Supplementary Material
CLASH: A Benchmark for Cross-Modal Contradiction Detection

The supplementary material is organized as follows:

- Implementation details (App. A)
- The prompts for data generation, quality check and model evaluation (App. B)
- Additional experiments and results on CLASH (App. C)
- Comprehensive list of the changed words (from original to conflict in caption) categorized into objects/attributes (App. D and App. E)
- Qualitative examples of CLASH question types (App. F)
- Details about the human validation of the test set (App. G)
- Broader impact (App. H)

A. Implementation details

A.1. Relaxed string matching for multiple choice tasks

Our evaluation protocol employs a hierarchical matching approach that accommodates various response formats. The matching process follows three sequential steps:

(i) **Bracketed choices (strict matching)**: Responses like (A), (B), (C), (D) are matched directly to answer categories. If multiple bracketed options appear, the response is marked incorrect.

(ii) **Single letters**: Responses like A–D (without brackets) are normalized and matched to the expected choices.

(iii) **Keyword matching**: If neither of the above applies, we check whether the response contains the exact answer text as a standalone word (case-insensitive, using word boundaries).

A.2. Relaxed string matching for open-ended tasks

The open-ended question answering task presents models with conflicting image-text pairs and asks them to provide free-form responses explaining what they observe. We classify model responses into four mutually exclusive categories.

Four response categories

1. Conflict detection: responses that correctly identify the contradiction between image and text content.
2. Image-grounded: responses that describe or rely primarily on visual information.
3. Text-grounded: responses that align with or prioritize textual information.
4. Incorrect: responses that are unrelated to the content, incoherent, or fail to address the question meaningfully.

We implement a multi-step normalization pipeline to handle the natural variability in open-ended responses:

- Tokenization: extract word tokens and convert to lowercase.
- Articles removal: filter out articles (the, a, an).
- Number normalization: map numeric tokens to their word equivalents (e.g., "1" → "one").
- Word stripping: remove common linguistic variations using a predefined mapping (e.g., "wooden" → "wood", "brightly" → "bright").
- Stemming: apply Porter stemming to reduce words to their root forms, handling morphological variations.

For each response category, we perform substring matching on the normalized text:

- Contradiction detection: check for presence of stemmed versions of "conflict" or "contradict" using word boundary matching.
- Modality-specific responses: match against normalized versions of the expected image-only or text-only answers for each sample.
- Boundary matching: use regex word boundaries to ensure whole-word matches and avoid partial matches.

The automated evaluation system was manually validated on a subset of responses to ensure classification accuracy.

A.3. Finetuning with LoRA

To avoid the strong bias of flagging conflicts, we pick "conflicting caption" or "original caption" with equal probability during finetuning. The LoRA finetuning is conducted for LLaVA-1.5 and mPLUG-Owl-1, both in 1 epoch. Following their official repository [6], we adopt bf16 and gradient checkpointing for efficiency. All experiments are conducted in a single A100-64GB. Check LoRA settings and training hyperparameters in Table 4.

B. Prompts

B.1. Data generation

In this section, we present the prompts employed for data generation. These prompts were carefully designed to elicit high-quality outputs from the model while controlling for specific linguistic and visual attributes.

[6] LLaVA https://github.com/haotian-liu/LLaVA, mPLUG-Owl-1 https://github.com/X-PLUG/mPLUG-Owl

Table 4. LoRA hyperparameters for multimodal conflict detection finetuning.

	LLaVA-1.5	**mPLUG-Owl-1**
LoRA-r	128	8
LoRA-α	256	32
dropout	0.05	0.05
sequ. length	2048	2048
batch size	16	4
learning rate	2e-5	2e-5
scheduler	cosine	cosine
warmup	0.03 ratio	50 steps

Generate conflicting caption and question

You are an expert image caption editor and question generator. Your task is to modify existing image captions and then create subtle questions based on your modifications. Given an original image caption, you need to perform the following four steps:

Step 1: Create a conflicting caption.

- Take the provided original caption.
- Identify *one* key element (either a specific object or an attribute of an object, like its color, number, shape, material, or texture).
- Change *only this one element* to create a subtle, but noticeable, conflict or discrepancy. The rest of the caption must remain identical to the original.
- Ensure the conflict is a plausible, though incorrect, alternative (e.g., "red car" to "blue car," not "red car" to "flying car").
- Do not change words that have binary states (e.g., man–woman, open–close, dark-light, indoor–outdoor).

When changing an attribute:

- Only change **objective attributes** such as **color, number, shape, material, or texture**.
- Do not change **subjective or ambiguous attributes** such as beautiful, small, large, big, medium, moderate, modern, young, old, fast, slow, elegant, scary, tall, short, etc.

Step 2: Track the changed words.

- Identify the exact word(s) that were changed from the original caption.
- Record both the original word(s) and the replacement word(s).

Step 3: Identify the type of change.

- Determine whether the change made in Step 1 was to an "object" (e.g., "cat" changed to "dog") or an "attribute" (e.g., "white" cat changed to "black" cat).

Step 4: Generate a subtle question.

- Based on your *newly created conflicting caption*, formulate a question.
- This question must subtly hint at the conflicting element without directly stating that something is wrong or different.
- The question should encourage the user to focus on the changed element.

Here is an example. Provide your responses in the exact JSON format shown:

```
User: "A fluffy white cat sitting on a
    red couch."
Model:
{
  "conflicting_caption": "A fluffy black
      cat sitting on a red couch.",
  "question": "What color is the cat
      sitting on a couch?",
  "change_type": "attribute",
  "changed_words": {
    "original": "white",
    "conflicting": "black"
  }
}
```

Generate multiple choice answers

Your task is to write **three answer choices** for a multiple-choice question that highlights a subtle conflict between two captions.

Instructions

Generate the following three answer options:

1. **image_only_answer**: The answer that fits the original caption.
2. **text_only_answer**: The answer that matches the conflicting caption.
3. **irrelevant_but_plausible**: A plausible distractor that doesn't appear in either the original or the conflicting caption, but is contextually reasonable.
   - Make sure the distractor is NOT ambiguous or vague (avoid words such as *several, afternoon, medium, moderate, thing, stuff*).

Ensure that all answers are:

- Concise (preferably 1–3 words),
- Mutually exclusive,
- Plausible in context, but not synonymous to each other.

Only output a JSON with the following fields:

```
{
  "image_only": "...",
  "text_only": "...",
```

```
  "irrelevant_but_plausible": "..."
}
```

B.2. Data filtering

This section describes the data filtering procedures applied to ensure the quality and consistency of the generated dataset. We perform multiple checks, including attribute verification, object validation, answer consistency, and question clarity. Each filtering step is designed to identify and remove entries that are ambiguous, irrelevant, or inconsistent, thereby maintaining the reliability of the dataset for downstream evaluation and analysis.

Attribute check

You are given two words: `original` and `conflicting`. Perform the following two steps:

1. Decide if each word is an **attribute** (a descriptive property, e.g., red, tall, beautiful) or **not an attribute** (an object, e.g., car, tree).
2. If both are attributes, classify them as:
   - **Objective** = measurable, factual, observable (e.g., red, square, wooden, three).
   - **Subjective** = opinion-based or interpretive (e.g., beautiful, small, moderate, large, big, medium, modern, young, old, fast, slow, elegant, scary, tall, short, stylish, fancy, cheap, impressive).

Only output a JSON with the following fields:

```
{
  "change_is_attribute": "Yes/No",
  "change_is_objective": "Yes/No"
}
```

Object check

You are given two words: `original` and `conflicting`.
Your task is to check the quality of the conflicting word in relation to the original:

1. **Object check:** Determine whether each word is an object (a tangible or identifiable thing/entity, e.g., car, apple, chair).
   - Are the two words objects?
2. **Synonymy check:** Is the conflicting object a synonym or near-synonym of the original?
3. **Ambiguity check:** Is the conflicting object ambiguous or vague (e.g., "thing", "object", "stuff")?
4. **Contextual relevance:** Does the conflicting object make sense in the same scene as the original?

Only output a JSON with the following fields:

```
{
  "change_is_object": "Yes/No",
  "change_is_synonym": "Yes/No",
  "change_is_ambiguous": "Yes/No",
  "change_is_relevant": "Yes/No"
}
```

Answers check

You are given three words/phrases. For each word/phrase, check the following:

1. **Synonymy check:** Is one of the words/phrases a synonym or near-synonym of the other two?
2. **Ambiguity check:** Is any of the words/phrases ambiguous or vague (e.g., "several", "afternoon", "medium", "thing")?
3. **Contextual relevance:** Are all words/phrases contextually relevant and objective (not subjective or off-topic)?
4. **Visual check:** Can each word/phrase be directly observed in an image? Examples of visual words include:
   - Attributes of objects (number, color, shape, size, material)
   - Object categories (car, chair, dog)
   - Spatial relations (on top of, next to)
   - Scenes (beach, kitchen)

   Examples of non-visual words include:
   - Temporal concepts (afternoon, tomorrow)
   - Abstract states (freedom, happiness)
   - Non-observable attributes (brand, taste, temperature)
   - Subjective labels (beautiful, boring, large)

```
{
  "ans_is_synonym": "Yes/No",
  "ans_is_ambiguous": "Yes/No",
  "ans_is_relevant": "Yes/No",
  "ans_is_visible": "Yes/No"
}
```

Questions check

You are given a question, a set of changed words, and three possible answers. Your task is to check the question based on three criteria:

1. **Ambiguity check:** Is the **question** clear, specific, and unambiguous?
2. **Focus check:** Does the question explicitly ask about the **changed words**? *Examples:*
   - Question: "What is the color of something?"

Changed words: green, blue → Output: Yes
- Question: "How many items are there?" Changed words: three, four → Output: Yes

3. **Answerability:** Are all three candidate **answers** semantically and contextually compatible with the question? *Examples:*
   - Question: "What is the color?" Answers: green, blue, red → Output: Yes
   - Question: "What is the gender?" Answers: man, woman, child → Output: No

Only output a JSON object with the following fields:

```
{
  "question_is_clear": "Yes/No",
  "question_is_focused": "Yes/No",
  "question_is_answerable": "Yes/No"
}
```

B.3. Model evaluation

This section outlines the procedures used to evaluate model performance across different tasks. We describe the evaluation instructions for multiple-choice and open-ended questions, as well as the framework in which a large language model (LLM) acts as a judge to assess response quality.

Multiple choice instruction

You are given an image and a textual description of the image.
Use *both* the visual and textual information to answer the multiple-choice question below as accurately as possible.

- Choose the option that is best supported by **both sources**.
- If the image and the description contradict each other, or if one is missing key information, select the option "Conflicting information – cannot answer."
- Answer only with the letter corresponding to your choice in parentheses: (A), (B), (C), or (D).
- Do not include any other text.

Open-ended instruction

You are given an image and a textual description of the image.
Use *both* the visual and textual information to answer the question below as accurately as possible.

- Only answer if your response can be clearly supported by **both sources**.
- If the image and the description contradict each other, or if one is missing key information, output "Conflicting information – cannot answer".
- Otherwise, provide the answer (less than 15 words).

LLM-as-a-judge

You are an evaluator. Given two reference answers (`image_only` and `text_only`) and a model prediction, decide which category the prediction belongs to:

1. **IMAGE** — if the prediction semantically matches the `image_only` answer.
2. **TEXT** — if the prediction semantically matches the `text_only` answer.
3. **CONFLICT** — if the prediction explicitly refers to a contradiction, conflict, or states that both cannot be true.
4. **NONE** — if the prediction matches neither answer and does not indicate a conflict.

Ignore minor differences in phrasing, synonyms, plural/singular forms, or capitalization. Return only one label: **IMAGE**, **TEXT**, **CONFLICT**, or **NONE**.
Examples:

```
Image-only answer: polar bear
Text-only answer: brown bear
Prediction: Brown bear
Output: TEXT

Image-only answer: Black
Text-only answer: Blue
Prediction: Conflicting information.
Output: CONFLICT

Image-only answer: dog
Text-only answer: cat
Prediction: dog
Output: IMAGE

Image-only answer: red
Text-only answer: green
Prediction: yellow
Output: NONE
```

C. Experiments

In this section we report additional experiments and detailed analysis of model performance on CLASH. We provide comprehensive results across both multiple-choice and open-ended evaluation formats, including category-specific breakdowns and validation of our evaluation methodology through LLM-as-a-judge assessment.

C.1. Multiple choice question answering

Performance evaluation. Table 5 shows percentage of predictions matching the respective answer for the multiple-choice question answering task using relaxed string matching for evaluation. This format tests models' ability to recognize conflicts when provided with explicit options, including the correct "Conflicting information – cannot answer" choice.

Figure 6 depicts the modality preference of various models, revealing systematic biases toward either visual or textual information. Leading closed-source models (GPT-5, Gemini 2.5 Pro) show minimal bias, while their lighter variants (GPT-4.1 Mini, Gemini Flash Lite) demonstrate strong image preference. Open-source models show varying degrees of modality preference, with some strongly favoring text (InterVL-1.5, LLaVA-1.5-7b) and others showing more balanced distributions.

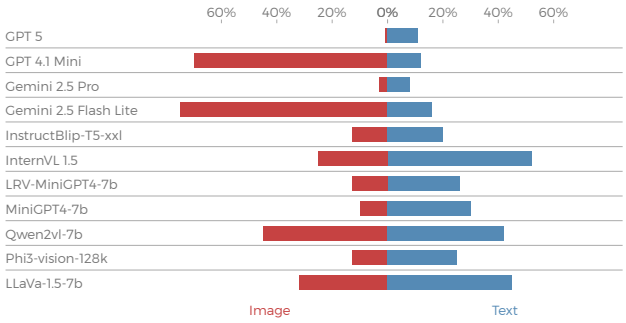


Figure 6. Modality preference patterns in multiple-choice QA. Most models exhibit systematic biases, favoring either visual (red) or textual (blue) information when faced with contradictions.

Evaluation with non-conflicting samples. To better understand the robustness of MM-LLMs in detecting multimodal contradictions, we evaluate models on data that includes both conflicting and non-conflicting samples. This experiment uses the same models and evaluation protocols but incorporates samples where the original caption (without modifications) is presented alongside the image, creating scenarios where no contradiction exists. In these non-conflicting cases, models should respond with the image–grounded answer.

We report the results of this experiment in Table 6, both with strict and relaxed string matching. The "Conflict" column shows accuracy on samples with visual-textual contradictions, "No conflict" column shows accuracy on samples with consistent information, and "Overall" represents the true positive rate across both conflicting and non-conflicting samples. InternVL1.5, despite low conflict detection (16%), achieves exceptional performance on non-conflicting samples (93%), suggesting the model model can appropriately respond to non-conflicting information but fails at conflict identification. Conversely, InstructBlip-T5xxl shows stronger conflict detection (64%) but weaker non-conflict performance (43%). LLaVA-1.5-7b achieves near-zero conflict detection but achieves moderate performance on non-conflicting samples (63.84%).

Comparing the strict vs. relaxed string matching evaluation reveals that most models demonstrate minimal performance changes. However, several models show substantial improvements in non-conflict performance under relaxed matching. Models like LLaVA1.5-7b and Phi3-vision-128k appear to understand task requirements but struggle with strict answer formatting, leading to substantial underestimation of their capabilities under strict evaluation.

Category-specific performance analysis. To understand how different types of contradictions affect performance, we analyze results across semantic categories. Tables 9 and 8 show an overview of the performance split across the object and attribute categories, respectively. This breakdown reveals whether models struggle more with certain types of contradictions (e.g., color vs. environmental characteristics) and helps identify systematic weaknesses in multimodal reasoning capabilities.

Spatial conflicts. While our main benchmark focuses on object- and attribute-level contradictions, we construct a small exploratory subset to assess whether models can detect spatial inconsistencies. We create 271 human-verified spatial conflict examples spanning three relation types: left-right positioning (151 samples), above-below vertical relationships (59 samples), and in-front-of/behind depth ordering (61 samples). These conflicts modify captions to reverse or contradict the spatial relationships present in images. The question and the distractor answer are generated with Gemini 2.5 Flash. Fig. 7 shows representative examples from each spatial conflict type.

Table 7 shows that spatial conflicts are substantially harder than object- and attribute-level inconsistencies for all models. GPT-5 achieves 68.24% spatial conflict detection, a notable drop from its 86.78% performance on object and attribute conflicts (Table 1). Open-source models show even bigger degradation: InstructBlip-T5xxl drops from 63.87% on objects and attributes to 31.23% spatial conflict detection, while InternVL1.5 nearly completely fails at spatial conflicts, achieving only 4.42% detection compared to 16.71% on object and attribute-level conflicts.

These results confirm our design rationale: if models cannot reliably detect "simple" object- and attribute-level mismatches, introducing spatial complexity only amplifies existing failure modes. Moreover, adding spatial conflicts conflates multiple failure modes—models could fail due to object recognition errors, inability to parse spatial relations

Table 5. Percentage of predictions matching the respective answer for the multiple-choice question answering task using relaxed string matching. Last column denotes cases where no match with any of the answers was found.

Model	Conflict (↑)	Image (↓)	Text (↓)	Distractor (↓)	Incorrect (↓)
Phi3-vision-128k	$1.29_{\pm 0.28}$	$27.51_{\pm 1.10}$	$53.65_{\pm 1.35}$	$1.19_{\pm 0.28}$	$16.30_{\pm 0.96}$
MiniGPT4-7b	$2.18_{\pm 0.38}$	$23.50_{\pm 1.13}$	$51.11_{\pm 1.27}$	$9.88_{\pm 0.75}$	$20.50_{\pm 1.04}$
mPLUG-Owl-2	0.00	$0.72_{\pm 0.22}$	$77.92_{\pm 1.06}$	$0.20_{\pm 0.12}$	$21.10_{\pm 1.05}$
LRV-MiniGPT4-7b	$2.16_{\pm 0.39}$	$24.10_{\pm 1.06}$	$42.77_{\pm 1.24}$	$10.69_{\pm 0.78}$	$30.41_{\pm 1.21}$
InternVL 1.5	$17.05_{\pm 0.99}$	$26.35_{\pm 1.12}$	$53.65_{\pm 1.31}$	$2.34_{\pm 0.38}$	$0.80_{\pm 0.23}$
InstructBlip-T5xxl	$63.86_{\pm 1.23}$	$12.67_{\pm 0.84}$	$19.89_{\pm 1.02}$	$3.30_{\pm 0.47}$	$0.06_{\pm 0.06}$
LLaVA-1.5-7b	$0.13_{\pm 0.09}$	$38.98_{\pm 1.23}$	$57.89_{\pm 1.29}$	$3.09_{\pm 0.43}$	0.00
Qwen2vl-instruct-7b	$1.27_{\pm 0.29}$	$50.12_{\pm 1.31}$	$47.64_{\pm 1.27}$	$1.01_{\pm 0.25}$	0.00

Table 6. Performance of various models on multiple-choice QA including conflicting and non-conflicting samples. *Conflict* shows accuracy on contradictory samples, *No conflict* on consistent samples, and *Overall* is the true positive rate across both.

Model	Conflict (↑)	No conflict (↑)	Overall (↑)
Strict string matching			
InstructBlip-T5xxl	$63.54_{\pm 1.36}$	$42.66_{\pm 1.34}$	$53.18_{\pm 0.98}$
InternVL1.5	$16.25_{\pm 1.02}$	$92.51_{\pm 0.73}$	$54.27_{\pm 0.95}$
MiniGPT4-7b	$2.32_{\pm 0.42}$	$35.39_{\pm 1.33}$	$18.80_{\pm 0.78}$
MiniGPT4-7b-LRV	$2.18_{\pm 0.42}$	$36.01_{\pm 1.32}$	$19.04_{\pm 0.77}$
Phi3-vision-128k	$0.94_{\pm 0.27}$	$48.77_{\pm 1.40}$	$24.80_{\pm 0.86}$
Qwen2vl-7b	$1.25_{\pm 0.30}$	$94.55_{\pm 0.65}$	$47.82_{\pm 0.98}$
LLaVA1.5-7b	$0.07_{\pm 0.08}$	$63.84_{\pm 1.36}$	$31.95_{\pm 0.92}$
Relaxed string matching			
InstructBlip-T5xxl	$63.59_{\pm 1.35}$	$42.70_{\pm 1.42}$	$53.17_{\pm 0.98}$
InternVL1.5	$16.71_{\pm 0.99}$	$92.98_{\pm 0.70}$	$54.71_{\pm 1.01}$
MiniGPT4-7b	$2.34_{\pm 0.42}$	$63.13_{\pm 1.33}$	$32.62_{\pm 0.96}$
MiniGPT4-7b-LRV	$2.17_{\pm 0.41}$	$55.71_{\pm 1.36}$	$28.89_{\pm 0.87}$
Phi3-vision-128k	$1.08_{\pm 0.30}$	$89.45_{\pm 0.83}$	$45.13_{\pm 1.02}$
Qwen2vl-instruct-7b	$1.31_{\pm 0.31}$	$97.02_{\pm 0.47}$	$49.04_{\pm 0.96}$
LLaVA1.5-7b	$0.08_{\pm 0.08}$	$90.44_{\pm 0.81}$	$45.24_{\pm 0.99}$

from text, failures in visual grounding of entities, or weakness in comparing relational structures across modalities– undermining the diagnostic value of the benchmark.

C.2. Open-ended question answering

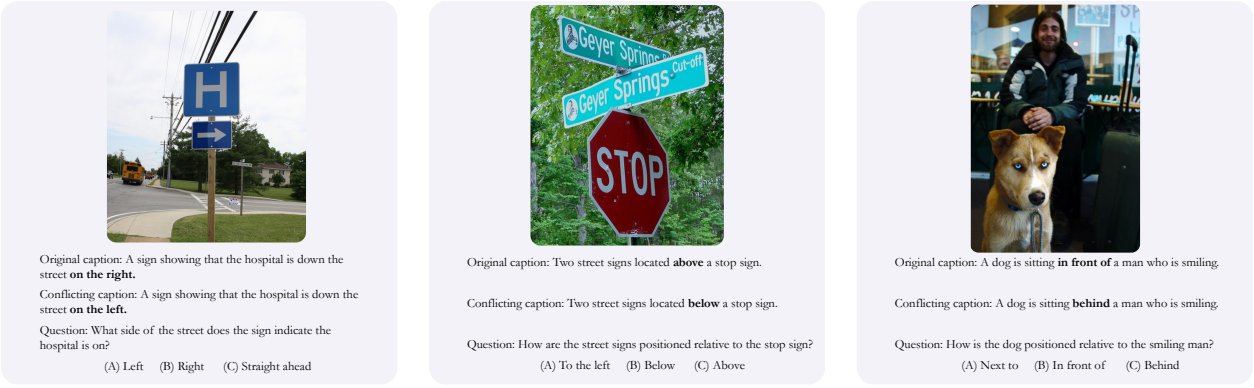
Unlike multiple-choice tasks, where models select from given options, open-ended questions require models to formulate their own responses. This format more closely mirrors real-world deployment scenarios where models must generate explanations or decisions without explicit guidance about potential conflicts.

We evaluate whether models can naturally identify contradictions between visual and textual inputs, or whether they default to following one modality while ignoring the other. The evaluation protocol categorizes responses into four classes: correctly identifying **conflicts**, following **image**-based information, adhering to **text**-based descriptions, or producing **incorrect**/unintelligible responses.

Overall performance on open-ended tasks. This section presents a comprehensive analysis of open-ended conflict detection performance, summarizing key results from the main paper. Table 10 provides the complete breakdown of response categories, while Table 2 and Figure 5 in the main text focus on conflict detection rates and modality bias patterns separately.

LLM-as-judge. To assess the consistency and reliability of our relaxed string matching procedure for evaluating the open-ended task, we ran an LLM-as-a-judge using two models: Gemini 2.5 Pro and GPT-5. For each sample, we determined whether each model's prediction matched one of the four categories: CONFLICT, IMAGE, TEXT, or NONE. We then recorded:

1. **Gemini 2.5 Pro "Yes"** – number of samples where Gemini assigned the category.
2. **GPT-5 "Yes"** – number of samples where GPT-5 assigned the category.
3. **At least one "Yes"** – number of samples where either model assigned the category.
4. **Both "Yes"** – number of samples where both models agreed on the category.



Left vs. right | Above vs. below | In front of vs. behind

Figure 7. Qualitative examples from the spatial conflict subset. We show left-right, above-below, and in-front-of/behind conflicts where captions contradict the spatial relationships visible in the images.

Table 7. Percentage of predictions matching the respective answer on the multiple-choice task using **spatial** conflict types. The last column denotes cases where no match with any of the answers was found. The error bars$_{\pm}$ show standard deviation.

Model	Conflict (↑)	Image	Text	Distractor	Incorrect
GPT 5	$68.24_{\pm 2.92}$	$1.46_{\pm 0.71}$	$17.21_{\pm 2.32}$	0.00	$12.88_{\pm 2.08}$
InstructBlip-T5xxl	$31.23_{\pm 2.83}$	$7.66_{\pm 1.64}$	$43.62_{\pm 2.95}$	$17.39_{\pm 2.34}$	0.00
InternVL1.5	$4.42_{\pm 1.26}$	$21.84_{\pm 2.51}$	$67.80_{\pm 2.89}$	$4.09_{\pm 1.18}$	$1.85_{\pm 0.85}$

This setup allows quantifying both individual model performance and inter-model agreement. Results are presented in Table 11. The results reveal several trends, discussed below.

High consistency across evaluation methods. The comparison between the string matching evaluation in Table 10 and LLM-as-judge evaluation in Table 11 revels remarkably consistent results, demonstrating the reliability of both approaches. Across all models, the differences between string matching and LLM-based evaluation are minimal, typically within 1-3 percentage points. For instance, InternVL 1.5's conflict detection accuracy shows only a 1.79 point difference (18.72% string matching vs. 20.51% Gemini), while modality bias patterns remain nearly identical.

Inter-judge agreement and reliability. The strong agreement between Gemini 2.5 Pro and GPT-5 as judges (differences $< 1\%$ across most metrics) validates the robustness of LLM-based evaluation. This consistency suggests that both judge models apply similar semantic understanding when categorizing responses, reducing concerns about judge-specific biases.

These findings strengthen confidence in our evaluation methodology and suggest that either approach can reliably assess model performance on CLASH.

Performance across object and attribute categories on open-ended tasks. We further analyze category-specific performance in the open-ended setting to understand how different types of contradictions affect free-form reasoning capabilities. This breakdown across object and attribute categories reveals whether the patterns observed in multiple-choice evaluation persist when models must generate their own explanations rather than select from predefined options. Tables 12 and 13 performance of several open source and closed source models on the open-ended task across object and attribute categories. The evaluation is performed using relaxed string matching.

C.3. Finetuning

While our finetuning demonstrates clear improvements on conflict detection in §4.4, we evaluate performance on standard vision-language benchmarks to understand the broader impact of task-specific adaptation. We emphasize that our training data is purposefully constructed for the conflict detection task and does not aim to improve general vision-language capabilities. POPE [24] tests object hallucination via yes/no questions about object presence across random, popular, and adversarial settings, totaling $9k$ samples. OKVQA [31] and GQA [18] ($5k$ and $12k$ samples) measure general visual question answering capabilities requiring external knowledge and spatial reasoning, respectively.

Tables 14 and 15 reveal consistent patterns across benchmarks. mPLUG-Owl-1, which had poor baseline performance, shows consistent improvements across nearly all metrics after finetuning. In contrast, for LLaVA-1.5-7b,

Table 8. Object category performance breakdown for multiple choice QA. We report mean ± standard deviation. Results across five major categories reveal category-specific strengths and weaknesses in multimodal conflict detection. Performance variations suggest that different object types pose varying difficulty levels for conflict detection, potentially due to visual saliency, semantic complexity, or training data distribution.

Model	Animals	Vehicles	Food	Sports	Household	Other
GPT 5	$97.32_{\pm 1.33}$	$93.91_{\pm 2.97}$	$83.88_{\pm 3.92}$	$89.07_{\pm 3.15}$	$78.38_{\pm 3.25}$	$77.58_{\pm 4.15}$
GPT 4.1 Mini	$3.47_{\pm 1.49}$	$10.38_{\pm 3.75}$	$14.89_{\pm 3.72}$	$1.06_{\pm 1.07}$	$8.54_{\pm 2.39}$	$8.80_{\pm 2.75}$
Gemini 2.5 Pro	$98.65_{\pm 0.90}$	$88.06_{\pm 4.03}$	$88.56_{\pm 3.46}$	$89.12_{\pm 3.22}$	$77.74_{\pm 3.60}$	$82.62_{\pm 3.66}$
Gemini 2.5 Flash Lite	$8.25_{\pm 2.28}$	$6.12_{\pm 2.90}$	$10.36_{\pm 3.38}$	$3.37_{\pm 1.91}$	$10.58_{\pm 2.60}$	$4.84_{\pm 2.10}$
InstructBlip-T5xxl	$76.66_{\pm 3.36}$	$71.70_{\pm 5.52}$	$66.60_{\pm 4.98}$	$60.82_{\pm 5.11}$	$64.52_{\pm 3.85}$	$59.88_{\pm 4.83}$
InternVL1.5	$21.16_{\pm 3.31}$	$27.00_{\pm 5.59}$	$27.70_{\pm 4.77}$	$20.91_{\pm 4.31}$	$11.17_{\pm 2.56}$	$18.95_{\pm 3.97}$
Phi3-vision-128k	$2.05_{\pm 1.17}$	0.00	0.00	$1.17_{\pm 1.14}$	$0.70_{\pm 0.73}$	$2.05_{\pm 1.36}$
MiniGPT4-7b	$1.35_{\pm 0.96}$	$2.95_{\pm 2.10}$	$3.41_{\pm 1.93}$	$3.43_{\pm 1.89}$	$2.05_{\pm 1.20}$	$1.94_{\pm 1.36}$
mPLUG-Owl-2	0.00	0.00	0.00	0.00	0.00	0.00
LRV-MiniGPT4-7b	$2.81_{\pm 1.43}$	$1.49_{\pm 1.47}$	$1.15_{\pm 1.17}$	$1.10_{\pm 1.13}$	$1.41_{\pm 1.00}$	$2.01_{\pm 1.40}$
InternVL1.5	$21.16_{\pm 3.31}$	$27.00_{\pm 5.59}$	$27.70_{\pm 4.77}$	$20.91_{\pm 4.31}$	$11.17_{\pm 2.56}$	$18.95_{\pm 3.97}$
LLaVA-1.5-7b	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2vl-instruct-7b	$1.40_{\pm 1.00}$	$1.52_{\pm 1.51}$	0.00	0.00	$2.10_{\pm 1.21}$	$3.84_{\pm 1.97}$

Table 9. Attribute category performance breakdown for multiple choice QA. We report mean ± standard deviation. Results demonstrate how models handle different descriptive properties, from concrete visual attributes (colors, materials) to more abstract characteristics (environmental conditions, physical properties). Notable performance gaps emerge between attribute categories, with colors generally being easier to detect than environmental descriptors.

Model	Colors	Numbers	Materials	Physical	Environmental	Other
GPT 5	$95.49_{\pm 1.24}$	$75.70_{\pm 2.95}$	$93.06_{\pm 3.38}$	$84.04_{\pm 8.09}$	$57.88_{\pm 14.19}$	$73.22_{\pm 8.41}$
GPT 4.1 Mini	$31.59_{\pm 2.62}$	$6.87_{\pm 1.75}$	$11.82_{\pm 4.18}$	$5.06_{\pm 5.31}$	$16.74_{\pm 10.83}$	$10.12_{\pm 5.49}$
Gemini 2.5 Pro	$93.50_{\pm 1.37}$	$85.83_{\pm 2.45}$	$91.24_{\pm 3.67}$	$89.21_{\pm 7.22}$	$66.58_{\pm 13.63}$	$73.34_{\pm 7.89}$
Gemini 2.5 Flash Lite	$9.42_{\pm 1.66}$	$4.42_{\pm 1.47}$	$6.78_{\pm 3.24}$	$20.96_{\pm 9.44}$	$8.63_{\pm 7.99}$	$10.15_{\pm 5.57}$
InstructBlip-T5xxl	$61.35_{\pm 2.84}$	$56.73_{\pm 3.46}$	$63.93_{\pm 6.32}$	$68.03_{\pm 10.83}$	$48.87_{\pm 14.95}$	$47.11_{\pm 8.87}$
InternVL1.5	$16.40_{\pm 2.13}$	$3.94_{\pm 1.44}$	$15.60_{\pm 4.73}$	$20.59_{\pm 9.24}$	0.00	$16.35_{\pm 7.05}$
Phi3-vision-128k	$1.57_{\pm 0.72}$	0.00	0.00	0.00	0.00	$3.31_{\pm 3.29}$
MiniGPT4-7b	$2.58_{\pm 0.89}$	$2.42_{\pm 1.04}$	$1.76_{\pm 1.71}$	0.00	0.00	$3.28_{\pm 3.27}$
mPLUG-Owl-2	0.00	0.00	0.00	0.00	0.00	0.00
LRV-MiniGPT4-7b	$2.27_{\pm 0.87}$	$2.96_{\pm 1.16}$	$1.66_{\pm 1.72}$	0.00	0.00	$6.59_{\pm 4.36}$
LLaVA-1.5-7b	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2vl-instruct-7b	$0.97_{\pm 0.55}$	$0.48_{\pm 0.49}$	0.00	0.00	0.00	0.00

Table 10. Evaluation of models on open-ended QA using relaxed string matching. Columns indicate the percentage of predictions matching Conflict, Image or Text, with Incorrect denoting responses that match none of them.

Model	Conflict (↑)	Image (↓)	Text (↓)	Incorrect (↓)
GPT 5	$81.21_{\pm 1.08}$	$1.11_{\pm 0.30}$	$9.49_{\pm 0.81}$	$8.21_{\pm 0.78}$
GPT 4.1 Mini	$40.03_{\pm 1.37}$	$40.39_{\pm 1.35}$	$19.59_{\pm 1.09}$	$5.53_{\pm 0.61}$
Gemini 2.5 Pro	$91.61_{\pm 0.76}$	$0.84_{\pm 0.25}$	$7.78_{\pm 0.70}$	$0.30_{\pm 0.15}$
Gemini 2.5 Flash Lite	$20.94_{\pm 1.12}$	$49.73_{\pm 1.35}$	$21.64_{\pm 1.18}$	$9.39_{\pm 0.82}$
InternVL1.5	$18.72_{\pm 1.11}$	$22.67_{\pm 1.19}$	$55.51_{\pm 1.42}$	$6.31_{\pm 0.70}$
mPLUG-Owl-1	$6.57_{\pm 0.69}$	$15.91_{\pm 1.03}$	$24.17_{\pm 1.22}$	$61.50_{\pm 1.38}$
mPLUG-Owl-2	$1.40_{\pm 0.33}$	$1.02_{\pm 0.29}$	$97.03_{\pm 0.48}$	$0.93_{\pm 0.26}$
LLaVA-1.5-7b	0.00	$34.67_{\pm 1.29}$	$62.10_{\pm 1.39}$	$4.97_{\pm 0.63}$

finetuning leads to substantial drops on most tasks. This indicates that the specialized training serves as general capability enhancement for initially weak models, while potentially causing capability regression in stronger base-

Table 11. LLM-as-judge evaluation of predictions from four evaluated open-ended models, using Gemini 2.5 Pro and GPT-5. Results show high consistency between judges and close alignment with string matching results (Table 10).

Evaluated Model	Judge	Conflict (↑)	Image (↓)	Text (↓)	Incorrect (↓)
InternVL1.5	Gemini	$20.51_{\pm 1.08}$	$23.83_{\pm 1.15}$	$52.24_{\pm 1.37}$	$3.43_{\pm 0.49}$
	GPT-5	$20.54_{\pm 1.14}$	$23.20_{\pm 1.21}$	$52.04_{\pm 1.38}$	$4.23_{\pm 0.56}$
	≥ 1	$21.33_{\pm 1.09}$	$24.29_{\pm 1.18}$	$52.30_{\pm 1.39}$	$4.43_{\pm 0.58}$
	Both	$19.75_{\pm 1.14}$	$22.81_{\pm 1.20}$	$51.94_{\pm 1.42}$	$3.19_{\pm 0.49}$
mPLUG-Owl-1	Gemini	$2.67_{\pm 0.45}$	$14.24_{\pm 0.99}$	$19.70_{\pm 1.11}$	$61.85_{\pm 1.35}$
	GPT-5	$2.43_{\pm 0.43}$	$13.49_{\pm 0.94}$	$19.23_{\pm 1.07}$	$64.89_{\pm 1.29}$
	≥ 1	$3.79_{\pm 0.53}$	$14.75_{\pm 0.99}$	$20.51_{\pm 1.14}$	$65.95_{\pm 1.40}$
	Both	$1.25_{\pm 0.31}$	$12.97_{\pm 0.95}$	$18.46_{\pm 1.11}$	$60.79_{\pm 1.38}$
mPLUG-Owl-2	Gemini	$1.40_{\pm 0.33}$	$0.78_{\pm 0.25}$	$96.69_{\pm 0.50}$	$1.08_{\pm 0.29}$
	GPT-5	$1.41_{\pm 0.34}$	$0.71_{\pm 0.24}$	$96.79_{\pm 0.47}$	$1.10_{\pm 0.30}$
	≥ 1	$1.38_{\pm 0.34}$	$0.77_{\pm 0.24}$	$97.04_{\pm 0.48}$	$1.33_{\pm 0.32}$
	Both	$1.41_{\pm 0.33}$	$0.71_{\pm 0.23}$	$96.49_{\pm 0.50}$	$0.85_{\pm 0.25}$
LLaVA-1.5-7b	Gemini	$0.24_{\pm 0.13}$	$35.89_{\pm 1.32}$	$60.31_{\pm 1.36}$	$3.56_{\pm 0.51}$
	GPT-5	$0.16_{\pm 0.11}$	$36.11_{\pm 1.36}$	$60.95_{\pm 1.34}$	$2.70_{\pm 0.45}$
	≥ 1	$0.31_{\pm 0.16}$	$36.22_{\pm 1.38}$	$61.17_{\pm 1.36}$	$3.66_{\pm 0.53}$
	Both	$0.08_{\pm 0.08}$	$35.76_{\pm 1.35}$	$60.20_{\pm 1.32}$	$2.64_{\pm 0.46}$

Table 12. Object category performance of several open source and closed source models on the open-ended task using relaxed string matching evaluation. We report mean ± standard deviation.

Model	Animals	Vehicles	Food	Sports	Household	Other
GPT 5	$96.55_{\pm 1.50}$	$89.56_{\pm 3.74}$	$74.73_{\pm 4.55}$	$86.82_{\pm 3.47}$	$64.48_{\pm 4.12}$	$67.53_{\pm 4.42}$
GPT 4.1 Mini	$54.82_{\pm 4.21}$	$43.27_{\pm 6.16}$	$38.86_{\pm 5.29}$	$38.13_{\pm 4.92}$	$33.52_{\pm 3.98}$	$31.38_{\pm 4.53}$
Gemini 2.5 Pro	$98.67_{\pm 0.93}$	$95.58_{\pm 2.39}$	$92.99_{\pm 2.68}$	$95.56_{\pm 2.12}$	$84.72_{\pm 3.00}$	$82.34_{\pm 3.82}$
Gemini 2.5 Flash Lite	$21.24_{\pm 3.41}$	$13.46_{\pm 4.17}$	$11.34_{\pm 3.33}$	$8.67_{\pm 2.93}$	$16.66_{\pm 3.04}$	$20.63_{\pm 4.08}$
InternVL1.5	$26.10_{\pm 3.63}$	$31.37_{\pm 5.59}$	$21.71_{\pm 4.17}$	$31.54_{\pm 4.55}$	$14.64_{\pm 2.93}$	$20.56_{\pm 3.96}$
mPLUG-Owl-1	$4.77_{\pm 1.77}$	$7.36_{\pm 3.01}$	$10.31_{\pm 3.22}$	$3.27_{\pm 1.83}$	$5.55_{\pm 1.94}$	$4.91_{\pm 2.19}$
mPLUG-Owl-2	0.00	0.00	$2.38_{\pm 1.63}$	$1.02_{\pm 1.04}$	$0.70_{\pm 0.69}$	$2.91_{\pm 1.69}$
LLaVA-1.5-7b	0.00	0.00	0.00	0.00	0.00	0.00

Table 13. Attribute category performance of several open source and closed source models on the open-ended task using relaxed string matching evaluation. We report mean ± standard deviation.

Model	Colors	Numbers	Materials	Physical	Environmental	Other
GPT 5	$93.52_{\pm 1.41}$	$67.33_{\pm 3.39}$	$91.24_{\pm 3.81}$	$68.17_{\pm 10.46}$	$49.61_{\pm 12.07}$	$76.63_{\pm 8.31}$
GPT 4.1 Mini	$66.18_{\pm 2.75}$	$3.36_{\pm 1.19}$	$27.33_{\pm 6.00}$	$5.45_{\pm 5.21}$	$31.73_{\pm 11.54}$	$50.01_{\pm 9.71}$
Gemini 2.5 Pro	$96.48_{\pm 1.08}$	$86.50_{\pm 2.34}$	$96.60_{\pm 2.37}$	$89.16_{\pm 7.11}$	$62.86_{\pm 12.15}$	$80.85_{\pm 7.62}$
Gemini 2.5 Flash Lite	$34.21_{\pm 2.65}$	$16.10_{\pm 2.58}$	$22.24_{\pm 5.63}$	$16.05_{\pm 8.49}$	$12.22_{\pm 8.40}$	$26.83_{\pm 8.70}$
InternVL1.5	$16.80_{\pm 2.06}$	$13.15_{\pm 2.39}$	$10.32_{\pm 3.87}$	0.00	$18.91_{\pm 9.80}$	$7.74_{\pm 5.07}$
mPLUG-Owl-1	$7.79_{\pm 1.47}$	$5.81_{\pm 1.60}$	$6.87_{\pm 3.29}$	$15.73_{\pm 8.64}$	0.00	$3.87_{\pm 3.87}$
mPLUG-Owl-2	$0.31_{\pm 0.31}$	$1.46_{\pm 0.85}$	$10.09_{\pm 3.89}$	$5.17_{\pm 5.13}$	0.00	0.00
LLaVA-1.5-7b	0.00	0.00	0.00	0.00	0.00	0.00

line models. We note that benchmarks like OKVQA and GQA provide their own training sets that could be used for finetuning – for building a general-purpose expert model, we would recommend finetuning on a mixture of diverse datasets rather than solely on CLASH.

These results serve as a reminder that task-specific finetuning, while effective for the target task, may require additional considerations – such as multi-task training or regularization strategies – to maintain performance on out-of-distribution evaluation scenarios during deployment.

C.4. Spatial conflicts

The open-ended evaluation on spatial conflicts presented in Table 16 reveals similar patterns to the multiple-choice results in App C.1, with even more pronounced degradation. GPT-5 conflict detection drops from 81.21% on ob-

Table 14. Accuracy (%) on POPE-COCO across random/popular/adversarial settings. We report mean ± standard deviation computed via bootstrap resampling with 1000 iterations.

Model	Random	Popular	Adversarial
LLaVA-1.5-7b	$89.60_{\pm 0.31}$	$86.22_{\pm 0.40}$	$79.68_{\pm 0.50}$
LLaVA-1.5-7b-ft	$83.42_{\pm 0.44}$	$84.06_{\pm 0.49}$	$81.02_{\pm 0.55}$ ↑
mPLUG-Owl-1	$35.32_{\pm 0.73}$	$31.98_{\pm 0.71}$	$32.28_{\pm 0.72}$
mPLUG-Owl-1-ft	$53.07_{\pm 0.81}$ ↑	$46.91_{\pm 0.84}$ ↑	$45.37_{\pm 0.85}$↑

Table 15. Performance on OKVQA and GQA.

Model	OKVQA	GQA
LLaVA-1.5-7b	$60.84_{\pm 0.56}$	$58.96_{\pm 0.21}$
LLaVA-1.5-7b-ft	$43.93_{\pm 0.54}$	$44.58_{\pm 0.20}$
mPLUG-Owl-1	$35.39_{\pm 0.54}$	$28.05_{\pm 0.17}$
mPLUG-Owl-1-ft	$44.18_{\pm 0.55}$ ↑	$35.06_{\pm 0.18}$ ↑

ject and attribute conflicts to 56.20% on spatial relations. InternVL1.5 also shows a decrease in conflict detection (18.72% to 9.58%). InstructBlip-T5xxl achieves near-zero conflict detection (0.36%), predominantly producing text-based answers (58.98%) or invalid outputs (34.38%). These results demonstrate that spatial reasoning failures are consistent across evaluation formats.

D. Object Categories

This section presents the object categories used in our evaluation framework. The categories are organized into four main domains: animals, transportation, food, sports, and household items.

D.1. Animals

- **Domestic/Farm Animals:** dog, cat, sheep, cow, cows, horse, horses, pig, goat, goats, cattle, donkeys, chicken, bull
- **Wild Animals:** elephant, zebra, zebras, giraffe, giraffes, bear, rhino, rhinoceros, rhinoceroses, rhinos, bird, birds, monkey, camel, antelope, deer, bison, wildebeest, hippopotamus, hippos, polar bear, brown bear, parrot, owl, crow, pigeon, butterfly, octopus, shark, fish, worm
- **Multiple/General:** elephants, dogs, cats, ducks, animals, kitten, puppy

D.2. Transportation

- **Land Vehicles:** skateboard, skate board, bicycle, car, bus, scooter, motorcycle, train, truck, cars, tractor, automobile, motorcycles, bicycles, scooters, fire truck, police car, snowmobile, tow truck, pickup truck, train car, tour buses, bullet train
- **Air Vehicles:** plane, airplane, air plane, helicopter, fighter jet, commercial plane, fighter jets, commercial jets
- **Water Vehicles:** boat, boats, surf board, surfboard, surf boards, kayak, wakeboard
- **Transportation-Related:** bike, bikes, commercial jet, engine, trunk, trucks, road, tracks, track, windsail, parking meter

D.3. Food

- **Fruits:** apples, bananas, banana, fruits, fruit, oranges, apple, pear, strawberries, blueberries, strawberry, banana peel, apple core
- **Vegetables:** vegetables, carrots, potatoes, broccoli, olives, tomatoes, carrot, cauliflower, onion rings, mushrooms, peppers, peas, spinach
- **Prepared Food:** pizza, burger, burgers, cake, hot dog, hot dogs, pastry, sandwich, donuts, cookies, food, noodles, hamburgers, hotdogs, cheese, sauce, sandwiches, quiche, meat, mead, beef, eggs, pasta, french fries, bread, hamburger, rice, cheeses, meats, fries, rice cake, cookie, pickle, piece of cake, slice of pizza, sausage
- **Drinks:** wine, beer, coffee, wine bottle, beer bottle, coffees, drinks, milk
- **Food Descriptors/Toppings:** toppings, sauces, greens, pepperoni
- **Food-Related Items:** bar b que, blender, bottle, bottles, bowls, chicken, dishes, fork, knife, olive, oysters, pears, pie, spoon, water, snails

D.4. Sports

- **Sports Equipment:** ball, frisbee, snowboard, snow board, skis, ski, snowboards, baseball bat, baseball, tennis racket, tennis racquet, basketball, kite, kites, tennis ball, football, glove, bat, tennis, surfboard, surf board, surf boards, soccer, soccer ball, soccer balll, golf club, golf, golf ball, racket, racquet, shuttlecock, a snowboard, skateboard, skate board, bicycle
- **Sports Participants:** snowboarder, skier, skiier, batter, catcher, snowboarders, skiers, skateboarder, cyclist, surfer, kayaker, tennis players, basketball players, tennis player, basketball player, baseball player, football player, baseball players, football players, umpire, skateboarders, cyclists
- **Sports Venues:** tennis court, basketball court, skate park, ski lift
- **Sports-Related:** base ball, cricket, pitch, pitcher, ski lift, skis, surfer, track, wakeboard

D.5. Household Items

- **Furniture:** chair, table, bed, sofa, couch, bench, coffee table, dining table, computer desk, nightstand, shelf, counter
- **Room Identifiers:** kitchen, bathroom, bedroom, living room, dining room
- **Bathroom Items:** toilet, sink, bathtub, shower, tooth-

Table 16. Percentage of predictions matching the respective answer on the open ended task using **spatial** conflict types. The last column denotes cases where no match with any of the answers was found. The error bars$_{\pm}$ show standard deviation.

Model	Conflict (↑)	Image	Text	Incorrect
GPT 5	$56.20_{\pm 3.01}$	$0.77_{\pm 0.54}$	$11.34_{\pm 1.92}$	$31.61_{\pm 2.91}$
InstructBlip-T5xxl	$0.36_{\pm 0.37}$	$17.44_{\pm 2.39}$	$58.98_{\pm 2.97}$	$34.38_{\pm 2.81}$
InternVL1.5	$9.58_{\pm 1.78}$	$10.01_{\pm 1.79}$	$48.41_{\pm 3.07}$	$32.89_{\pm 2.76}$

brush, tooth brush, toilet tissue, soap, mirror, toilet bowl, shower curtain, towel rack, handicap bar
- **Kitchen Items:** fork, bowl, bowls, spoon, plate, plates, knife, knif, cup, trays, pots
- **Storage/Containers:** bag, trash can, laundry basket, baskets, mason jar, coffee cup, bottle, bottles, vases
- **Technology/Electronics:** phone, cell phone, laptop, laptops, keyboard, mouse, tablet, tablets, television, camera, phones, cellphones, wii, wii console, playstation, playstation console, xbox, remote, game remote, controller, refrigerator, oven, stove, microwave, dishwasher, washer, dryer, clothes washer, clothes dryer, blender, ice machine, coffee machine, printer, monitors, screen, clock, bell, ipod, microphones, speakers, equipment
- **Decor/Furnishing:** lamp, paintings, painting, picture frame, pillow, carpet, rug, sculptures, sculpture, statues, statutes, crosses, flags, flag
- **General Household:** furniture, window, door, towels, dishes, appliances, comb, rope, chain, scarf, mask, ties, scarves, backpack, coat, shirt, belt, hairbrush, aluminum foil, plastic wrap, stand, cart, books, book, glasses, handle, backrest, toys, toy, doll, accessories, clothing, swimsuit, dress, skirts, pants, roses, tulips, flowers, plants, leaves, branches, umbrella, umbrellas, hat, changing table, fire place, fireplace, pacifier, refrigerator magnet, urinal

E. Attribute Categories

This section presents the comprehensive attribute categories used in our evaluation framework. The attributes are organized into five main domains: colors, numbers, materials, physical properties, and environmental conditions.

E.1. Colors

- **Single Colors:** blue, white, red, black, brown, green, yellow, orange, pink, grey, gray, purple, silver, tan, beige, cream, gold
- **Color Combinations:** black and white, blue and white, black and yellow, green and yellow, brown and white, black and red, black and gray, white and gray
- **Color Descriptors:** light blue, dark red, mint green, colorful, rainbow colored, monochrome, dark, light, color, colored, different colors, colors, browns, whites, rosy, colorfully, red-haired, blonde-haired, ginger, creamy

E.2. Numbers

- **Basic Numbers:** one, two, three, four, five, six, seven
- **Written Numbers:** 2, 3, 25, 50
- **Ordinals:** first, second, third
- **Quantities:** a, another, solo, whole
- **Prices:** 11.98, 10.99

E.3. Materials

- **Materials:** wooden, wood, metal, plastic, glass, ceramic, concrete, stainless steel, tile, brick, cement, marble, leather, fabric, steel, granite, stone, plywood, paper
- **Surface Qualities and Textures:** striped, polka dot, polka-dotted, polka dotted, tiled, plain, painted, polished, scratched, printed, stripped

E.4. Physical Properties

- **Shapes:** square, round, circular, oval, rectangular, triangular
- **Physical Descriptors:** thick, thin, stuffed, sliced, ripe, unripe, wet, dry, clean, muddy, squares, wedges, opaque, clear, edge, back, duck shaped, fish shaped, horned, antlered

E.5. Environmental Conditions

- **Weather:** sunny, snowy, cloudy, overcast, stormy, wet
- **Landscape/Terrain:** grassy, grass covered, snow covered, rocky, sandy, lush, dry, desert, tropical, remote, green, fenced
- **Water Depth:** knee deep, ankle deep
- **Light Conditions:** dim, bright

F. Qualitative examples

In this section, we present qualitative examples from CLASH, illustrating the variety of object and attribute categories. The dataset covers five object categories – animals, transportation, food, sports, and household items, and five attribute categories – colors, numbers, materials, physical properties, and environmental conditions. Fig. 8 shows qualitative examples from each category in CLASH.

G. Human validation

Annotators assess three components for each sample: conflicting captions (verifying single-element modifications involving plausible, objective properties while avoiding impossible or subjective changes like man-to-woman), questions (confirming clarity, unambiguity, and focus on the changed element), and answers (checking for distinctiveness, objectivity, and visual observability while identifying problematic vague terms like ”medium” or ”beautiful”). Based on this assessment, annotators provide a single accept/reject decision for each sample. The exact instructions provided to the annotators are shown below, while Fig. 9 depicts a few examples of accepted and rejected samples. Fig. 10 shows examples from the human verification interface. Annotators evaluate each sample using binary accept/reject votes to ensure the benchmark's reliability.

Human validation instructions

You will evaluate three things for each example:

1. **Conflicting Caption**
   - Did the caption change only one clear attribute or object (the change is marked in bold)?
   - Is the change plausible and objective (e.g., color, number, shape, material, texture)?
2. **Question**
   - Is the question clear and unambiguous?
   - Does it focus on the changed attribute or object?
   - Can it be answered using the given options?
3. **Answers**
   - Are the answers distinct and not synonyms?
   - Are they objective and visual (e.g., colors, numbers, objects)?

Your task: For each sample, mark whether it is sensible (Yes/No).

H. Broader impact

This work contributes to the development of more reliable multimodal AI systems by exposing critical limitations in conflict detection capabilities. Improved conflict detection could enhance AI safety in applications like medical diagnosis, autonomous systems, and content verification. However, the focus on synthetic contradictions may not fully represent the complexity of real-world misinformation or adversarial scenarios. We encourage future work to extend these findings to more diverse contradiction types and real-world deployment contexts.

Original caption: A herd of **sheep** are slowing drivings down on the road.

Conflicting caption: A herd of **cows** are slowing drivings down on the road.

Question: What type of animals are causing a delay for the drivers on the road?

(A) Sheep (B) Cows (C) Goats

Animals

Original caption: A vintage **automobile** sitting on the flat bed of a tow truck.

Conflicting caption: A vintage **motorcycle** sitting on the flat bed of a tow truck.

Question: What type of vehicle is being transported on the flat bed?

(A) Automobile (B) Motorcycle (C) Van

Vehicles

Original caption: A large amount of **bananas** that are sitting inside of a large tin.

Conflicting caption: A large amount of **apples** that are sitting inside of a large tin.

Question: What fruit is plentifully places within the large tin?

(A) Pears (B) Bananas (C) Apples

Food

Original caption: Guy stands in the snow posing with his **snowboard**.

Conflicting caption: Guy stands in the snow posing with his **skis**.

Question: What kind of equipment is the guy posing with in the snow?

(A) Snowboard (B) Snowshoes (C) Skis

Sports

Original caption: A bedroom with a bed and a **computer desk.**

Conflicting caption: A bedroom with a bed and a **nightstand.**

Question: What type of furnishing is situated beside the bed?

(A) Computer desk (B) Nightstand (C) Dresser

Household items

Original caption: A man standing beside a **robot** with a camera around his neck.

Conflicting caption: A man standing beside a **dog** with a camera around his neck.

Question: What kind of creature is standing next to the man?

(A) Robot (B) Cat (C) Dog

Other

Original caption: A **red** fire hydrant outside of an old brick building.

Conflicting caption: A **yellow** fire hydrant outside of an old brick building.

Question: What is the color of the fire hydrant in front of the brick structure?

(A) Red (B) Blue (C) Yellow

Colors

Original caption: **Two** giraffes are standing near each other in tall bush.

Conflicting caption: **Three** giraffes are standing near each other in tall bush.

Question: How many giraffes are visible among the tall bush?

(A) Two (B) Four (C) Three

Numbers

Original caption: A woman holding her hand out over a **wood** fence.

Conflicting caption: A woman holding her hand out over a **metal** fence.

Question: What material is the fence made of?

(A) Wood (B) Metal (C) Stone

Materials / Texture

Original caption: A pizza with various toppings is sliced into **wedges.**

Conflicting caption: A pizza with various toppings is sliced into **squares.**

Question: What gemetric form do the pizza slices take?

(A) Rectangles (B) Wedges (C) Squares

Physical properties

Original caption: A crowd in a city on a **sunny** day.

Conflicting caption: A crowd in a city on a **cloudy** day.

Question: What is the atmospheric condition in the city where people are gahered?

(A) Rainy (B) Sunny (C) Cloudy

Environmental conditions

Original caption: A **double decker** bus is parked near the curb.

Conflicting caption: A **single decker** bus is parked near the curb.

Question: What type bus is situated by the curb?

(A) School bus (B) Single decker (C) Double decker

Other

Figure 8. Qualitative examples from CLASH, illustrating the diversity of object and attribute categories. The dataset spans five object categories (top two rows): animals, transportation, food, sports, and household items, and five attribute categories (bottom two rows): colors, numbers, materials, physical properties, and environmental conditions.

(a) Changed object (b) Changed attribute (color) (c) Changed attribute (number)

(d) Problematic conflicting word (e) Problematic question (f) Problematic answers

Figure 9. Examples of accepted (**top row**) and rejected (**bottom row**) samples during human validation. Positive examples illustrate cases where the conflicting caption, question, and answers are clear and unambiguous. Negative examples highlight typical sources of rejection: (1) conflicting words, e.g., the change is ”cake → table“ but the image contains both a table and a cake; (2) problematic questions, e.g., the change is ”baby → dog“ but the question implies an animal; and (3) problematic answers, e.g., the color of the field could be described as being ”red“ (distractor answer).

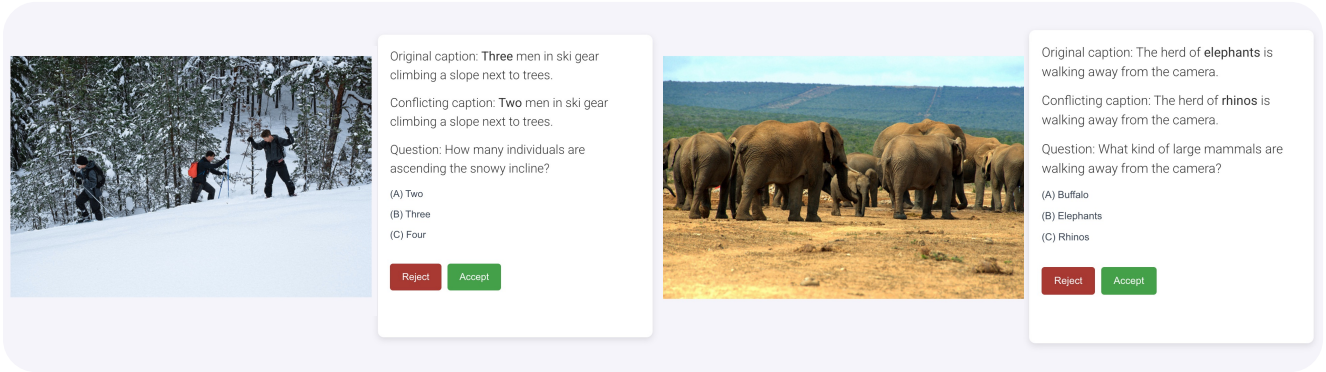


Figure 10. Examples from the human verification interface. Each sample includes an original caption from MS COCO, a conflicting caption that introduces a controlled contradiction, a targeted question designed to test conflict detection, and multiple-choice answers. Annotators use binary accept/reject voting to validate the quality and clarity of each sample, ensuring the reliability of the benchmark’s diagnostic test set.