

FINEMATCH: Aspect-based Fine-grained Image and Text Mismatch Detection and Correction (Supplementary Material)

Hang Hua¹, Jing Shi², Kushal Kafle², Simon Jenni², Daoan Zhang¹,
John Collomosse², Scott Cohen², and Jiebo Luo¹

¹ University of Rochester

² Adobe Research

{hhua2, jluo}@cs.rochester.edu, dzhang52@ur.rochester.edu
{jingshi, kkafle, jenni, collomos, scohen}@adobe.com

1 Explanations and Examples for Mismatch Aspects in FINEMATCH

We present the definition of the four aspects in FINEMATCH. **Entities** are objects within an image that can be recognized by VLMs. In FINEMATCH, we define the entities using the same concepts in Flickr 30K Entities [2]. The terms **Attributes** and **Relations** follow the definitions provided by the Visual Genome (VG) Attribution/Relation framework [1]. However, the categories are limited in VG, while in FINEMATCH the attributes and relations are open-set. **Numbers** refers to the count of entities in the image. We also provide the examples for each aspect in Figure 1.

2 Statistical Results for FINEMATCH

We provide extended statistical results for the quantity and quality analysis across all data sources within FINEMATCH. Figure 2 illustrates the weights and relations distribution among the data source, data domain, and the number of mismatched aspects within the captions. Additionally, Figure 3 shows the distribution of the CLIP scores between the captions and the images in FINEMATCH. Analysis from Figure 3 indicates that the average CLIP score across the three data sources ranges between 0.3 and 0.35. This suggests a comparatively high similarity level between the mismatched captions and the images, making it challenging for models to discern discrepancies between the captions and images.

3 Evaluation Framework

In this study, the BERT Score is computed using the `bert-base-uncased` model weights, with the threshold T set as 0.55 for these conditions. Below, we present the pseudo-code for ITM-IoU. The notation for all variables is consistent with that outlined in Section 3.



Caption: A woman is standing under a pink umbrella.
Numbers, one (number of woman), two (number of women)



Caption: A brightly colored military helicopter flying above a city.
Entities, city, airport;
Relations, flying above (helicopter flying above city), sitting at (helicopter sitting at airport);
Attributes, brightly colored (helicopter is brightly colored), camouflaged (helicopter is camouflaged)

Fig. 1: Illustration of the four mismatch aspects in FINEMATCH.

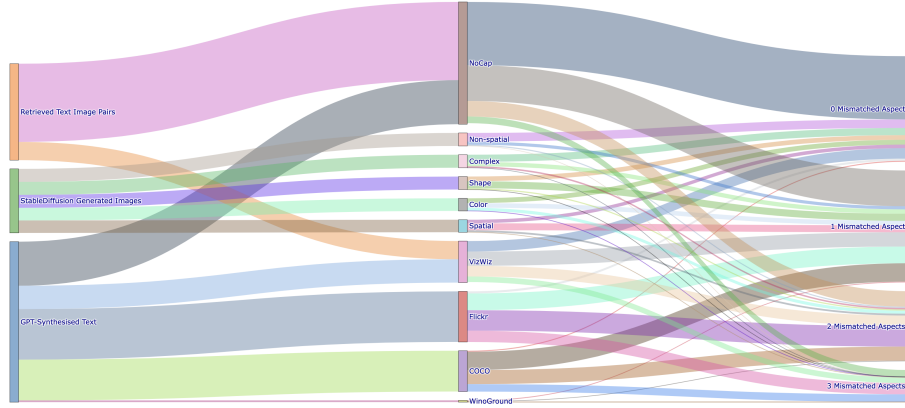


Fig. 2: Distribution of demonstrations in FINEMATCH across source, domain, and the number of mismatched aspects in the captions.

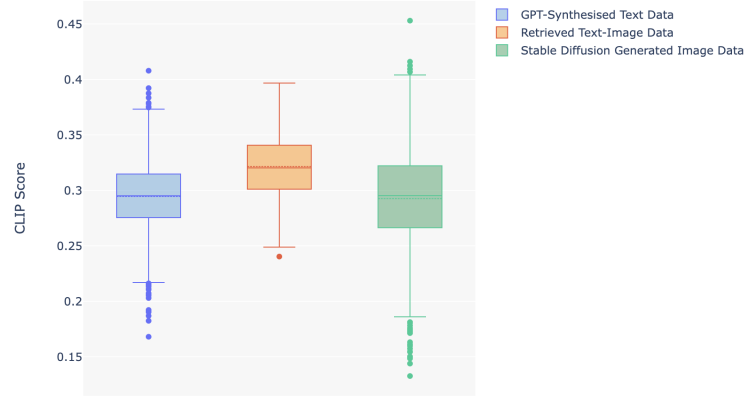


Fig. 3: CLIP score distribution across different data sources in FINEMATCH.

Algorithm 1 ITM-IoU

Input: The predicted aspect tuples/triplets $P_i = \{c_j, p_j, o_j\}_{j=1}^M$ over the test set and the ground truth $G_i = \{c'_j, p'_j, o'_j\}_{j=1}^{M'}$, $i \in \{1, 2, \dots, |\mathcal{D}|\}$.

Output: The average ITM-IoU score.

```

1:  $Score_{ITM-IoU} \leftarrow \{\}$ 
2: for each  $P_i, G_i$  do
3:    $Scores \leftarrow \{\}$ 
4:    $Score_{Aspect_i} \leftarrow \{\}$ 
5:   for  $j = 0$  to  $M'$  do
6:      $Score \leftarrow 0$ 
7:     if  $c_j = c'_j$  then
8:        $Score \leftarrow W_{Ca}$ 
9:     end if
10:    if Mismatch Correction then
11:       $Score \leftarrow Score + W_{De} \times Score_{Dj} + W_{Co} \times Score_{Cj}$ 
12:    else
13:       $Score \leftarrow Score + W_{Co} \times Score_{Cj}$ 
14:    end if
15:     $Score_{Aspect_i} \leftarrow Score_{Aspect_i} \cup Score$ 
16:  end for
17:  if  $\max(Score_{Aspect_i}) \geq T$  then
18:     $Scores \leftarrow Scores \cup \max(Score_{Aspect_i})$ 
19:    Calculate  $P_i \cap G_i$ 
20:  else
21:     $Scores \leftarrow Scores \cup 0$ 
22:  end if
23:  Calculate  $P_i \cup G_i$ 
24:   $Score_{ITM-IoU} \leftarrow \text{mean}(Scores) \times \frac{|P_i \cap G_i|}{|P_i \cup G_i|}$ 
25: end for
26: return  $\text{mean}(Score_{ITM-IoU})$ 

```

4 Prompts for Aspect Graph Parsing and Node Replacing for GPT-4V

Figures 4 and 5 illustrate the instructions and context examples for aspect graph parsing and node replacing for GPT-4V, respectively.

5 Details for Experiments

5.1 Details for Supervised Learning Experiments

We conduct supervised learning experiments on FINEMATCH under the visual instruction tuning settings, where we design the instructions to train models for fine-grained image-text mismatch detection and correction. In the experiments, we set the batch size $\in \{128, 256, 512\}$, the learning rate $\in \{1e-5, 3e-5, 5e-5\}$, and we train each model for 3 epochs. We maintain all other hyperparameters at their default values as specified in the official code base. All the models are finetuned on 8 Nvidia A100 GPUs. The predictions of each model are post-processed for calculating ITM-IoU.

5.2 Prompts for In-Context Learning Experiments

We present the prompts for both GPT-4V and Gemini Pro Vision in Figure 6 for the in-context learning experiments. For each other white-box model, we provide 6 context examples randomly sampled from the training set.

6 Examples for AUTOALIGN

In this section, we provide examples to illustrate that FINEMATCH can help reduce the hallucination for T2I generation. Figure 7 and Figure 8 show examples of single-turn mismatch detection and correction. Figure 9 shows the examples for multi-turn mismatch detection and correction. We also provide the image editing generation prompt for GPT-4V in Figure 10.

7 Human Annotation

We design an annotation interface for human experts to annotate the mismatched aspects between the images and the captions. The interface is shown in Figure 11. The entire annotation process is conducted on the LabelBox platform.

The given caption describes a scene of a photo.

Your goal is to parse the caption to a scene graph that contains all the entities, relations, attributes, and numbers (Non-specific numeral-related words should be only included in the 'Attribute' column) in the image.

The parsing requirements are:

All the parsed entities, relations, attributes, and numbers should be atomic concepts (noun word, adjective word, etc.).

Follow the JSON format like the given reference examples for the target caption.

The passive of some relations, for example, the relation 'carry' and 'carried by', just keep the active relation, the carried by should be removed.

The caption may need the coreference resolution process.

Reference Examples:

Caption: A group of women is playing the piano in the room.

Scene graph: {'Entities': ['woman', 'room'],

'woman': {'Relations': ['play piano', 'in, the room'], 'Attributes': ['a group of, Non-specific quantity of woman'], 'Numbers': ['None']},

'room': {'Relations': ['None'], 'Attributes': ['None'], 'Numbers': ['None']}]}

Caption: A Chihuahua runs after a child on a bicycle.

Scene graph: {'Entities': ['Chihuahua', 'child', 'bicycle'],

'Chihuahua': {'Relations': ['runs after, Chihuahua runs after child'], 'Attributes': ['Chihuahua, Chihuahua is a breed of dog'], 'Numbers': ['one, number of Chihuahua']},

'child': {'Relations': ['on, child on bicycle'], 'Attributes': ['None'], 'Numbers': ['one, number of child']},

'bicycle': {'Relations': ['on, child on bicycle'], 'Attributes': ['None'], 'Numbers': ['one, number of bicycle']}]}

Caption: A Delta Boeing 777 taxiing on the runway.

Scene graph: {'Entities': ['Boeing 777', 'runway'],

'Boeing 777': {'Relations': ['taxiing on, the runway'], 'Attributes': ['Boeing 777 belongs to Delta'],

'Numbers': ['one, A Delta Boeing 777 is one plane']},

'runway': {'Relations': ['on, plane on the runway'], 'Attributes': ['None'], 'Numbers': ['None']}]}

Caption: An office kitchen with open windows and no food.

Scene graph: {'Entities': ['office kitchen', 'windows', 'food'],

'office kitchen': {'Relations': ['has, office kitchen has windows', 'no, office kitchen has no food'],

'Attributes': ['None'], 'Numbers': ['one, number of office kitchen']},

'windows': {'Relations': ['in, windows in the office kitchen'], 'Attributes': ['open, (windows are open)'],

'Numbers': ['None']},

'food': {'Relations': ['None'], 'Attributes': ['None'], 'Numbers': ['zero, no food']}]}

Caption: A kitchen with wooden furniture and a vase missing red flowers.

Scene graph: {'Entities': ['kitchen', 'furniture', 'vase', 'flowers'],

'kitchen': {'Relations': ['has, (kitchen has wooden furniture)', 'has, kitchen has vase'], 'Attributes':

['None'], 'Numbers': ['one, number of kitchen']},

'wooden furniture': {'Relations': ['in, wooden furniture in the kitchen'], 'Attributes': ['wooden, furniture is wooden'], 'Numbers': ['None']},

'vase': {'Relations': ['in, vase in the kitchen', 'missing, vase missing red flowers'], 'Attributes': ['None'],

'Numbers': ['one, number of vase']},

'flowers': {'Relations': ['None'], 'Attributes': ['red, flowers are red'], 'Numbers': ['zero, missing red flowers']}]}

Caption: Two little bears climbing on a weathered wood chair.

Scene graph: {'Entities': ['bears', 'chair'],

'bears': {'Relations': ['climbing on, bears climbing on wood chair'], 'Attributes': ['little, bears are little'],

'Numbers': ['two, number of bears']},

'chair': {'Relations': ['None'], 'Attributes': ['weathered, chair is weathered', 'wood, chair is wood'],

'Numbers': ['one, number of wood chair']}]}

Fig. 4: Prompts and context examples for aspect graph parsing with GPT-4.

Given the caption {}, your goal is to change the aspect (Entity, Relation, Attribute, Number): '{}' to the opposite meanings. The opposite meaning may include: other related but not semantic equivalent descriptions, and counterfactual statements against aspect {}.

Swapping the subject and object of a verb.
generating a new phrase that's not in the sentence and/or add the new object next to the selected object to make a new sentence.

Requirements:

The rewritten part should keep the same pos tag with '{}', and the other parts of the sentence should keep unchanged.
The rewritten sentence may be contradicted in the aspect phrase {} with the reference sentence '{}'.
The new sentence must make logical sense.
Here are some aspect changing examples for reference {}.
Please return multiple examples without any extra explanation.

Fig. 5: Prompts for node replacement with GPT-4. {} in the prompt will be replaced with the initial caption, the node to be replaced $\times 4$, the references, and the context examples.

GPT-4V

Please find all the aspect phrase in the caption which is mismatch with the image and return the results in the triplets format. The first element in the triplet should be one of the Entities, Attributes, Relations, or Numbers. The second element should be the mismatched phrases in the caption, and the third element should be the corrections of the mismatched phrases. The captions may contains 0-3 mismatched aspects. When there is no mismatch, the results should be None. Here are some examples of the possible aspects triplets: {'Numbers', 'two number of toilets', 'one number of toilet'}; {'Entities', 'Several motorcycles', 'non specific quantity of car parked'}. The caption is: {}. Please only return the triplets without any other information.

Gemini Pro Vision

Given a caption and image pair, your goal is to find out all the mismatched phrases in the caption and correct them. Please alsoe give the category of the mismatched phrase. Please return the results with the triplets format.

The first element in the result triplet should be one of the Entities, Attributes, Relations, or Numbers, which indicated the category of the mismatched phrase.

The second element should be the mismatched phrases in the caption, and the third element should be the corrections of the mismatched phrases.

The captions may contains 0-3 mismatched aspects. When there is no mismatch, the results should be None.

Here are some examples

Example1: The caption is: a diagram depicting stillness from left to right. Mismatched Aspects: {'Entities, stillness, movement'}, {'Relations, depicting -- (diagram depicting stillness), showing -- (diagram showing movement)};

Example2: The caption is: The dog holds in its mouth what someone would normally wear as a hat. Mismatched Aspects: {'Relations, held in -- (held by dog), worn by -- (worn by dog)}, {'Entities, mouth, dog};

Example3: The caption is: they drank water then they worked out. Mismatched Aspects: None; Please only return the triplets without any other information.

The Caption is:

Fig. 6: Prompts for the in-context learning experiments with GPT-4V and Gemini Pro Vision.

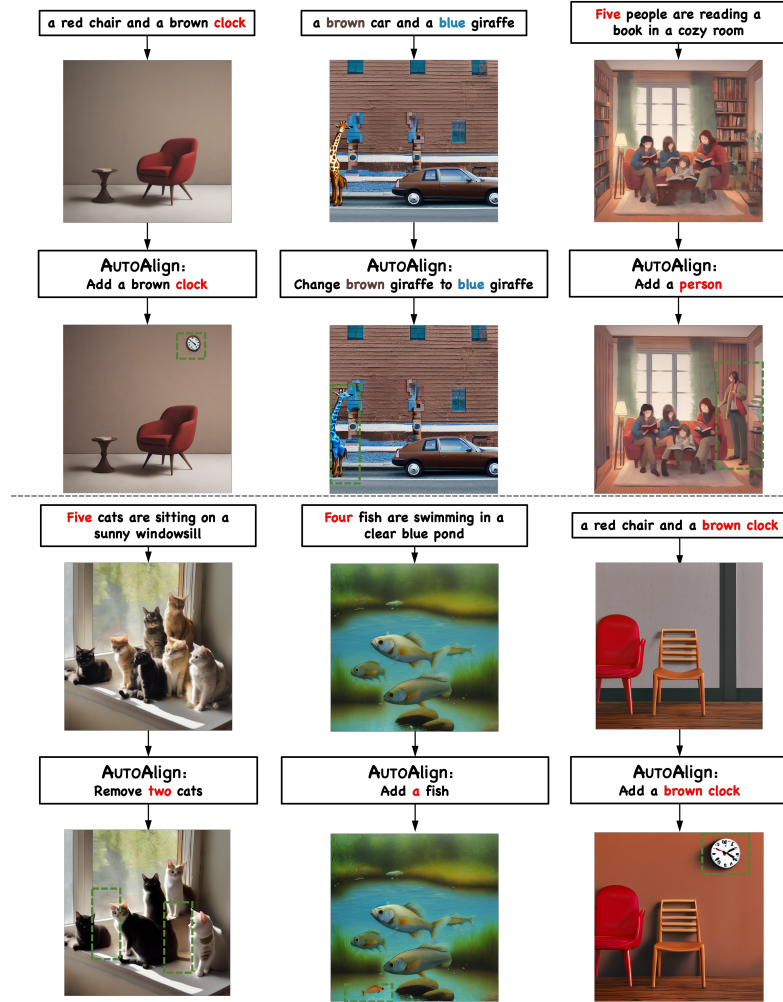


Fig. 7: Examples demonstrating how AUTOALIGN helps reduce the hallucination for T2I generation.

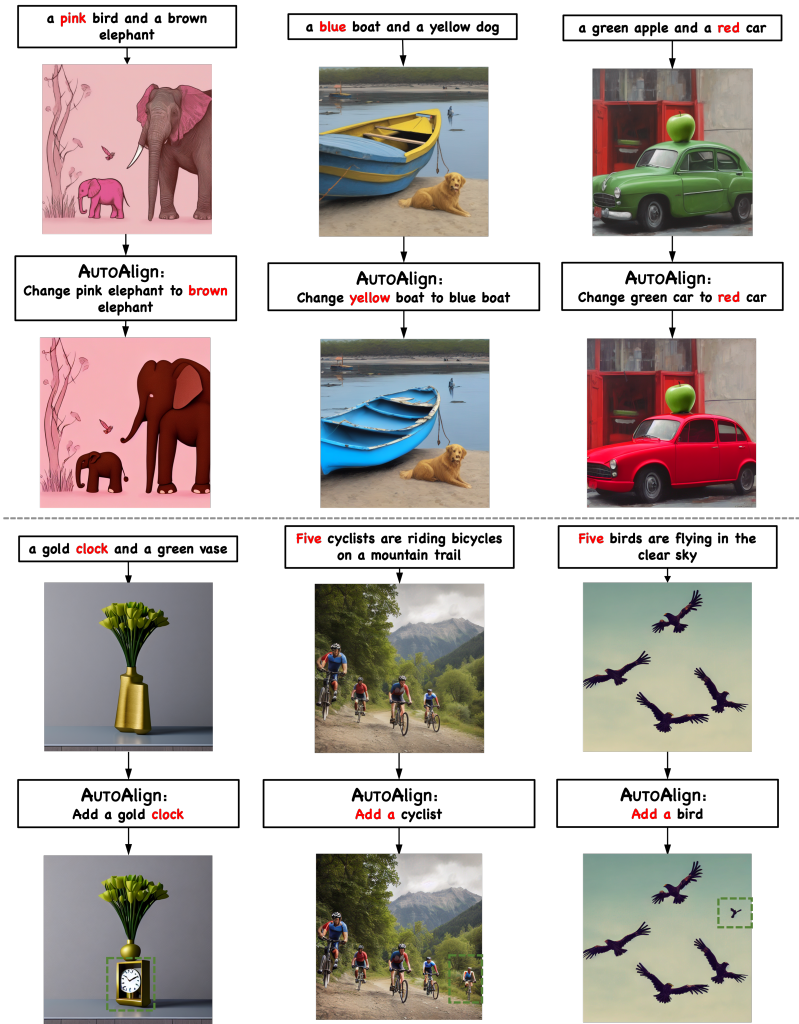


Fig. 8: Examples demonstrating how AUTOALIGN helps reduce the hallucination for T2I generation.

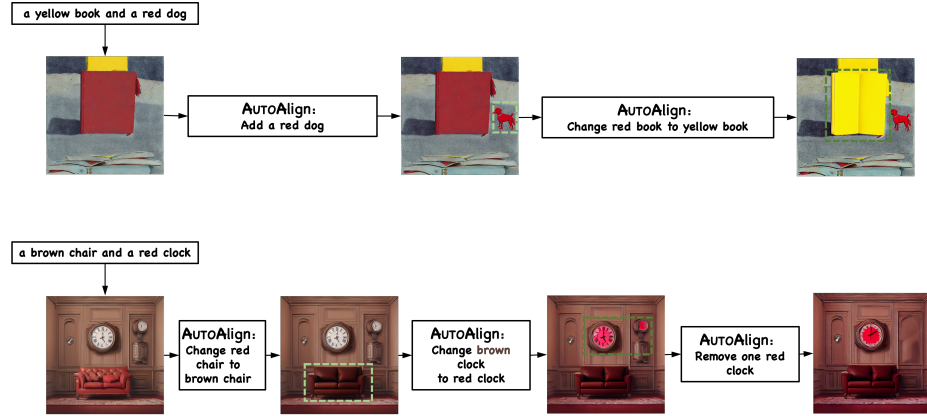


Fig. 9: Examples for multi-turn mismatch detection and correction with FINEMATCH.

Let's think step by step, your goal is to generate an instruction to edit the Target Image to be aligned with the Caption according to the Contradiction prompt.
 The Contradiction prompt reflects the mismatch of the Target Image and the Caption. The only editable part should be the Editing Aspect. Please first recognize the editing aspect using the schema: 'things to be changed -> change target', the changing direction is Caption -> Target Image. Please give your one-sentence instruction:

Caption: a pug dog with some cookies on its head.
 Target Image: a pug dog eat a cookie.
 Contradiction: the relation 'eat' does not exists in the source image.
 Editing Aspect: -> dog eat cookie
 Instruction: change the dog to eat the cookie.

Caption: a child and a woman sitting at a table eating and talking on the phone.
 Target Image: a man and a woman sitting at a table eating and talking on the phone.
 Contradiction: man does not exist in the source image but is expected by the Target Image.
 Editing Aspect: child -> man
 Instruction: Change the child to a man

Caption: A city view of LA.
 Target Image: A night view of LA.
 Contradiction: the attribute 'night, view of city' does not exists in the source image but is expected by the Target Image.
 Editing Aspect: view -> night view
 Instruction: Change the background to a night view

Caption: A group of children play in a park.
 Target Image: two child and a dog playing in the park.
 Contradiction: the number 'two, (number of child)' does not exists in the source image but is expected by the Target Image.
 Editing Aspect: a group of child -> two child
 Instruction: change the number of the child and remains only two children


Caption: A woman holding a hammer and a box on the floor.
 Target Image: A woman holding a hammer and a box on the wall.
 Contradiction: the relation 'on, (box on the wall)' does not exists in the source image but is expected by the Target Image.
 Editing Aspect: -> box on the wall
 Instruction: Change the position of the box to be on the wall.

Caption: A woman sitting on a sofa is reading a book.
 Target Image: A woman wearing glasses is reading a book.
 Contradiction: 'glasses' does not exist in the source image but is expected by the Target Image.
 Editing Aspect: -> glasses
 Instruction: Add glasses to the woman.

Caption: a man and a woman sitting at a table eating and talking on the phone
 Target Image: a man and a dog sitting at a table eating and talking each other
 Contradiction: dog does not exist in the source image but is expected by the Target Image.
 Editing Aspect: woman -> dog
 Instruction: Change the woman to a dog.

Caption: a table with a plate of red apple
 Target Image: a green apple on the table
 Contradiction: the attribute 'green' does not exist in the source image but is expected by the Target Image.
 Editing Aspect: red apple -> green apple
 Instruction: Change the color of the apple from red to green.

Fig. 10: Prompts and context examples for image editing instruction generation with GPT-4.



Caption:

Black and white photo of three men walking with bikes.

Aspect	Mismatched Phrases	Corrected Phrases
Entities <input type="text"/>	<input type="text"/>	<input type="text"/> <input type="button" value="Delete"/>

Fig. 11: The human annotation interface for FINEMATCH.

References

1. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **123**, 32–73 (2017) [1](#)
2. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV* **123**(1), 74–93 (2017) [1](#)