 68.42 | 73.93 |

Results Insights from Table 2 highlight key capabilities and limitations of current models. Text scores generally surpass image scores, especially in MLLMs, where text scores are often two to three times higher. In contrast, foundational VLMs show a modest discrepancy between image and text scores. We hypothesize that for MLLMs, the image score is lower compared to the text score because they lack training with multiple images, and simple vertical concatenation does not provide sufficient visual signals, leading to suboptimal alignment with captions. Notably, Idefics2, which supports multi-image processing, performs similarly on text and image scores, underscoring the importance of multi-image data during pretraining. Foundational VLMs’ higher text scores suggest that distinguishing between captions is easier than between images, highlighting the need for our visual minimal change benchmark. Interestingly, foundational VLMs generally outperform MLLMs due to the latter’s lower image scores.

All models perform well on Object and Attribute splits, indicating that understanding semantic changes correlates strongly with recognition capabilities. Models excelling in image classification tend to perform better, reflecting a foundational understanding that does not require advanced reasoning. For instance, Idefics2, using the SigLip (ViT-L/16) vision encoder, performs worse with strong LLMs compared to its foundational VLM counterpart, likely due to limited multi-image understanding in MLLMs. Notably, CogVLM shows superior performance over other MLLMs in understanding object and attribute changes. This advantage likely stems from its integration of grounding tasks and high-quality human-annotated datasets like Flickr30K Entities [29], RefCOCO [12], and VisualGenome [13] in its pretraining. On the other hand, the Spatial Relation split relies heavily on reasoning capabilities, with MLLMs outperforming foundational models. This suggests that LLMs can parse object relationships through reasoning. However, existing VLMs struggle with spatial relations, often scoring below random chance, indicating potential biases in models and highlighting an area for further research.

³For single-image models, such as Llava, we combine the two images vertically with a clear border.

We document human baseline performance on our benchmark via Amazon Mechanical Turk (see Appendix A.4.2). Humans generally outperform models on image scores, except in the attribute category where GPT4-o excels. Models typically surpass humans in text scores, especially with attributes and objects. However, in spatial relations and counting, humans significantly outperform models in group scores, highlighting areas for model improvement and the robustness of human scene comprehension.

6 Enhancing fine-grained understanding in VLMs

We use a synthetic minimal-change dataset to enhance fine-grained understanding through additional fine-tuning of VLMs. Training with pairs of images and captions with minimal differences provides a richer training signal, improving model performance in fine-grained understanding tasks. We demonstrate improvements on top of both foundational VLMs and MLLMs by conducting extensive evaluations across various benchmarks: (1) **Single image benchmarks** test the alignment between single images and multiple captions: VSR [23], CountBench [26], VALSE [27], SPEC [28], and Sugarcrepe [8]. (2) **Multiple image benchmarks** test the alignment between multiple images and captions: ImageCode [14], MMVP [37], Whatsup [11], Winoground [36], EQBEN [38], and our VisMin benchmark.

6.1 Fine-tuning Foundational VLMs

Enhancement with Minimal-Change Data Our approach uses a synthetic minimal-change dataset to improve visual representation without altering the training methodology. We construct training batches with both source and edited image-text pairs: In the original CLIP training, a mini-batch is $\mathcal{B} = \{(C_1, I_1), (C_2, I_2), \dots, (C_n, I_n)\}$, with pairs randomly sampled from the dataset as random negatives. With minimal-change data, we add edited image-text pairs as hard negatives, resulting in $\mathcal{B} = \{(C_1, I_1), (C'_1, I'_1), (C_2, I_2), (C'_2, I'_2), \dots\}$, where (C'_n, I'_n) is the edited pair of (C_n, I_n) . We use a total batch size of 128 with 4 A100 GPUs and retain other training protocols and hyperparameters as default from OpenCLIP [3], including a learning rate of 1e-05, weight decay of 0.2, Adam β_1 of 0.9, β_2 of 0.98, an eps of 1e-06, and a cosine scheduler. The training runs for 5 epochs, and we select checkpoints based on a separate VisMin validation set.

We fine-tuned pre-trained CLIP on our minimal-change data, calling it VisMin-CLIP. For comparison, we implemented three models using the same pre-trained CLIP: NegCLIP [42], CounterCurate-CLIP [45], and SPEC-CLIP [28]. In NegCLIP, we fine-tuned CLIP with automatically generated hard-negative captions and nearest-neighbor images paired with human-written captions. For CounterCurate-CLIP, we used hard-negative data (attribute, position, counting) but trained one model on all three types, unlike the original approach of training separate models. SPEC-CLIP was fine-tuned on six combined category-specific splits (size, spatial, existence, count). Hard-negatives were included in all batch constructions, with excess negatives rolled into the next batch if needed. All models used ViT-L/14 as the backbone and the original CLIP loss, initialized from OpenAI checkpoints. Best checkpoints were selected from validation sets for NegCLIP and CounterCurate-CLIP, and from average benchmark performance for SPEC-CLIP. This controlled comparison evaluates the impact of different hard-negative data approaches on improving CLIP’s fine-grained understanding.

Table 3: Performance of fine-tuned CLIP and Idefics2 across categories on the VisMin Dataset. The [†] symbol indicates the reproduced model checkpoints based on their respective training data.

| | Object | | | Attribute | | | S. Relation | | | Count | | | AVG |
|---------------------------------|--------|-------|--------------|-----------|-------|--------------|-------------|-------|-------------|-------|-------|--------------|--------------|
| | T | I | G | T | I | G | T | I | G | T | I | G | |
| CLIP(ViT-L/14) | 87.56 | 83.59 | 78.07 | 74.49 | 69.73 | 57.82 | 9.16 | 4.66 | 1.45 | 37.01 | 30.56 | 18.17 | 46.02 |
| NegCLIP [†] | 87.74 | 87.05 | 80.66 | 81.63 | 80.27 | 71.77 | 10.13 | 4.66 | 1.13 | 55.01 | 57.72 | 42.28 | 55.00 |
| CounterCurate-CLIP [†] | 89.81 | 91.02 | 84.46 | 82.99 | 80.27 | 72.79 | 20.1 | 11.41 | 7.4 | 49.24 | 45.16 | 31.92 | 55.54 |
| SPEC-CLIP [†] | 86.53 | 86.01 | 78.58 | 78.57 | 71.77 | 63.95 | 9.16 | 5.31 | 1.13 | 45.5 | 47.71 | 32.43 | 50.55 |
| VisMin-CLIP | 91.54 | 91.19 | 86.36 | 85.03 | 83.67 | 75.85 | 11.9 | 3.38 | 1.29 | 82.34 | 79.97 | 72.33 | 63.74 |
| Idefics2 | 95.4 | 69.4 | 67.6 | 89.1 | 71.4 | 67.0 | 18.6 | 18.8 | 4.8 | 72.2 | 50.6 | 47.0 | 55.99 |
| VisMin-Idefics2 | 96.5 | 95.7 | 93.3 | 91.2 | 91.8 | 86.7 | 83.0 | 76.0 | 69.3 | 85.4 | 87.8 | 80.5 | 86.43 |

Results We evaluated these models on the VisMin benchmark (results in Table 3). Fine-tuning with minimal-change data significantly improves CLIP’s performance on the Object, Attribute, and Count categories, demonstrating the usefulness of our minimal-change data in enhancing the fine-grained

Table 4: Evaluation on other single and multi-image **visual** fine-grained understanding benchmarks. All models adopt **ViT-L-14** as the vision encoder. **CB** refers CountBench, **SG** refers to Sugarcrepe, **IC** refers to Imagecode. I2T and T2I indicate standard standard image-to-text and text-to-image retrieval metrics. Best-performing models in the CLIP-family are highlighted in blue, and best-performing MLLM models are highlighted in green.

| | #Samples | SINGLE-IMAGE | | | | MULTI-IMAGE | | | | | | | | | | | | | |
|---------------------------------|----------|--------------|-------|-------|-------|-------------|-------|-------|-------|-------|------------|-------|------|-------|-------|-------|--------|-------|-------|
| | | VSR | CB | VALSE | SG | Whatsup | SPEC | | IC | MMVP | Winoground | | | EQBEN | | | VisMin | | |
| | | | | | | | I2T | T2I | | | T | I | G | T | I | G | T | I | G |
| CLIP (ViT-L/14) | - | 58.33 | 33.65 | 69.1 | 73.0 | 37.7 | 32.85 | 30.86 | 61.47 | 19.26 | 27.5 | 11.0 | 8.5 | 35.71 | 33.57 | 21.43 | 52.05 | 47.13 | 38.88 |
| NegCLIP [†] | 118K | 56.56 | 40.0 | 75.41 | 85.73 | 41.2 | 37.73 | 35.45 | 67.33 | 29.63 | 25.25 | 12.0 | 7.0 | 42.86 | 40.0 | 30.0 | 58.63 | 57.42 | 48.96 |
| CounterCurate-CLIP [†] | 241k | 56.74 | 30.79 | 68.47 | 83.66 | 44.29 | 37.99 | 35.24 | 65.81 | 25.19 | 28.0 | 13.25 | 9.0 | 45.0 | 33.57 | 28.57 | 60.88 | 56.68 | 49.1 |
| SPEC-CLIP [†] | 637k | 64.54 | 32.06 | 68.75 | 79.34 | 43.35 | 87.04 | 88.08 | 66.25 | 30.37 | 22.5 | 7.75 | 4.75 | 41.43 | 40.0 | 30.71 | 54.94 | 52.7 | 44.02 |
| VisMin-CLIP | 65k | 58.69 | 49.84 | 72.24 | 81.44 | 43.99 | 44.28 | 39.98 | 66.81 | 32.59 | 32.75 | 14.75 | 9.75 | 54.29 | 40.71 | 33.57 | 67.7 | 64.55 | 58.96 |
| Idefics2 | - | 77.3 | 91.11 | 88.91 | 90.45 | 68.04 | 74.37 | 60.5 | 64.4 | 48.15 | 47.25 | 33.75 | 22.5 | 62.88 | 33.33 | 25.76 | 68.83 | 52.56 | 46.6 |
| Vismin-Idefics2 | - | 80.32 | 93.97 | 86.08 | 91.14 | 74.42 | 76.2 | 76.58 | 70.7 | 48.89 | 47.0 | 35.75 | 22.5 | 64.39 | 54.55 | 49.24 | 89.01 | 87.83 | 82.44 |

understanding of foundational VLMs such as CLIP. VisMin-CLIP consistently outperforms NegCLIP, CounterCurate-CLIP and SPEC-CLIP across all categories except spatial relations. This suggests that the visual minimal-change data is more helpful in improving the fine-grained understanding capabilities of the CLIP model compared to the nearest neighbor images in NegCLIP and not fully minimally-changed CounterCurate and SPEC data.

We further conduct a zero-shot evaluation of the fine-tuned CLIP models on other fine-grained understanding benchmarks (beyond our VisMin benchmark) to test their generalization capabilities (see Table 4). VisMin-CLIP performs best in 11 out of 18 tasks, while NegCLIP, CounterCurate-CLIP, and SPEC-CLIP lead in 3, 1, and 3 tasks, respectively. All models outperform the pre-trained CLIP model across benchmarks. For counting and spatial reasoning tasks, VisMin training data shows significant improvements. On CountBench, we observed 9%, 19%, and 17% gains over NegCLIP, CounterCurate-CLIP, and SPEC-CLIP, respectively. Similarly, on spatial reasoning benchmarks (SPEC, Whatsup, VSR), VisMin-CLIP showed an average 7.79% improvement over NegCLIP and 5.21% over CounterCurate-CLIP. While SPEC-CLIP performed best on the in-distribution SPEC benchmark, VisMin-CLIP still outperformed other models, indicating its effectiveness in improving fine-grained understanding. For VALSE and SugarCrepe, NegCLIP performed best, likely due to similarities in the textual hard-negative generation process used in those benchmarks and NegCLIP’s fine-tuning data.

Furthermore, our minimal-change data significantly outperforms others in multi-image understanding benchmarks. Fine-tuning on this data enhances the model’s ability to distinguish between similar images, showing improvements on challenging benchmarks like Winoground, MMVP, and EQBEN, which test compositional reasoning and fine-grained understanding. VisMin-CLIP improved Text scores by 6% on Winoground and 18% on EQBEN over baseline CLIP, demonstrating effective alignment of visual and textual feature spaces. VisMin-CLIP also outperforms other models on most tasks and achieves comparable results on remaining benchmarks (except SPEC), despite using fewer samples (e.g., 65K in VisMin vs. 637K in SPEC).

Figure 4: VisMin fine-tuned models show greater improvements with larger models. The circle radius reflects the number of model parameters.

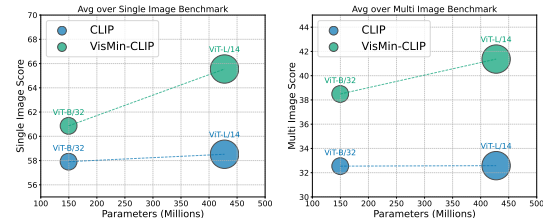
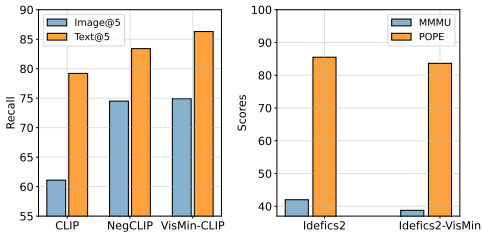


Figure 5: (left) Recall results with ViT-L/14 on COCO benchmark (right) standard benchmark results on Idefics2.



Additional Findings Further experiments reveal several key findings: (1) **Scalability**: As illustrated in fig. 4, we evaluated varying sizes of OpenAI’s CLIP models– B/32 and L/16. Larger models demonstrated improved performance across both single-image and multi-image benchmarks after training on our synthetic data. This improvement is likely because understanding minimal changes is a complex task, demanding robust model capabilities. For instance, the smallest tested model,

ViT-B/32 (149.62M parameters), exhibited improvements of 2.37 and 3.24 in single and multiple image benchmarks, respectively, when comparing VisMin-CLIP against the baseline CLIP. When the model’s capacity was expanded to ViT-L/14 (427.62M parameters), the improvements increased to 6.88 and 9.21, respectively. These results highlight the scalability and efficacy of our data in enhancing model performance. (2) **Enhanced Original Capabilities:** In addition to improvements in fine-grained understanding tasks, training on our data also enhances performance in standard retrieval tasks, as shown in fig. 5. This suggests that models achieve better alignment from training on minimal change tasks, indicating that our data is generally applicable across various cross-modal tasks.

6.2 Fine-tuning Multimodal Large Language Models (MLLMs)

We utilized Idefics2 [15] to improve fine-grained understanding, employing our instruction-formatted dataset. Given its proficiency in multimodal interactions and advanced multi-image processing, Idefics2 was chosen for its open-source accessibility, model size and leading zero-shot performance.

Dataset and QLoRa Fine-tuning Our dataset, *VisMin Instruct-IT*, includes image-text pairs created using a rule-based approach (see A.2.3 for details). We reformulated these pairs for MLLMs, where the task is to select the correct image from two options based on a given caption or choose the appropriate caption for an image from two possibilities. While the base Idefics2 model was trained with a variable number of images in a sequence, we limited it to two images to include one positive and one hard negative example from VisMin. We fine-tuned the Idefics2-8B model using the QLoRa technique [4], updating adapters in the language model and modality connector including perceiver resampler with 1 A100 80GB GPU. We used 4-bit quantization, with $r = 64$ and $\alpha = 16$ for LoRa, and a learning rate of $1e - 5$. The model was fine-tuned for one epoch with an accumulated batch size of 64.

Results The fine-tuned Idefics2 model shows significant improvement on VisMin (see Table 3) across all categories, comparable to GPT4-o (see Table 2), demonstrating the effectiveness of our minimal-change data in enhancing MLLM fine-grained understanding. Notably, in the Spatial Relation category, gains of 64.4%, 57.2%, and 64.5% were observed for Text, Image, and Group, respectively, unlike CLIP, where fine-tuning did not improve spatial understanding. Fine-tuning Idefics2 also transfers well to other fine-grained benchmarks, achieving over 5% overall improvement (see Table 4). To assess generalization, we evaluated its zero-shot performance on non-fine-grained tasks (MMM U [41] and POPE [19]; results in Fig. 5 (right)). The model maintained comparable performance on POPE but dropped on MMM U, likely due to the binary-choice format in our fine-tuning data. This suggests further gains could be made by combining our data with instruction-tuning datasets, though this wasn’t pursued due to the GPU demands of fine-tuning an 8B model with additional data.

In addition to our main results, we provide further analysis in the Appendix A.3, including detailed evaluations on additional multimodal benchmarks, zero-shot image classification task, and video understanding tasks.

7 Conclusion and Limitations

We present VisMin , a benchmark for evaluating fine-grained visual understanding in VLMs such as CLIP, SigLIP, LLaVA, and Idefics2. While these models perform well in object and attribute recognition, they struggle with counting and spatial relationships. To address this, we fine-tuned CLIP and Idefics2 on our minimal-change dataset, yielding significant improvements in objects, attributes, and counting. For spatial relations, CLIP showed limited gains, while Idefics2 made notable progress. Fine-tuning also enhanced CLIP’s image-text alignment, demonstrated in COCO retrieval tasks, underscoring our dataset’s value as a training resource for VLMs. **Limitations:** Despite automatic filtering, the dataset contains noise, including image deformations and text-image mismatches due to diffusion model limitations. Future diffusion model advancements should improve minimal-change editing. Additionally, our model may inherit social biases from the base models (CLIP/Idefics2), as no specific mitigation measures were applied during fine-tuning. Our use of uniform prompts for evaluation may have influenced performance variably.

Acknowledgments

We express our gratitude to Mojtaba Faramarzi and Yash Goyal for their constructive feedback throughout the different stages of the project. We also acknowledge the valuable feedback provided by Qian Yang, Kanish Jain, and Oscar Manas on the early draft. The technical support extended by the Mila IDT team in managing the computational infrastructure is greatly appreciated. Additionally, Aishwarya Agrawal received support from the Canada CIFAR AI Chair award throughout this project.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- [3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.
- [4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.
- [5] Sivan Dohav, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023.
- [6] Sivan Dohav, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [7] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs, 2023.
- [8] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*, 2023.
- [9] 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.
- [10] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [11] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568. URL <https://aclanthology.org/2023.emnlp-main.568>.
- [12] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

- [13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [14] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.241. URL <https://aclanthology.org/2022.acl-long.241>.
- [15] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.
- [16] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.
- [17] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat seng Chua, Siliang Tang, and Yueting Zhuang. Fine-tuning multimodal LLMs to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=BXY6fe7q31>.
- [18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [20] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *ArXiv*, abs/2301.07093, 2023. URL <https://api.semanticscholar.org/CorpusID:255942528>.
- [21] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.
- [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [23] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [25] 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.
- [26] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten, 2023.
- [27] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.567. URL <https://aclanthology.org/2022.acl-long.567>.
- [28] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding, 2023.
- [29] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017.
- [30] 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. *arXiv preprint arXiv:2307.01952*, 2023.
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [32] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.
- [33] Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality, 2023.
- [34] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.
- [35] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.
- [36] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, June 2022.
- [37] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024.
- [38] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11998–12008, October 2023.
- [39] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.
- [40] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [41] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.
- [42] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

- [43] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=KRLUvxh8uaX>.
- [44] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.
- [45] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. *arXiv preprint arXiv:2402.13254*, 2024.
- [46] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic fine-grained understanding. *arXiv preprint arXiv:2306.08832*, 2023.
- [47] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-demos.4>.

A Appendix

A.1 Societal Impact

Our work shares the broader positive societal implications of vision-language research, particularly in improving the accuracy of vision-language models (VLMs). These advancements could benefit various applications, such as helping visually impaired individuals better interpret their surroundings, creating interactive learning tools for children, and enabling more effective interaction with in-home robots. However, like most technologies, accurate vision-language systems also present potential risks. For instance, they could be misused to extract sensitive information from public CCTV footage or inadvertently leak personal data, such as credit card information, when assisting visually impaired users. To address these concerns, robust privacy safeguards are essential to mitigate these risks.

A.2 Prompt Templates

A.2.1 Prompts Used for Edit Instruction Generation

We utilize the Mixtral 47B LLM to generate edit suggestions across selected categories: objects, attributes, counting, and spatial relations. The specific prompts used for each category are detailed in the Boxes below. These prompts guide the LLM in creating minimal, precise modifications to captions, ensuring the generation of high-quality hard-negative instances.

In-context demonstrations for objects edit suggestions.

As an AI language model, your task involves making minimal, targeted edits to a caption that describes an image, guiding corresponding visual changes to be made in the edited version of the image.

Follow these structured steps:

- Suggest Edit: Identify and suggest a specific edit from the given caption. Your edit should only change an object in the scene.
*The edit must adhere to the following criteria:
Object changes should be visually distinct and mutually exclusive. Do not replace an object with its synonyms or closely related sub-categories. The replacement object must fit within the original object's region and be visually plausible within the scene. Maintain a one-to-one replacement ratio; do not introduce additional objects.
- General Criteria for Edits: Avoid any changes related to action, spatial relations, counting, or the size of objects. Edits must be visually and contextually plausible, ensuring the scene remains coherent and the edit does not introduce inconsistencies in other parts of the image. Create an Edited Caption: Draft a new caption reflecting the image post-edit.
- Specify Edit Category: Clearly state that your suggested edit falls under 'Object'.

In-context demonstrations for attributes edit suggestions.

As an AI language model, your task involves making minimal, targeted edits to a caption that describes an image, guiding corresponding visual changes in the edited version of the image.

Follow these structured steps:

- Suggest Edit: Identify and suggest a specific edit from the given caption if it is editable. Your edit should modify an attribute of existing objects in the caption.
*The edit must adhere to the following criteria: Change only the attributes of an object from one of the allowed sub-category: color, pattern, shape, appearance, material, state, condition. Replacing an object itself or introducing new attributes is not allowed. Attribute

changes must be distinct, mutually exclusive, and not synonymous. The edited attribute should make visual and contextual sense within the image without altering its overall composition.

- General Criteria for Edits: Do not suggest attribute changes related to spatial relations, counting, or the size of objects. Edits must be visually and contextually plausible, ensuring the scene remains coherent and the edit does not introduce inconsistencies. Create an Edited Caption: Draft a new caption reflecting the image post-edit.
- Specify Edit Category: Clearly state that your suggested edit falls under 'Attribute'. If you cannot change an attribute due to a missing attribute in the given caption, please output 'EMPTY'.

In-context demonstrations for counting image description generation.

Task Instruction Counting

Your goal is to suggest a textual description for creating a visual scene within a 512x512 pixel canvas, focusing specifically on a counting task involving objects numbered between two and nine. Follow these steps to ensure comprehensive output:

1. ****Input****: Propose a scene with groups of distinct objects, ensuring each group contains a different number of items ranging from two to nine. These objects should be easily countable and distinguishable. The objects should be spread out across the canvas; cluttering or overlapping is prohibited. Detail the arrangement of these objects within the scene, considering their visibility for accurate counting. Do not mention any activity in the scene.
2. ****Bounding Boxes and Object Counts****: For each group of objects, provide bounding box data in the format 'object name', [x, y, width, height]. This data should reflect the object's position and size within the 512x512 image, aiding in identifying the exact number of objects present.
3. ****Background Prompt****: Suggest a fitting and simple background that does not detract from the main task of counting the objects.
4. ****Negative Prompt****: List elements that should be avoided in the image to ensure clarity and focus on the counting task.
5. ****Category (sub-category)****: Identify the most relevant category and sub-category for the scene, based on the objects and their arrangement. Choose categories that enhance the clarity and focus of the counting task.

In-context demonstrations for relation image description generation.

Task Instruction Relation

Your goal is to suggest a textual description and structured data for creating a visual scene within a 512x512 pixel canvas. Follow these steps to ensure a comprehensive output:

1. **Input**: Suggest a description with pairs of distinct objects that commonly occur together in a scene. These objects could be swapped in their positions if needed later on. Avoid suggesting common background objects in this step. You may occasionally mention the relative size, distance and orientation between objects if applicable.
2. **Bounding Boxes and Object Pairs**: Detail each identified object with its bounding box in the format 'object name', [x, y, width, height], indicating the object's position and size within the 512x512 image. Be specific about the spatial relationship of objects, such as whether they are on the 'top', 'down', 'above', 'below', 'left', or 'right' of each

other.

3. Background Prompt: Based on the scene described in the input, suggest a fitting and straightforward background.
4. Negative Prompt: List elements to be excluded from the image.
5. Category (sub-category): Identify the most relevant category and sub-category for the scene based on the objects and their arrangement. Limit the options for sub-categories based on the provided reference examples.

A.2.2 Prompts Used for Edit Instruction Verification

We employ the Mixtral47B LLM to verify edit suggestions in the object and attribute categories. The specific prompts used for this verification process are illustrated in the Box A.2.2. Verification is crucial because initial suggestions by the LLM can occasionally be implausible or violate the intended edit category. By thoroughly verifying the edits, we ensure the generation of accurate and high-quality hard-negative instances, which are essential for effectively testing and improving VLMs.

In-context demonstrations for LLM suggested edit verification for object and attribute category.

****Task Instruction Object****

Determine the acceptability of edits to an image caption based on following criteria:

- Edits must only involve object changes within a specified scene region.
- No changes related to attributes, spatial relations, counting, or any unrelated alterations.
- The new object must be visually distinct and not closely related to the original. Edits should be confined to one region without needing adjustments elsewhere for scene consistency.
- Adhere to a one-to-one replacement rule, introducing no additional objects. Reject edits that add unnecessary specificity or make assumptions not clear from the original caption.
- Reject if information is insufficient or uncertain. Strict filtering: When in doubt, reject the edit.

****Task Instruction Attribute****

Evaluate edits to an image caption that change an object's attributes based on following criteria:

- Reject changes involving the objects themselves, their spatial relations, counting, or size.
- Acceptable modifications are distinct changes in color, pattern, shape, texture, material, state, or condition. Attribute changes must be distinct and not synonymous.
- Reject edits suggesting synonym replacements or adding new attributes.
- Reject if information is insufficient or uncertain. Strict filtering: When in doubt, reject the edit.

A.2.3 VisMin Instruct-It Dataset

In the VisMin Instruction dataset, a sample consists of a source image-caption pair (I_0, C_0) and its minimally edited version (I_1, C_1) . To construct the VisMin Instruct-IT dataset, we create four instruction samples from each VisMin sample:

- Two samples for the Image-Task (selecting the best matching image given a caption):
 1. Given C_0 , select between I_0 and I_1
 2. Given C_1 , select between I_0 and I_1
- Two samples for the Text-Task (selecting the best matching caption given an image):

1. Given I_0 , select between C_0 and C_1
2. Given I_1 , select between C_0 and C_1

Please refer to Table 5 in the rebuttal PDF for the exact set of instruction templates used. We randomly sample one template for each task type when creating the instruction samples.

In total, we create 65k instruction samples for training. Additionally, we include 16k randomly selected image-caption pairs (either (I_0, C_0) or (I_1, C_1)) to retain the model’s general image captioning ability.

Table 5: VisMin Instruct-It dataset creation templates

| Template Type | Template Examples |
|---------------|--|
| Image-Task | “You are given two images. Which one better aligns with the description: {caption}? The first or the second image?” |
| | “Question: You are given two images. Which one better aligns with the description? {caption} Choices: First, Second.” |
| | “Based on the description: {caption}, please select one of the given options that best matches the image. Choices: First, Second.” |
| Text-Task | “Question: Does this image depict: (A) {candidate text A}, or (B) {candidate text B}? Choices: A, B.” |
| | “Does this image best represent: (A) {candidate text A}, or (B) {candidate text B}?” |
| | “Which of the following best matches the image: (A) {candidate text A}, or (B) {candidate text B}?” |

A.3 Additional Results

We report additional results to demonstrate the impact of fine-tuning on our dataset across different aspects of scene understanding. We evaluated our fine-tuned models, VisMin-CLIP and VisMin-Idetics2, on various tasks designed to assess distinct elements of multimodal comprehension.

Table 6: Performance breakdown of EQBEN video splits for both CLIP fine-tuned and Idetics2 models. The table also includes results on additional multimodal benchmarks for Idetics2 family of models and ImageNet results for various CLIP models. We leave the CLIP results on the additional multimodal benchmarks and the Idetics2 results on ImageNet blank due to the task-format not being suitable for evaluation.

| | EQ-YouCook2 | | | EQ-GEBC | | | EQ-AG | | | Additional Multimodal Benchmarks | | | | ImageNet | |
|-----------------|-------------|-------|-------|---------|-------|-------|-------|-------|-------|----------------------------------|---------|---------|-------|----------|-------|
| | T | I | G | T | I | G | T | I | G | MathVista | MMBench | TextVQA | VQAv2 | top1 | top5 |
| CLIP(ViT-L/14) | 50.00 | 65.00 | 40.00 | 15.00 | 10.00 | 0.00 | 15.00 | 25.00 | 10.00 | - | - | - | - | 75.53 | 94.59 |
| NegCLIP | 65.00 | 60.00 | 50.00 | 20.00 | 20.00 | 0.00 | 15.00 | 35.00 | 10.00 | - | - | - | - | 65.50 | 90.00 |
| VisMin-CLIP | 80.00 | 70.00 | 65.00 | 30.00 | 10.00 | 5.00 | 30.00 | 25.00 | 10.00 | - | - | - | - | 68.20 | 91.81 |
| Idetics2 | 84.62 | 53.85 | 53.85 | 25.00 | 40.00 | 15.00 | 52.63 | 47.37 | 31.58 | 53.00 | 76.18 | 73.40 | 78.82 | - | - |
| VisMin-Idetics2 | 69.23 | 84.62 | 61.54 | 35.00 | 20.00 | 15.00 | 57.89 | 36.84 | 31.58 | 47.80 | 72.74 | 71.89 | 79.75 | - | - |

Multimodal Benchmarks: We assessed performance of multimodal LLMs (Idetics2 and Vismin-Idetics2) on benchmarks like MathVista, MMBench, TextVQA, and VQAv2 to understand how fine-tuning on our data impacts tasks requiring both general scene understanding and specialized skills (e.g., OCR and math) (see Table 6, middle block).⁴ Results show slight declines on specialized tasks (MathVista, MMBench, TextVQA) but notable improvements on VQAv2 that tests for general scene understanding, supporting our hypothesis that our dataset strengthens general scene comprehension while not losing performance much on more specialized tasks.

Zero-Shot Image Classification: To evaluate the robustness of learned representations of the contrastive models, we included evaluation on zero-shot image classification on ImageNet (see Table 6, last block). Both NegCLIP and VisMin-CLIP underperform compared to the base CLIP model. However, VisMin-CLIP shows a smaller performance drop compared to NegCLIP, indicating that it retains a stronger generalization capability even after fine-tuning.

Performance on Video Understanding Tasks: Finally, we tested if fine-tuning the models on our minimal-change dataset also improves their performance on video understanding tasks since video understanding tasks require identifying subtle frame-to-frame variations (see Fig. 6 for some examples). To test this, we evaluated the models on the EQBEN [38] benchmark, particularly the following splits: EQ-YouCook2, which focuses on cooking procedures from instructional videos; EQ-GEBC, which identifies event boundaries in kinetic videos; and EQ-AG, which captures changes between objects and their pairwise relationships during actions.

⁴Contrastive models with two-tower encodings are not suitable for these benchmarks, as these benchmarks require generating text outputs rather than just measuring similarity between encoded images and texts.

We would like to note that our fine-tuned models excel at discerning subtle *semantic* differences, but neighboring video frames often exhibit *low-level* rather than semantic changes. Abrupt changes between frames, like a bus turning into a car, a shirt changing color, or objects swapping positions, are rare. Additionally, our dataset does not cover action changes, which are typical in videos. Therefore, we expect finetuning on our dataset to result in limited improvements for nearby video frames.

This was confirmed in our results (see Table 6, first block), where fine-tuning CLIP and Idefics2 showed significant improvements on EQ-YouCook2 (in group scores) due to its higher presence of object-based changes, while performance remained similar on EQ-GBEC and EQ-AG, which mostly feature action shifts. Thus, our dataset is most effective for enhancing video understanding in scenarios involving subtle semantic adjustments rather than dynamic actions.

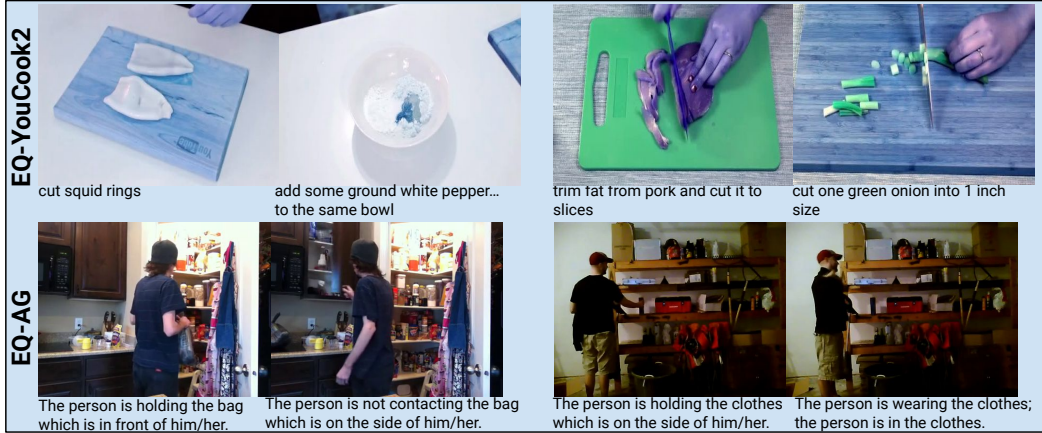


Figure 6: Examples from different video splits from the EQBEN benchmark.

A.4 Annotation Platform and Benchmark Quality

A.4.1 Human Verification of Benchmark

The human verification process involves four main steps. **1) Naturalness and Image-Text Matching Verification:** in this step, humans evaluate the data samples using the following three criteria: a) whether the image looks natural or not, b) whether the caption sounds sensible or not, and c) whether the image matches the caption or not, addressing limitations of automatic filtering and maintaining robustness against manipulation [8] (see Figure 7). Only 26% of synthetic images pass this step, highlighting the need for human verification. The low acceptance rate is mainly associated with the high rejection rate of criterion a, where most counting and spatial relation synthetic images do not appear natural. Please refer to Appendix Table 7 for detailed acceptance rates for each criterion. Please note that we aimed for a balanced benchmark. To address the higher rejection rates for the categories of counting and spatial relation, we began with a larger number of samples in these categories. **2) Visual Edit Verification** confirms that the edited images accurately reflect the specified edits without additional changes, ensuring visual minimal change, with an acceptance rate of 80% (see Figure 8). **3) Edit Instruction Verification** checks that LLM-generated edit instructions are minimal and targeted, altering only one aspect of the object, attribute, counting, or spatial relation in captions and images, with an acceptance rate of 84%. (see Figure 9). **4) Textual Edit Verification** ensures the edited sentence accurately reflects the specified edit instruction without additional changes and classifies the change type, with an acceptance rate of 95% (see Figure 10).

A.4.2 Human baseline

We gather human annotations using Amazon Mechanical Turk (AMT) to determine human performance on our benchmark, replicating the exact tasks given to models as described in Section 5. Additionally, annotators can choose “none” or “both” options (please see Figure 11 and Figure 12); we deliberately included these options to accurately estimate the best model performance when tasks are difficult for humans. We collect five annotations per sample for each of the four scenarios in the image-text matching task. We compute image score, text score, and group score as described in

Table 7: Acceptance rates for all criteria across different categories.

| Criterion | Object | Attribute | Counting | S. Relation | Overall |
|---|---------------|------------------|-----------------|--------------------|----------------|
| Step 1-a: Naturalness of Image | 64% | 80% | 41% | 22% | 37% |
| Step 1-b: Sensicality of Caption | 84% | 79% | 75% | 70% | 74% |
| Step 1-c: Image-Text Matching | 83% | 89% | 52% | 65% | 65% |
| Step 2: Visual Edit Verification | 81% | 79% | 81% | 78% | 80% |
| Step 3: Edit Instruction Verification | 81% | 86% | 85% | 84% | 84% |
| Step 4: Textual Edit Verification | 95% | 94% | 94% | 95% | 95% |
| Step 4: Automatic Categorization Verification | 95% | 91% | 99% | 100% | 97% |

Section 5. If the answer “both” or “none” is chosen by three or more annotators, or if one annotator chooses “none” or “both” while the other options each receive exactly two votes, we randomly assign a 50% match to maintain consistency with the model evaluation setup.

Instructions

In this HIT, you will be shown an image and a description associated with the image. Your task is to answer following questions:









1. Does the description sound sensical, i.e. is it fluent and plausible? Please see examples below.

| Sounds sensical ✓ | Does not sound sensical ✗ |
|---|---|
| ✓ A scene with four vehicles. | ✗ Non-fluent: A group of cat. |
| ✓ A covered box below a motorbike. | ✗ Non-fluent: A street sign is parked on left and a car on the right. |
| ✓ Bear playing hockey. (Although this phenomenon is unlikely to be observed in the real world, one can imagine this phenomenon and hence we consider such descriptions as sensical) | ✗ Non-fluent: It heaving at a city. |
| ✓ A toy car on the left and a teddy bear on the right. | ✗ Implausible: The bush speaking in the garden. |
| ✓ An apple on the top shelf and a banana on the bottom shelf. | ✗ Implausible: Olives and grapes inside a plate. |
| ✓ Elephant standing on top of a table. (Although this phenomenon is unlikely to be observed in the real world, one can imagine this phenomenon and hence we consider such descriptions as sensical) | ✗ Implausible: Grass eating horse. |





2. Does the image look natural, i.e. it does not have any major odd phenomenon inconsistent with the reality of the world? When evaluating, please keep these two scenarios in mind:

1. Animatic and artistic painting should be considered as natural as long as they adhere to fundamental real-world principles.
2. Minor deformations of objects can be acceptable as long as they align with the intended concept and do not appear overly unnatural or disruptive to the viewer.

Please review the examples below. We've provided sets of natural-looking images and sets of images that are not considered natural, along with descriptions explaining why each is categorized as such.

| Natural-looking images ✓ | Not natural-looking images ✗ |
|--|--|
|  <p>✓ Despite being a painting, this image feels natural because it follows basic real-world rules and doesn't have any major oddities.</p> |  <p>✗ This image looks unnatural because the person is missing their full body and head.</p> |
|  <p>✓ This image gives off a natural impression, with no significant problems evident.</p> |  <p>✗ This image seems unnatural because it mixes the body of a zebra with the legs of a giraffe.</p> |
|  <p>✓ This image simply looks natural as it doesn't depict any major odd phenomena or disruptions.</p> |  <p>✗ This image seems unnatural because the surface where the three people are sitting doesn't appear to be supporting them, and their hands look deformed.</p> |
|  <p>✓ This image appears natural since it adheres to fundamental principles without major inconsistencies.</p> |  <p>✗ This image seems unnatural because several fruits are stuck together in a disruptive and unexpected manner.</p> |

3. Does the description accurately describe the contents of the image? Please see the examples below, for each image, we have provided one incorrect description and one correct description.

| | |
|---|--|
|  <p>✗ Incorrect description: A large bear sitting and eating a bamboo plant. ✓ Correct description: A large koala sitting and eating a bamboo plant.</p> |  <p>✗ Incorrect description: A male tennis player holding a green racket. ✓ Correct description: A male tennis player holding a red racket.</p> |
|  <p>✗ Incorrect description: A young woman is eating food. ✓ Correct description: A young man is eating food.</p> |  <p>✗ Incorrect description: Several women gathered together posing with pizzas in take out boxes. ✓ Correct description: Several women gathered together posing with cupcakes in take out boxes.</p> |

For a detailed look at the image, please hover over it.

Please answer the questions below to the best of your knowledge. Careless work will be rejected. Thank you for your careful attention to detail and your valuable contribution!

Figure 7: Instructions for naturalness and image-text matching verification.

Instructions

In this HIT, you will be shown an image and an instruction that specifies some edits to be made to the image (we call this "edit instruction"). You will also be shown the edited image. Your task is to evaluate the correctness of the edited image. The edited image is judged to be correct if:

1. it accurately reflects the edit specified in the edit instruction, and
2. it does not contain any other changes beyond those specified in the edit instruction.

If you deem the edited image to be incorrect, please specify why the edited image is not correct.

Please see the examples below to understand the task better:

- Example 1:** **Edit instruction: Change grassy shore to rocky shore.**




Original Image Edited Image

Is the edited image correct?

Yes
 No

In case you selected "No" for the above question, please specify why the edited image is not correct (please select all that apply):

It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
- Example 2:** **Edit instruction: Change tents to yurts.**





Original Image Edited Image

Is the edited image correct?

Yes
 No

In case you selected "No" for the above question, please specify why the edited image is not correct (please select all that apply):

It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
- Example 3:** **Edit instruction: Change three men to three women.**





Original Image Edited Image

Is the edited image correct?

Yes
 No

In case you selected "No" for the above question, please specify why the edited image is not correct (please select all that apply):

It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
- Example 4:** **Edit instruction: Change "A yellow school bus on the right and a green sports car on the left." to "A green sports car on the right and a yellow school bus on the left."**




Original Image Edited Image

Is the edited image correct?

Yes
 No

In case you selected "No" for the above question, please specify why the edited image is not correct (please select all that apply):

It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
- Example 5:** **Edit instruction: Change triangular bathroom sink to round sink.**





Original Image Edited Image

Is the edited image correct?

Yes
 No

In case you selected "No" for the above question, please specify why the edited image is not correct (please select all that apply):

It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
- Example 6:** **Edit instruction: Change Two skateboards to two surfboards.**






Original Image Edited Image

Is the edited image correct?

Yes
 No

In case you selected "No" for the above question, please specify why the edited image is not correct (please select all that apply):

It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
- Example 7:** **Edit instruction: Change "A toy crane is above a toy car." to "A toy car is above a toy crane."**

Original Image Edited Image

Is the edited image correct?

Yes
 No

In case you selected "No" for the above question, please specify why the edited image is not correct (please select all that apply):

It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.

The edited image does not accurately reflect the edit specified in the edit instruction because the toy car is not above the toy crane; it is behind it. Note that, in this example, the original image is also incorrect since the toy crane is not above the toy car, it is behind; but that does not make the original image correct. Both images are incorrect.

For a detailed look at the image, please hover over it.

Please answer the questions below to the best of your knowledge. Careless work will be rejected. Thank you for your careful attention to detail and your valuable contribution!

Figure 8: Instructions for visual edit verification after applying edit instructions.

Instructions

In this HIT, you will be shown a sentence and an instruction that specifies some edits to be made to the sentence (we call this "edit instruction"). Your task is to verify the validity of the edit instruction. The edit instruction is considered valid if all of the following conditions holds for it:

- The suggested edit results in mutually exclusive concepts, e.g., changing dog to puppy is not a valid edit instruction since puppy is a type of dog, and hence they are not mutually exclusive. Please see the examples below to understand mutually exclusive concepts better:
 - Changing "Poodle" to "Bulldog", "Rose" to "Orchid", "woman" to "man", "car" to "truck", "cold drink" to "warm drink", "dog" to "cat", and "sofa" to "bed", are **mutually exclusive** changes.
 - Changing "dog" to "puppy", "Rose" to "flower", "woman" to "person", "car" to "vehicle", and "cold drink" to "drink" are **not mutually exclusive** changes. Please note that any replacement of words with their synonyms is also not considered mutually exclusive, e.g., changing "kids" to "children", changing "drink" to "beverage", and changing "insect" to "bug".
- The suggested edit modifies **only** one aspect of the sentence, chosen from one of the following categories only:
 - Object:** modifying an object (e.g., changing dog to cat).
 - Attribute:** modifying the properties of an object, such as its color, pattern, shape, or material (e.g., changing round table to rectangular table).
 - Counting:** modifying the count of an object (changing three dogs to two dogs).
 - Spatial Relationship:** modifying the spatial relationship between two objects (e.g., changing "A cat to the left of the dog." to "A dog to the left of the cat.").
- The modification type should fall within one of the **four specified categories:** object, attribute, counting, or spatial relationship. For example, changing verbs is not allowed.

If you deem the edit instruction to be invalid, please specify why the edit instruction is not valid.

Please see the examples below to understand the task better:

- Example 1:**
Original sentence: A dog jumps for a frisbee over a pool.
Edit instruction: Change frisbee to ball.
 Is the edit instruction valid?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edit instruction is not valid (please select all that apply):
 It does not result in mutually exclusive concepts.
 It modifies more than one aspect of the sentence within the categories of object, attribute, counting, and spatial relationship.
 The modification type does not fall within any of the four specified categories: object, attribute, counting, or spatial relationship.
- Example 2:**
Original sentence: A picture of a couple who got married.
Edit instruction: Change couple to bride and groom.
 Is the edit instruction valid?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edit instruction is not valid (please select all that apply):
 It does not result in mutually exclusive concepts.
 It modifies more than one aspect of the sentence within the categories of object, attribute, counting, and spatial relationship.
 The modification type does not fall within any of the four specified categories: object, attribute, counting, or spatial relationship.
- Example 3:**
Original sentence: A group of children riding bicycles along a scenic path in the park.
Edit instruction: Change bicycles to bikes.
 Is the edit instruction valid?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edit instruction is not valid (please select all that apply):
 It does not result in mutually exclusive concepts.
 It modifies more than one aspect of the sentence within the categories of object, attribute, counting, and spatial relationship.
 The modification type does not fall within any of the four specified categories: object, attribute, counting, or spatial relationship.
- Example 4:**
Original sentence: A picture of a white cat.
Edit instruction: Change white cat to black dog.
 Is the edit instruction valid?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edit instruction is not valid (please select all that apply):
 It does not result in mutually exclusive concepts.
 It modifies more than one aspect of the sentence within the categories of object, attribute, counting, and spatial relationship.
 The modification type does not fall within any of the four specified categories: object, attribute, counting, or spatial relationship.
- Example 5:**
Original sentence: A purple bicycle chained to a metal tree enclosure.
Edit instruction: Change purple bicycle to red bicycle.
 Is the edit instruction valid?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edit instruction is not valid (please select all that apply):
 It does not result in mutually exclusive concepts.
 It modifies more than one aspect of the sentence within the categories of object, attribute, counting, and spatial relationship.
 The modification type does not fall within any of the four specified categories: object, attribute, counting, or spatial relationship.
- Example 6:**
Original sentence: Three apples sitting in a bowl.
Edit instruction: Change 3 apples to 2 apples.
 Is the edit instruction valid?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edit instruction is not valid (please select all that apply):
 It does not result in mutually exclusive concepts.
 It modifies more than one aspect of the sentence within the categories of object, attribute, counting, and spatial relationship.
 The modification type does not fall within any of the four specified categories: object, attribute, counting, or spatial relationship.
- Example 7:**
Original sentence: A flower vase on the table and a cake on the right.
Edit instruction: Change "A flower vase on the table and a cake on the right." to "A cake on the table and a flower vase on the right."
 Is the edit instruction valid?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edit instruction is not valid (please select all that apply):
 It does not result in mutually exclusive concepts.
 It modifies more than one aspect of the sentence within the categories of object, attribute, counting, and spatial relationship.
 The modification type does not fall within any of the four specified categories: object, attribute, counting, or spatial relationship.

Original sentence: The kids joyfully dancing together in the sandbox.
Edit instruction: Change dancing to playing.
 Is the edit instruction valid?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edit instruction is not valid (please select all that apply):
 It does not result in mutually exclusive concepts.
 It modifies more than one aspect of the sentence within the categories of object, attribute, counting, and spatial relationship.
 The modification type does not fall within any of the four specified categories: object, attribute, counting, or spatial relationship.

Please answer the questions below to the best of your knowledge. Careless work will be rejected. Thank you for your careful attention to detail and your valuable contribution!

Figure 9: Instructions for edit instruction verification.

Instructions

In this HIT, you will be shown a sentence and an instruction that specifies some edits to be made to the sentence (we call this 'edit instruction'). You will also be shown the edited sentence. Your task is to evaluate the correctness of the edited sentence. The edited sentence is judged to be correct if:

1. It accurately reflects the edit specified in the edit instruction, and
2. It does not contain any other changes beyond those specified in the edit instruction.

If you deem the edited sentence to be incorrect, please specify why the edited sentence is not correct.

Finally, you must select the **type** of the edit being specified in the edit instruction, from the following types:

- **Object:** modifying an object (e.g., changing dog to cat).
- **Attribute:** modifying the properties of an object, such as its color, pattern, shape, or material (e.g., changing round table to rectangular table).
- **Counting:** modifying the count of an object (changing three dogs to two dogs).
- **Spatial Relationship:** modifying the spatial relationship between two objects (e.g., changing "A cat to the left of the dog." to "A dog to the left of the cat.")

Please see the examples below to understand the task better:

• **Example 1:**
Original sentence: A black and white cat is sitting in front of a laptop that is on a desk.
Edit instruction: Change laptop to tablet.
Edited sentence: A black and white cat is sitting in front of a tablet that is on a desk.
 Is the edited sentence correct?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edited sentence is not correct (please select all that apply):
 It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
 Finally, select the type of the edit being specified in the edit instruction:
 Object
 Attribute
 Counting
 Spatial Relationship

• **Example 2:**
Original sentence: A white park bench sitting on a middle of a forest.
Edit instruction: Change white park bench to red park bench.
Edited sentence: A red park bench sitting on a middle of a forest.
 Is the edited sentence correct?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edited sentence is not correct (please select all that apply):
 It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
 Finally, select the type of the edit being specified in the edit instruction:
 Object
 Attribute
 Counting
 Spatial Relationship

• **Example 3:**
Original sentence: A dog sleeping on top of a red chair with two books.
Edit instruction: Change two books to one book.
Edited sentence: A red chair without a dog sleeping on it and without the two books nearby.
 Is the edited sentence correct?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edited sentence is not correct (please select all that apply):
 It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
 Finally, select the type of the edit being specified in the edit instruction:
 Object
 Attribute
 Counting
 Spatial Relationship

• **Example 4:**
Original sentence: A bike to the left of a door.
Edit instruction: Change "A bike to the left of a door." to "A door to the left of a bike."
Edited sentence: A door to the left of a bike.
 Is the edited sentence correct?
 Yes
 No
 In case you selected "No" for the above question, please specify why the edited sentence is not correct (please select all that apply):
 It does not accurately reflect the edit specified in the edit instruction.
 It contains other changes beyond those specified in the edit instruction.
 Finally, select the type of the edit being specified in the edit instruction:
 Object
 Attribute
 Counting
 Spatial Relationship

Please answer the questions below to the best of your knowledge. Careless work will be rejected. Thank you for your careful attention to detail and your valuable contribution!

Figure 10: Instructions for textual edit verification.

Instructions

In this HIT, you will be shown a description along with two images. Your task is to select the image that accurately depicts the description. If you believe both images accurately depict the description, please select both. However, if you find that neither image accurately depicts the description, choose none.

Please see the examples below to understand the task better:

• **Example 1:**

Description: A plate that has three donuts on it next to a cup.



First image



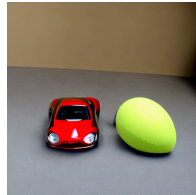
Second image

Which image(s) accurately depicts the description?

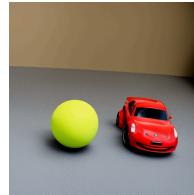
- First image
- Second image
- Both
- None

• **Example 2:**

Description: A ball lying to the left of a toy car.



First image



Second image

Which image(s) accurately depicts the description?

- First image
- Second image
- Both
- None

• **Example 3:**

Description: A person in a kitchen with an open refrigerator.



First image



Second image

Which image(s) accurately depicts the description?

- First image
- Second image
- Both
- None

• **Example 4:**

Description: Two giraffes standing by a barbed wire fence.



First image



Second image

Which image(s) accurately depicts the description?

- First image
- Second image
- Both
- None

For a detailed look at the image, please hover over it.

Please answer the questions below to the best of your knowledge. Careless work will be rejected. Thank you for your careful attention to detail and your valuable contribution!


Figure 11: Instructions for choosing the best matching image.

Instructions

In this HIT, you will be shown an image along with two descriptions. Your task is to select the description that accurately describes the image. If you believe both descriptions accurately describe the image, please select both. However, if you believe neither description accurately describes the image, choose none.

Please see the examples below to understand the task better:


- Example 1:**



Description 1: A cat that is laying down on a laptop.
Description 2: A cat lying down, not on a laptop.

Which description(s) accurately describes the image?


Description 1
 Description 2
 Both
 None
- Example 2:**



Description 1: 2 sheep lay in the grass in a field.
Description 2: One sheep lay in the grass in a field.

Which description(s) accurately describes the image?


Description 1
 Description 2
 Both
 None
- Example 3:**



Description 1: A young man dressed in a blazer, dress shirt and skinny tie.
Description 2: A young man dressed in a jacket dress shirt and skinny tie.

Which description(s) accurately describes the image?

Description 1
 Description 2
 Both
 None
- Example 4:**



Description 1: Two skateboards sitting on the rug in a room.
Description 2: Two surfboards are lying on the rug in a room.

Which description(s) accurately describes the image?

Description 1
 Description 2
 Both
 None

For a detailed look at the image, please hover over it.

Please answer the questions below to the best of your knowledge. Careless work will be rejected. Thank you for your careful attention to detail and your valuable contribution!

Figure 12: Instructions for choosing the best matching description.

A.5 Random qualitative samples from training and VisMin benchmark

In this section, we present random qualitative samples from both the training set and the VisMin benchmark to illustrate the quality and diversity of data used in our experiments. Specifically, we provide examples across object, attribute, relation, and counting categories to highlight the four strategic categories represented in our dataset.

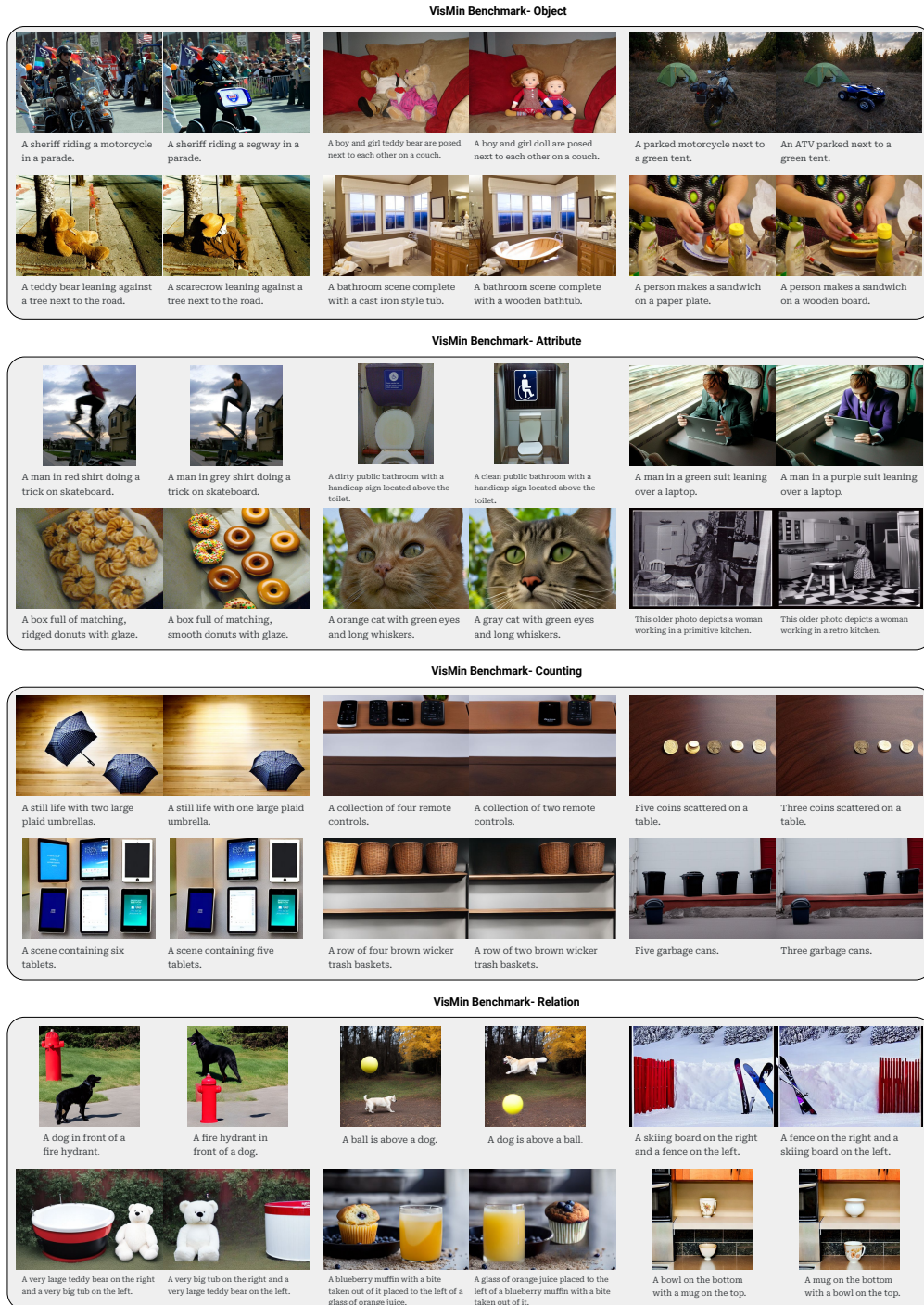
















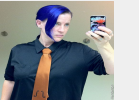









Figure 13: Random qualitative samples from VisMin.













Training- Object

| | | | | | |
|---|---|---|---|---|---|
|  |  |  |  |  |  |
| A man hitting a tennis ball with a tennis racquet. | A woman hitting a tennis ball with a tennis racquet. | Two buses are stopped on the side of the road. | Two taxis are stopped on the side of the road. | A plate of roast beef in barbecue sauce with broccoli and carrots. | A plate of roast beef in barbecue sauce with broccoli and peas. |
|  |  |  |  |  |  |
| The microwave is near the person. | The coffee maker is near the person. | The person is in front of the zebra. | The person is facing the lion. | Woman in red skirt and floral top about to throw a Frisbee. | A woman in a red skirt and floral top is about to throw a football. |

Training- Attribute

| | | | | | |
|---|---|---|---|--|---|
|  |  |  |  |  |  |
| A tv sits on top of a wooden table near plants. | A tv sits on top of a glass table near plants. | A man in a tie has a short stubby beard. | A man in a tie has a long beard. | A woman with blue hair taking a picture of herself with a phone. | A woman with blonde hair taking a picture of herself with a phone. |
|  |  |  |  |  |  |
| A beach scene with a beach chair decorated with the Canadian flag and surfers walking by with their surfboards. | A beach scene with a beach chair decorated with the American flag and surfers walking by with their surfboards. | A cat is curled up on a brown couch in this living room. | A cat is curled up on a green couch in this living room. | A personal pepperoni pizza on a white plate. | A personal veggie pizza on a white plate. |

Training- Counting

| | | | | | |
|---|---|---|---|--|---|
|  |  |  |  |  |  |
| Five cooking pots on a shelf. | Four cooking pots on a shelf. | Two water bottles and a mug on a table. | One water bottle and a mug on a table. | A collection of two wrapped gifts. | A collection of one wrapped gift. |
|  |  |  |  |  |  |
| A scene containing four small blue signs. | A scene containing two small blue signs. | Three antelopes running in a field. | Two antelopes running in a field. | Two yellow teddies. | One yellow teddy. |

Training- Relation






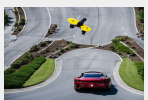



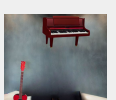
| | | | | | |
|---|---|---|---|---|---|
|  |  |  |  |  |  |
| A laptop is on the bottom and a printer is on the top. | A printer is on the bottom and a laptop is on the top. | A teddy bear below a bed and a night lamp above the bed. | A night lamp below a bed and a teddy bear above the bed. | A plant on the left and a bookshelf on the right. | A bookshelf on the left and a plant on the right. |
|  |  |  |  |  |  |
| A drone on the top and a car on the bottom. | A car on the top and a drone on the bottom. | A sunflower is located to the left of a rose. | A rose is located to the left of a sunflower. | A guitar on the top of a piano. | A piano on the top of a guitar. |

Figure 14: Random qualitative samples from the training set.

Distribution of Categories and Subcategories in the Training Set

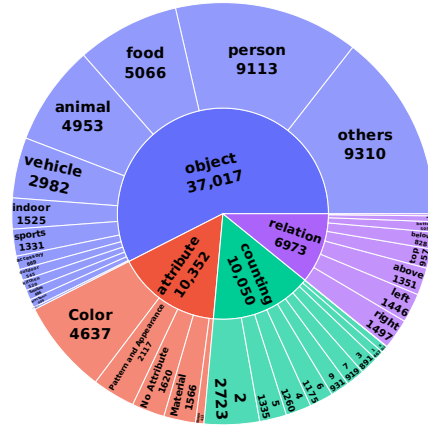


Figure 15: Categories and subcategories distribution in the training split.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

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

Answer: [Yes]

Justification: Yes, the abstract is summary of the content of the paper.

Guidelines:

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

2. Limitations

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

Answer: [Yes]

Justification: In the last paragraph of Conclusion section.

Guidelines:

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

3. Theory Assumptions and Proofs

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

Answer: [NA]

Justification: No theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

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

Answer: [Yes]

Justification: We provide details of training and inference in section 6.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: We will release the data and code and benchmark once accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.

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

6. Experimental Setting/Details

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

Answer: [Yes]

Justification: yes, provided in section 5 and 6.

Guidelines:

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

7. Experiment Statistical Significance

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

Answer: [No]

Justification: We simply benchmark existing available models, with greedy decoding for MLLMs e.g. Idefics2 and fixing seed for foundation models e.g. CLIP.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

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

Answer: [No]

Justification: We perform the task with one gpu, as it's simple benchmarking and finetuning.

Guidelines:

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

9. Code Of Ethics

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

Answer: [Yes]

Justification: Yes, we do.

Guidelines:

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

10. Broader Impacts

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

Answer: [NA]

Justification: No relevant.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [NA]

Justification: No relevant.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: We follow protocol.

Guidelines:

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

13. New Assets

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

Answer: [Yes]

Justification: We provide comprehensive analysis of our new benchmark and will release soon.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [\[Yes\]](#)

Justification: We provide in Appendix.

Guidelines:

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

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

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

Answer: [\[NA\]](#)

Justification: Not relevant.

Guidelines:

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