# SPECI : Fine-grained Evaluation of Specificity for Image Captioning

**Son T. Luu**[1,2,*], **Hiep Nguyen**[3], **Trung Bui**[4], **Viet Lai**[4], **Le-Minh Nguyen**[3,†]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[4]Adobe Research, San Jose, California, USA
sonlt@uit.edu.vn, hiepnkv@jaist.ac.jp, {bui,daclai}@adobe.com, nguyenml@jaist.ac.jp

## Abstract

Image Captioning lies at the center of synthetic data generation in computer vision. The quality of the caption defines the behavior of the resulting model trained on image-caption pairs generated by Image Captioning. Measuring the specificity of image captioning plays an important role in supplying high-quality captions for the downstream tasks. To address the challenge of automatically assessing the quality of the caption, we propose **SPECI**, an automatic metric that decomposes dense captions into atomic facts and evaluates specificity across visual criteria such as granularity, spatial detail, color, and linguistic focus using WordNet-based hyponym levels. Empirical results on the benchmark datasets demonstrate significant correlation with human judgments, and its application in image captioning and image reconstruction tasks highlights the practical utility of SPECI. The source code of SPECI is available at https://github.com/sonlam1102/SPECI.

## 1 Introduction

Image Captioning is the powerhouse behind many multi-modality models where the caption was used in downstream tasks (Nguyen et al., 2023) or being directly used to train large multi-modality language models (Deitke et al., 2025). As such, measuring the quality of image captioning plays an important role in supplying high-quality captions for the downstream tasks (Cui et al., 2025). Improving grounding and coverage (Yan et al., 2024; Lee et al., 2024) through reducing hallucination (Leng et al., 2024) has led to significant development of question answering (Wang et al., 2023; Yang et al., 2025a) and image generation (Yang et al., 2025b). However, these tasks suffer from various errors such as lack of detail (Ye et al., 2025) and ambiguity (Jian et al., 2025). This can be solved by improving the quality of caption through a different lens, namely **specificity** (Jas and Parikh, 2015).




Figure 1: Original image (upper) and three generated images (lower) generated from three captions with different level of detail reveals the needs for high level of specificity in image caption.

Specificity measures the level of detail engaged in the text (Gao et al., 2019). Ideally, it leads to more precise and concise captions through more details and better lexical choices. For example, given three non-hallucinated short captions for a picture below (Figure 1):
(i) *"a car in front of a house"*
(ii) *"a white car in front of a house"*
(iii) *"a white sedan in front of a house"*
The latter captions are more specific than the former captions because they provide more information (e.g., "*white*") and are more specific/detailed (e.g., "*sedan*" vs "*car*"). Thus, it leads to more faithful generated image.

The specificity of captions is essential to assess the performance of computer vision models and evaluate the efficiency of the downstream tasks. However, a caption usually contains a wealth of information, including objects, attributes, i.e., color, texture, lighting, and their relations, i.e., spatial

[*]This work was partially conducted at JAIST.
[†]Corresponding author



*Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 140–156
November 6-10, 2026 ©2026 Association for Computational Linguistics



SPECI Score: **0.0222**

A brown dog, likely a working breed, is captured mid-leap over a textured agility ramp, showcasing its muscular build and focused expression. The dog's mouth is open in a wide bark, revealing its teeth, which adds a sense of excitement to the scene. The ramp, covered in a sandy material, is positioned in a grassy area, suggesting an outdoor agility course. In the background, blurred figures of spectators can be seen, some wearing casual clothing, indicating a lively event atmosphere. The bright sunlight enhances the vivid colors of the scene, highlighting the dog's shiny coat and the green grass beneath.

SPECI Score: **0.0092**

A dog is jumping over a wooden ramp, with its mouth open and showing its teeth. The dog is competing in a dog show, and the ramp is part of the competition. The dog is likely trying to impress the judges and win the competition. There are several people in the background, watching the dog's performance.

Figure 2: An example of dense caption generated by two different models.

relations, and language style usage, indicating a challenge in assessing the specificity of a caption. For example, Figure 2 illustrates a picture with two captions. It can be seen that the caption in the blue box is more specific than the caption in the green box because the content in the blue-box caption describes better visual detail with descriptive language, such as *"captured mid-leap"*, *"textured agility ramp"*, *"blurred figures"*, and *"green grass beneath"*, while the caption in the green box use general words such as *"ramp"* or *"several people"*. Estimating the specificity score of the caption could help choose an appropriate description for the image with more enriched information.

In this paper, we introduce **SPECI**, a new fine-grained point-wise evaluation of an image caption 's specificity. SPECI quantifies the level of detail engaged in the text that describes the image, such as word phrases and entities, to estimate the specificity of the caption. Computing SPECI score involves (1) decomposing the long caption into a series of atomic visual cues, of which each describes a single object or objects' relation, (2) computing a real atomic specificity score for each visual cue, allowing assigning a single real overall specific score for the whole caption. To efficiently evaluate the specificity, we employ the hyponym level for assessment. Since a word has more hyponyms, it will become more general; conversely, it will become more specific. Leveraging this idea, we assess the specificity of each word in an atomic fact based on the hyponym level definition from Wordnet (Miller, 1994) to measure the specificity. To show the efficiency of SPECI, we evaluate the alignment between the SPECI score and the human score on three human-annotated datasets. The empirical results show that SPECI has significantly improved the correlation with the human preference for the specificity.

In summary, our contributions are as follows:

(1) We introduce SPECI, a new point-wise reference-free evaluation of specificity for image captioning using the decompose-and-compute paradigm. The model uses the hyponym level to compute the atomic-level score.

(2) We confirm the effectiveness of the method in improving alignment with human preferences on three open image captioning datasets.

(3) We demonstrate the effectiveness of SPECI in downstream tasks.

## 2 Related Works

To evaluate the quality of the image captioning, traditional metrics compare the predicted caption against a predefined set of golden captions. These method includes string matching (BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005)), and semantic matching (BERTScore (Zhang et al., 2019)). However, these metrics only focus on n-gram overlap or embedding similarity, hence, they often fail to capture fine-grained aspects like visual details, aesthetics, nuance, and comprehensiveness of the caption.

In the era of LLMs, using LLM-as-a-judge provided a new potential approach for automatically evaluating the generated output by assigning scores, producing rankings, or selecting the best options (Li et al., 2025). CapMas (Lee et al., 2024) proposes a multi-agent system where an LLM decomposes detailed captions into atomic propositions, an MLLM (multimodal LLM) verifies their truthfulness against the image, and another LLM revises the caption for factual accuracy. FAITHScore (Jing et al., 2024) employs reference-free evaluation by decomposing the caption into atomic claims, then verifying consistency with the input image using an LLM judge. CapArena (Cheng et al., 2025) introduces a new benchmark with human-voted pairwise comparisons of detailed captions from VLMs, using LLMs as judges to determine which one is better. Overall, these approaches offer a scalable, efficient method alternative to manual annotation and enable reference-free assessment in practical applications. However, they may inherit biases from the underlying LLMs and require careful prompt design to obtain consistency and fairness.

Prior studies have evaluated the specificity of captions in a reference-based manner. In CLAIR (Chan et al., 2023), the model compares the quality of the image caption between the candidate and reference captions. DCScore (Ye et al., 2025) decomposes the caption into primitive information units and verifies them based on the reference data. Both methods highlight the utility of fact decomposition for fine-grained evaluation but depend on human-annotated references for matching, limiting scalability in fully automated scenarios. Additionally, FineSureE (Song et al., 2024) shows a similar approach to CapMas for measuring the factuality and conciseness of the generated content, FINECAPTION (Hua et al., 2025), and GRANOLA-QA use multi-granularity for downstream tasks. Nevertheless, these current methods focus on estimating the consistency and completeness of details extracted from the image and text, without addressing the details of objects represented in the caption or proposing a proper way to measure specificity.

## 3 Method

SPECI is designed to measure the specificity by estimating the specific level of the word represented in the sentence. As illustrated in Figure 3, SPECI breaks down the caption into general atomic facts. In parallel, SPECI decomposes the caption text into four visual facts, including Visual Detail, Spatial Specificity, Color & Texture, and Interpretive Language. Then it computes the specificity of the text using WordNet. Finally, SPECI combined the two scores to give the final results that estimate the specificity of the caption.

### 3.1 WordNet-based Specific Score

Given a sentence $s$ that consists of $K$ tokens, the specific score of the sentence $s$ is computed as:

$$\text{wns}(s) = \frac{1}{\sum_{i=1}^{|K|} |synset_level(k_i)|} + \frac{|K|}{max(|synset_level(K)|)} \quad (1)$$

where the $s$ is segmented to $K$ tokens by Stanza (Qi et al., 2020). The $synset_level(.)$ is extracted through WordNet by NLTK (Bird and Loper, 2004) for indicating the level of hyponym synset for a word. Given 2 words A and B, A is more specific than B if A has fewer hyponyms than B, which means that the cardinality of $synset_level(A)$ is less than $synset_level(B)$. Additionally, if the sentence contains many general tokens, which means the tokens have a large level of hyponyms, the overall specific score will be low.

**Algorithm 1** Finding the synset level of a word

**Require:** $WN$ ▷ WordNet
  **function** SYNS_LEVEL($w$)
    **if** $w$ is named_entity **then**
      **return** {}
    **end if**
    $synset \leftarrow \{\}$
    **for** $wd \in WN.synonym(w)$ **do**
      $synset \leftarrow synset \cup$ SYNS_LEVEL(wd)
    **end for**
    **return** $synset$
  **end function**

### 3.2 Atomic Specific Score

Since the dense caption is a paragraph that has many sentences, we decompose the paragraph into small pieces of information called atomic fact sentences that capture valuable information (Min et al., 2023). Given a caption $S$ for an image, $S$ is then decomposed into a list of atomic facts $A = [a_1, ..., a_n]$. The atomic specific score for the caption $c$ is the sum of the specific score on each

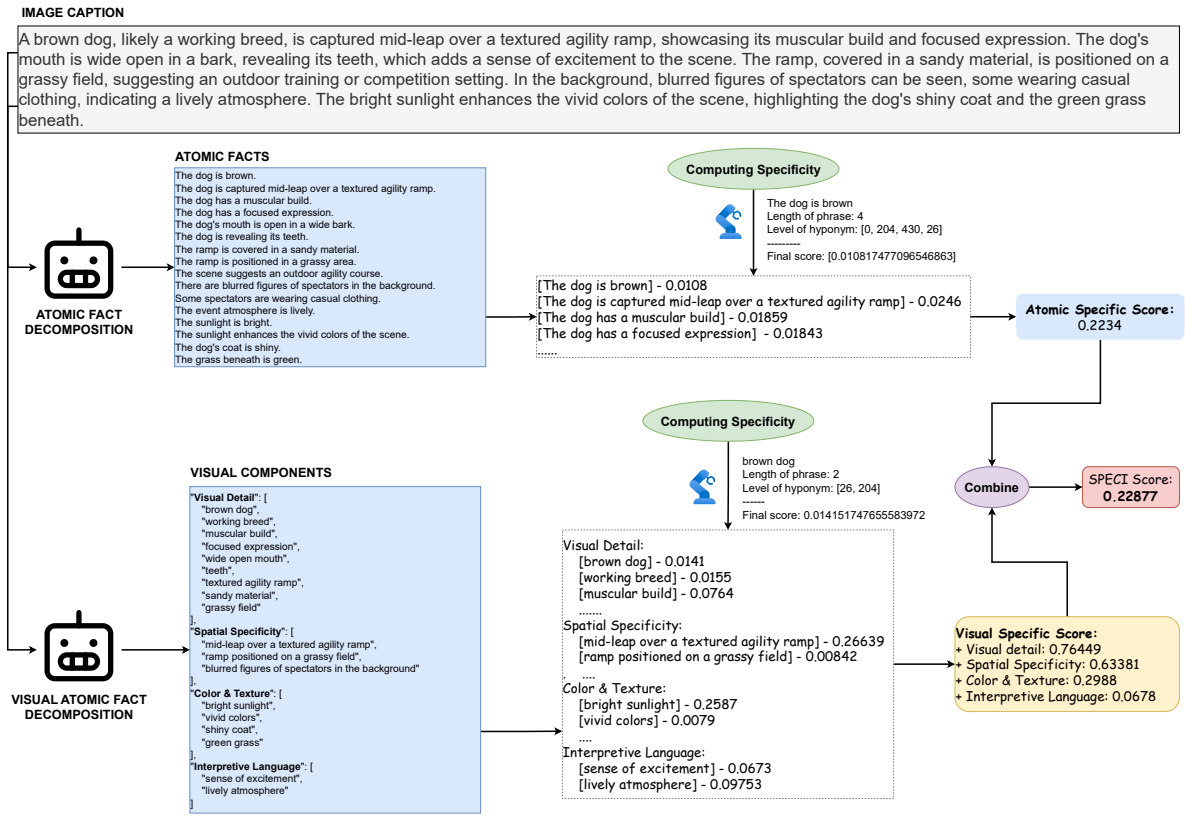



Figure 3: An overview of the SPECI pipeline for estimating the specificity of the caption.

atomic fact $a \in A$:

$$\text{atomic_speci}(c) = \frac{1}{|A|} \sum_{a_i \in A} \text{wns}(a_i) \quad (2)$$

To decompose the caption into atomic facts, we use the prompt below with the *GPT-4o-mini* API with the prompt as described in Appendix B.

### 3.3 Visual Specific Score

Since the decomposition of the atomic fact can capture valuable information in the caption, classifying them according to visual components gives us a better look into the details could enhance the interpretability in detecting the specificity of the caption. Similar to the idea in the Atomic Specificity score, we decompose the caption into four categories, including:

**Visual Detail:** specific object features, such as size, shape, quantity, material, pattern, or structural composition.

**Space and Composition:** position in the frame, order of elements, perspective, or how elements are arranged in space.

**Color & Texture:** color names, combinations, shades, or texture-related descriptors.

**Mood and Impression:** interpretive, subjective, or stylistic phrases that go beyond the surface visual information.

Given a caption $S$ that has been decomposed into visual categories. Let $C$ be the set of decomposed atomic phrases for the category $c$ among the above categories. The visual specific score for each category is computed as follows:

$$\text{visual_speci}_c(a_i) = \frac{1}{|C|} \sum_{a_i \in C} \text{wns}(a_i) \quad (3)$$

To decompose the caption into atomic facts according to each visual criterion, we employ the *GPT-4o-mini* API with the prompt as described in Appendix C.

### 3.4 SPECI Score

The SPECI score for estimating the specificity of a caption $S$ is a weighted average of the general atomic specific score (described in Equation 2) and visual specific scores (described in Equation 3):

$$\begin{aligned} \text{Speci}(S) = \alpha * \text{atomic_speci}(S) \\ + \sum \beta_c * \text{visual_speci}_c(S) \end{aligned} \quad (4)$$

| | DOCCI | DCI | IIW |
|---|---|---|---|
| # sample | 100 | 108 | 100 |
| max length (Caption A) | 576 | 640 | 399 |
| average length (Caption A) | 225 | 232 | 212 |
| min length (Caption A) | 59 | 85 | 66 |
| max length (Caption B) | 550 | 708 | 211 |
| average length (Caption B) | 134 | 159 | 122 |
| min length (Caption B) | 61 | 10 | 42 |

Table 1: Overall statistics about the IIW dataset. The length is computed according to the token level.

where $\alpha$ and $\beta$ are optimized on the development set. The details of the hyper-parameter values are shown in Appendix A.

# 4 Experiments

## 4.1 Dataset preparation

To compare the alignment between SPECI and human, we utilize the IIW dataset (Garg et al., 2024), which includes three subsets: DCI (Urbanek et al., 2024), DOCCI (Onoe et al., 2025), and IIW (Garg et al., 2024). In the IIW, there are a pair of captions with relevant human annotations about the specificity. Given two captions denoted as *A* and *B*, the annotation values are **2** if *the A is* ***substantially*** *more specific than the B*, **1** if *the A is* ***marginally*** *more specific than the B*, **0** if *the A and the B is similar in specific*, **-1** if *the B is* ***marginally*** *more specific than the A*, and **-2** if *the B is* ***substantially*** *more specific than the A*. Figure 5 shows an example of two captions for an image, and a human determines which one is better. In addition, Table 1 shows the overall analysis of the three datasets used for assessment in our experiments.

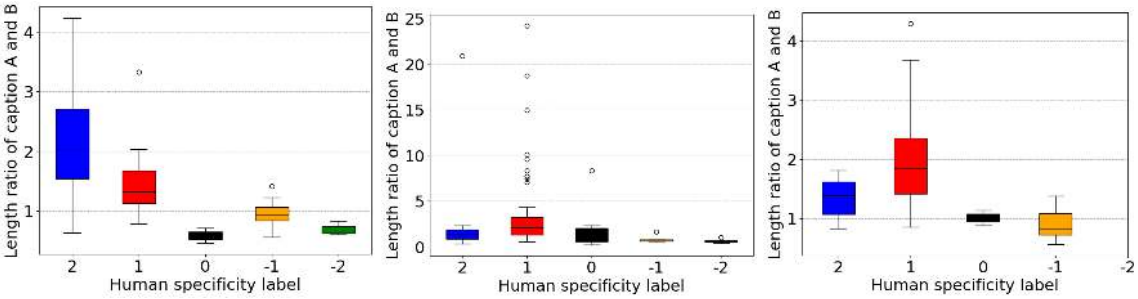
As shown in Table 1, the length between the source and target captions is not much different. The caption in the DCI is slightly longer than the two remaining, while **IIW** has shorter captions than DCI and DOCCI. To investigate the impact of the caption length on the decision of humans about the specificity, we compute the ratio of the length between caption A and B, and compare it with the human-specific score, as shown in Figure 4. The ratio will be high if the length of caption A is longer than that of B, and it will be lower if B is longer than A. It can be seen that in the DOCCI, the caption A tends to be more specific if the length of A is longer than that of B. Nevertheless, the level of specificity seems not to depend on "Caption A is longer than B" in DCI and IIW, where there are many cases that caption A shows a longer length than B are determined as "marginal specificity" (as shown in Figure 4b and 4b). Overall, from the analysis of the length between 2 captions *A* and *B* and the specificity label determined by humans, it can be seen that the specificity of the caption mostly depends on the length of text, but not all the times. The estimation of the specificity of the caption involves exploiting the detailed description of the caption for an image.

## 4.2 Experiment preparation

This section aims to evaluate the performance of SPECI with other available methods by comparing the correlation with humans in estimating the specific. We employ three human preference datasets, including DCI (Urbanek et al., 2024), DOCCI (Onoe et al., 2025), and IIW (Garg et al., 2024), to assess the performance. As defined in Section 4.1, each sample in the three datasets has a pair of captions, $(A, B)$, with a corresponding human-labeled score indicating which is more specific. In this experiment, each method assesses one caption with a score, then computes the difference between the two captions and calculates the correlation with the human score via correlation metrics, including *Spearman*, *Kendall*, and *Pearson*. For example, we employ SPECI for estimating the specificity of each caption A and B, then compute the difference in specific score $\text{SPECI}(A) - \text{SPECI}(B)$ to represent the specificity comparison between caption A and caption B. Similarly, we let the available LLMs such as GPT-4o, Gemini, Qwen3, and Llama3.1 to give a score assessing the specific of each caption A and B, then compute the abstraction between A and B to calculate the correlation with human-labeled (the prompt for scoring the specific of caption is described in Appendix E). Finally, we employ two automatic metrics for image and video evaluation – G-VEAL (Tong et al., 2025) and Polos (Wada et al., 2024), to compare the correlation with humans.

## 4.3 Results

**SPECI shows the highest correlation with human preference in estimating the specificity** Table 2 shows that current LLMs like GPT4-o and Gemini attain a moderate correlation with humans in estimating the specificity of the caption. For example, GPT-4o obtained 33.97 (Spearman), 30.89 (Kendall), and 30.89 (Pearson), which is better than Gemini's 23.48, 21.36, and 16.40, respectively. In contrast, the performance of GPT-4o and Gemini on DOCCI and IIW is worse than DCI, where GPT-




(a) DOCCI (b) DCI (c) IIW

Figure 4: The relation between the length of the caption and the specificity.

| Models | DCI | | | DOCCI | | | IIW | | |
|---|---|---|---|---|---|---|---|---|---|
| | Spearman | Kendall | Pearson | Spearman | Kendall | Pearson | Spearman | Kendall | Pearson |
| GPT-4o | 33.97 | 30.89 | 25.18 | 10.09 | 9.64 | 11.28 | -5.60 | -5.60 | 8.80 |
| Gemini | 23.48 | 21.36 | 16.40 | 3.20 | 3.06 | -2.85 | 14.09 | 13.57 | 12.68 |
| Qwen3-4B-Instruct | 12.96 | 12.01 | 10.10 | 12.61 | 12.05 | 10.46 | -1.12 | -1.09 | 1.55 |
| LLama3.1-8B-Instruct | 24.22 | 22.91 | 21.65 | 7.32 | 6.96 | 5.97 | 10.16 | 9.86 | 6.17 |
| G-VEAL | -7.52 | -5.76 | 1.12 | 12.65 | 9.65 | 8.63 | 2.96 | 2.40 | -12.20 |
| Polos | -6.10 | -4.78 | 2.49 | -3.77 | -2.70 | -2.98 | -11.51 | -9.09 | -12.71 |
| Atomic SPECI | 46.75 | 37.71 | 37.75 | **51.40** | **41.11** | **47.37** | 29.52 | 23.49 | 26.92 |
| Visual SPECI | 40.89 | 32.10 | 30.86 | 13.35 | 10.55 | 12.94 | 28.01 | 22.28 | 20.57 |
| **SPECI (our)** | **50.14** | **40.21** | **38.22** | 42.02 | 33.32 | 44.19 | **45.25** | **36.41** | **39.87** |

Table 2: Comparison of SPECI with human-annotated specificity



**Caption A (Prefered)**

*"A horizontally-oriented eye level image shows a residential concrete-like balcony destroyed by an explosion in Beirut. The image shows the balcony slightly slanted toward the background on the left side. Two parallel rectangular concrete-like sections have wrecked, twisted, and separated, missing metal safety grates. Torn and shredded white and red striped fabric hangs limply from partially separated metal grates still learning on the balcony. Behind the metal grates, window panes are missing and twisted strips of housing material are partially visible. Stacks of brown cardboard boxes are visible behind the leaning grates on the balcony on the right side. "*

**Caption B**

*"This is the side of a building. It shows mostly one floor and a little of the floor above it. It has a balcony. It had a metal railing but it is all broken and has a dirty, ripped and striped cloth draped in front of it. There are boxes piled haphazardly behind it. Part of the railing is leaning out. The higher floor seems better. The building is tan and dirty. "*

Figure 5: A sample from the dataset, in which a human evaluates two captions, A and B, to determine which one is better. More examples can be found in Appendix G.

4o achieved only 10.09 (Spearman), 9.64 (Kendall), and 11.28 (Pearson) on DOCCI, and Gemini obtained about 14.09 (Spearman), 13.57 (Kendall), and 12.68 (Pearson) on IIW. With open-source LLMs like Qwen3 and LLama3.1, the correlation results are worse than GPT4-o and Gemini on three datasets, indicating the challenges of LLMs in estimating the specificity of captions from scratch.

**Atomic SPECI shows a higher correlation with human preference than Visual SPECI** Atomic SPECI consistently outperforms Visual SPECI. For example, on the DCI dataset, Atomic SPECI achieves 46.75 (Spearman), 37.71 (Kendall), and 37.75 (Pearson), compared to Visual SPECI's 40.89, 32.10, and 30.86, respectively. Across nearly all metrics, SPECI which is an ensemble method of Atomic SPECI and Visual SPECI outperforms the other models, achieving the highest correlations, such as a Spearman score of 50.14, Kendall score of 40.21 and Pearson of 38.22 on DCI, and strong performance on IIW with 45.25 (Spearman), 36.41 (Kendall) and 39.87 (Pearson). This indicates that both Atomic SPECI and Visual SPECI contribute to the improvement on the DCI and IIW datasets. Overall, our method achieves the highest alignment with human preferences.

**Visual SPECI have impact in estimating the specificity** According to the results in Table 3,

the four visual criteria have a significant contribution to the Visual SPECI, where each component shows its efficiency in estimating the visual features represented in caption text. Additionally, we randomly take 149 examples from the human preferences dataset and conduct the ANOVA test among visual SPECI components, including Visual Detail (visual), Space and Composition (spatial), Color & Texture (color), and Mood and Impression (linguistic). Since the $p - value$ is lower than 0.05, we should reject the null hypothesis that all visual components are equal, which indicates the meaning of association among visual components in Visual SPECI . The details about the statistical testing results on the contribution of each visual component to the Visual SPECI are shown in Appendix F

| | Spearman | Kendall | Pearson |
|---|---|---|---|
| Visual Detail | 25.61 | 20.26 | 33.36 |
| Space & Composition | 11.06 | 8.77 | 18.55 |
| Color & Texture | 7.59 | 6.12 | 1.34 |
| Mood & Impression | 19.56 | 15.27 | 19.61 |
| **Full Visual SPECI** | **26.93** | **21.29** | 31.37 |

Table 3: Ablation study of Visual SPECI components.

**Other metrics are not designed for specificity** According to Table 2, G-VEAL (Tong et al., 2025), and Polos (Wada et al., 2024) demonstrate poor correlation with humans on three datasets in the specific estimation of the caption. This is attributed to that these metrics measures orthogonal aspects. In particular, G-VEAL measures the quality of caption through a chain-of-thought reasoning to comprehend the consistency and completeness of caption with the image (Tong et al., 2025), while Polos evaluates the quality of image captions by integrating human feed-back. These do not focus on the level of detail in the caption represented by words in the caption sentences like SPECI.

**SPECI is effective in image reconstruction task** To compare SPECI with other methods, we conduct an experiment in which we use SPECI for estimating the specificity of the generated caption for reconstructing the image. We use 100 image samples from Flickr30K-CFQ (Liu et al., 2024) and employ two image-caption models: GPT4o (OpenAI et al., 2024) and InstructBLIP-13B (Dai et al., 2023) to generate the captions. We set the temperature equal to 0 for GPT4o and $do_sample = False$ for InstructBLIP to reserve the deterministic of the generated content by the image captioning model.

To compare the performance of SPECI with other methods, we measure the prediction matching of SPECI versus other, which measures the proportion of agreed samples over the total. Given 2 captions, A and B, if SPECI estimates that "A is more specific than B" and another method also gives the same answer, it will be the "agreed". The Accuracy as shown in Table 4 indicate the proportion of "agreed" samples between SPECI with other metrics. Specifically, our comparison to the other methods includes:

**Caption-based comparison** : We use CapArena (Cheng et al., 2025) – a metric that leverages the LLMs in estimating the details between two captions.

**Image-based similarity comparison** : We use the StableDiffusion-3.5-medium[1] for generating new images $I_{GPT4}$ and $I_{instruct-blip}$ corresponding to $Cap_{GPT4}$ and $Cap_{instruct-blip}$, then compare the SPECI with the similarity of $I_{GPT4}$ and $I_{instruct-blip}$ to the original image via two image-embedding models: CLIP-large (Radford et al., 2021) and BGE-M3 Visualized (Zhou et al., 2024).

| Model | Acc |
|---|---|
| CapArena (Cheng et al., 2025) | 0.98 |
| CLIP-large (Radford et al., 2021) | 0.63 |
| BGE-M3 Visualized (Zhou et al., 2024) | 0.66 |

Table 4: Comparison between SPECI with other methods in estimating the specificity of image caption.

As shown in the Table 4, it can be shown that SPECI has a high matching with CapArena (Cheng et al., 2025) - a pair-wise battle framework in comparison of two machine-generated detailed image captions. The SPECI, which estimates the specificity of the caption directly, shows optimistic agreement with both pair-wise and point-wise image-detail evaluation methods. In addition, SPECI shows a moderate percent agreement in the text-to-image task. As shown in Table 4, the accuracy between the SPECI score with the similarity relationship of generated and original images computed through an image embedding model like CLIP (Radford et al., 2021) or BGE-M3 Visualized (Zhou et al., 2024) is more than 60%. Intuitively, from the results, it can be interpreted that a more

[1] https://huggingface.co/stabilityai/stable-diffusion-3.5-medium

| Models | DCI | | | DOCCI | | | IIW | | |
|---|---|---|---|---|---|---|---|---|---|
| | Spearman | Kendall | Pearson | Spearman | Kendall | Pearson | Spearman | Kendall | Pearson |
| G-VEAL | -7.52 | -5.76 | 1.12 | 12.65 | 9.65 | 8.63 | 2.96 | 2.40 | -12.20 |
| +SPECI | **19.24** ↑26.76 | **15.08** ↑20.84 | **14.24** ↑13.12 | **35.82** ↑23.17 | **28.03** ↑18.38 | **31.13** ↑22.50 | **44.53** ↑41.57 | **35.80** ↑33.40 | **32.04** ↑19.84 |
| POLOS | -6.10 | -4.78 | 2.49 | -3.77 | -2.70 | -2.98 | -11.51 | -9.09 | -12.71 |
| +SPECI | **19.24** ↑25.34 | **15.08** ↑19.86 | **14.24** ↑11.75 | **35.82** ↑39.59 | **28.03** ↑30.73 | **31.13** ↑34.11 | **44.53** ↑56.04 | **35.80** ↑44.89 | **32.04** ↑44.75 |

Table 5: The results of the combination between SPECI and other metrics for assessing the specificity of the image.

specific caption can be better used for reconstructing a new image that is mostly similar to the original image. Hence, the SPECI helps estimate the most specific caption in which it carries the most information from the original image (examples are revealed in Appendix D).

## 5 Discussion

### 5.1 SPECI improves other metrics in Specificity Evaluation

As Table 2 shows G-VEAL (Tong et al., 2025) and Polos (Wada et al., 2024) 's poor correlation with human preference, we investigate if SPECI can help improves their alignment. In particular, we combine the SPECI with the two metrics to investigate the impact of SPECI on others when estimating the specificity of the caption. The combination is computed by taking the average of the combined metrics. As demonstrated in Table 5, SPECI boost the correlation with humans of Polos and G-VEVAL significantly, even when the original results of both methods show an orthogonal correlation. It can be seen that SPECI demonstrates a significant impact in estimating the specificity of the caption. Moreover, SPECI performs computing the specificity of the caption directly by estimating the specificity of words in the caption without image reference, which is different from G-VEAL and Polos that require the involvement of the image in the input.

### 5.2 Caption length is a moderate indicator of specificity but not the best.

In this section, we investigate the impact of caption length on SPECI in estimating the specificity of the caption. As mentioned in Section 3, SPECI measures the specificity by estimating the specificity of the word used in the sentence. Hence, the longer caption may contain more specific words that help improve the detail of the caption. However, as demonstrated in Table 6, SPECI indicates a moderate to high correlation with the length of the caption (The length of the caption is calculated by token level using NLTK tokenizers), which shows that the length of the caption has a moderate impact on the estimation of the specificity scored by SPECI. Additionally, it can be seen that the specificity of human preference also demonstrates a moderate correlation with the length of the caption, while the hallucination determined by humans shows negative results, as illustrated in Table 6.

### 5.3 SPECI is not an indicator of hallucination

According to the results of hallucination by human preference described in Table 6, the hallucination of the caption does not seem to be completely impacted by the length of the caption, like the specificity. Since SPECI can capture the specificity well of the caption, it is difficult to consult with the hallucination that appeared in the caption.

| Length vs | Spearman | Kendall | Pearson |
|---|---|---|---|
| SPECI score | 62.55 | 45.84 | 61.86 |
| Specificity | 45.73 | 37.02 | 15.99 |
| Hallucination | -19.76 | -15.56 | -30.12 |

Table 6: The correlation between caption length versus SPECI score, human preferences of specificity, and human preference of hallucination on the DCI dataset.

## 6 Conclusion

This paper proposed **SPECI** - a method for automatically estimating the specificity of dense visual captions. By decomposing captions into atomic facts and categorizing them into visual components, SPECI leverages WordNet hyponym levels to quantify fine-grained specificity. Empirical results demonstrate that SPECI attains significant correlation with human-annotated specificity. Furthermore, SPECI shows strong agreement with other LLM-based evaluation methods and moderate alignment in image reconstruction tasks, showing its utility in assessing the specificity of captions for downstream applications. Our future work will extend SPECI to multilingual contexts by integrating with a multilingual lexicon database like Ba-

belNet (Navigli and Ponzetto, 2010) for extracting hyponym levels, as well as exploring SPECI application in real-time caption generation systems.

## Limitation

While SPECI provides an effective and scalable method for estimating the specificity of dense visual captions, it is subject to several limitations:

**Multilingual and Domain-specific** As WordNet is primarily an English lexical database, depending on WordNet for estimating hyponym-level specificity limits the utility in multilingual contexts or for captions involving non-English terms. Additionally, WordNet's coverage may be limited for specialized domains.

**Factuality and Hallucination** SPECI focuses exclusively on linguistic and structural specificity, but is not yet concerned about the factuality information represented in the caption. Therefore, SPECI should be used along with a faithfulness evaluation method like CapMAS (Lee et al., 2024) to better provide a comprehensive assessment for the image captioning task.

**Bias to LLMs for Decomposing** SPECI decomposes the captions into atomic facts and visual components, which depend on carefully crafted prompts. Errors in this decomposition process, like incomplete extraction, misclassification of phrases, or inconsistencies due to LLM calibration, could affect the final SPECI score.

## Acknowledgments

This work was partly supported by Japan Science and Technology Agency (JST) as part of Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE), Grant Number JPMJAP25B2 and JST CREST, Japan, Grant Number JPMJCR2554.

## References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

David M. Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, and John Canny. 2023. CLAIR: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, Singapore. Association for Computational Linguistics.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. 2025. CapArena: Benchmarking and analyzing detailed image captioning in the LLM era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14077–14094, Vienna, Austria. Association for Computational Linguistics.

Tianyu Cui, Jinbin Bai, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ye Shi. 2025. Evaluating image caption via cycle-consistent text-to-image generation. *arXiv preprint arXiv:2501.03567*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104.

Yifan Gao, Yang Zhong, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2019. Predicting and analyzing language specificity in social media posts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6415–6422.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. 2024. ImageInWords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 93–127, Miami, Florida, USA. Association for Computational Linguistics.

Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Soo Ye Kim, Zhifei Zhang, Yilin Wang, Jianming Zhang,

Zhe Lin, and Jiebo Luo. 2025. Finecaption: Compositional image captioning focusing on wherever you want at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24763–24773.

Mainak Jas and Devi Parikh. 2015. Image specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2727–2736.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638, Vienna, Austria. Association for Computational Linguistics.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, Miami, Florida, USA. Association for Computational Linguistics.

Saehyung Lee, Seunghyun Yoon, Trung Bui, Jing Shi, and Sungroh Yoon. 2024. Toward robust hyperdetailed image captioning: A multiagent approach and dual evaluation metrics for factuality and coverage. *arXiv preprint arXiv:2412.15484*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haoyu Liu, Yaoxian Song, Xuwu Wang, Xiangru Zhu, Zhixu Li, Wei Song, and Tiefeng Li. 2024. Flickr30k-cfq: A compact and fragmented query dataset for text-image retrieval. In *International Conference on Database Systems for Advanced Applications*, pages 419–434. Springer.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *Advances in neural information processing systems*, 36:22047–22069.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2025. Docci: Descriptions of connected and contrasting images. In *Computer Vision – ECCV 2024*, pages 291–309, Cham. Springer Nature Switzerland.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. 2025. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7419–7427.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26700–26709.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568.

Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. 2023. Filling the image information gap for VQA: Prompting large language models to proactively ask questions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2874–2890, Singapore. Association for Computational Linguistics.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, pages 37–53. Springer.

Shuo Yang, Caren Han, Siwen Luo, and Eduard Hovy. 2025a. MAGIC-VQA: Multimodal and grounded inference with commonsense knowledge for visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16967–16986, Vienna, Austria. Association for Computational Linguistics.

Zhantao Yang, Ruili Feng, Keyu Yan, Huangji Wang, Zhicai Wang, Shangwen Zhu, Han Zhang, Jie Xiao, Pingyu Wu, Kai Zhu, Jixuan Chen, Chen-Wei Xie, Yue Yang, Hongyang Zhang, Yu Liu, and Fan Cheng. 2025b. Bacon: Improving clarity of image captions via bag-of-concept graphs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14380–14389.

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. 2025. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. *arXiv preprint arXiv:2503.07906*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand. Association for Computational Linguistics.

## A Hyper-parameter configuration

**SPECI** $\alpha = 0.86044898$, $\beta = [0.06385769, -0.01287558, -0.05029409, 0.11765478]$, in which each value corresponds to *visual details*, *spatial specificity*, *color & texture*, and *interpretive language*, respectively.

**Open-source LLMs for specificity determination** $do_sample = False$, $max_new_tokens = 5$.

**Closed-source LLMs for specificity determination** $temp = 0.1$, $max_tokens = 10$.

**Atomic facts and visual details decomposition** $temp = 0.1$, $max_tokens = 2048$.

## B Atomic Fact Decomposition Prompt

The prompt for decomposing the paragraph into atomic facts is described as follows:

**Atomic Fact Decomposition**

Break down the following paragraph into a list of atomic facts. Each atomic fact must:
- Contain only one piece of information.
- Be clear, specific, and verifiable.
- Stand alone without relying on other facts.
- Avoid assumptions, opinions, or interpretations.

Return results as a list of strings such that each element is an atomic fact. Do not include any explanation, just return the list with format look likes ["atomic fact 1", "atomic fact 2", "atomic fact 3"].

Figure 6 shows an example of atomic facts extracted by the prompt.

## C Visual Atomic Fact Decomposition Prompt

The prompt for decomposing the paragraph into visual atomic facts is described as follows:

**Visual Atomic Decomposition**

You are given a paragraph that describes an image (caption).
Your task is to extract and categorize key **phrases** from the paragraph under the following four criteria:

1. Level of Visual Detail (Granularity): Extract phrases that describe specific object features, such as size, shape, quantity, material, pattern, or structural composition.

2. Spatial & Compositional Specificity: Extract phrases that describe spatial relationships, such as position in the frame, order of elements, perspective (e.g., top/bottom, left/right), or how elements are arranged in space.

3. Color & Texture Specificity: Extract phrases that describe precise color names, combinations, shades, or texture-related descriptors.

4. Linguistic Focus (Descriptive vs. Interpretive): Extract interpretive, subjective, or stylistic phrases (e.g., "visually appealing", "striking contrast", "reminiscent of", or metaphorical language) that go beyond objective visual description.

Caption:
input_caption

Return the result as a Python dictionary like this:

"Visual Detail": [...],
"Spatial Specificity": [...],
"Color & Texture": [...],
"Interpretive Language": [...]

Figure 7 shows an example of visual atomic facts (as visual components) extracted by the prompt.

## D Example of SPECI on text-to-image generation

The two examples in Figure 8 and 9 below illustrate the efficiency of SPECI in estimating the specificity of the caption. It can be seen that the caption with a higher SPECI score can generate an image similar to the original. Hence, if SPECI is used in practice, it will help select a better caption that captures much information from the original image.



A brown dog, likely a working breed, is captured mid-leap over a textured agility ramp, showcasing its muscular build and focused expression. The dog's mouth is open in a wide bark, revealing its teeth, which adds a sense of excitement to the scene. The ramp, covered in a sandy material, is positioned in a grassy area, suggesting an outdoor agility course. In the background, blurred figures of spectators can be seen, some wearing casual clothing, indicating a lively event atmosphere. The bright sunlight enhances the vivid colors of the scene, highlighting the dog's shiny coat and the green grass beneath.

Decomposing to atomic facts

The dog is brown.
The dog is captured mid-leap over a textured agility ramp.
The dog has a muscular build.
The dog has a focused expression.
The dog's mouth is open in a wide bark.
The dog is revealing its teeth.
The ramp is covered in a sandy material.
The ramp is positioned in a grassy area.
The scene suggests an outdoor agility course.
There are blurred figures of spectators in the background.
Some spectators are wearing casual clothing.
The event atmosphere is lively.
The sunlight is bright.
The sunlight enhances the vivid colors of the scene.
The dog's coat is shiny.
The grass beneath is green.

Figure 6: An example of atomic facts decomposition from the original caption text of the image based on the **Atomic Fact Decomposition Prompt**.



A brown dog, likely a working breed, is captured mid-leap over a textured agility ramp, showcasing its muscular build and focused expression. The dog's mouth is open in a wide bark, revealing its teeth, which adds a sense of excitement to the scene. The ramp, covered in a sandy material, is positioned in a grassy area, suggesting an outdoor agility course. In the background, blurred figures of spectators can be seen, some wearing casual clothing, indicating a lively event atmosphere. The bright sunlight enhances the vivid colors of the scene, highlighting the dog's shiny coat and the green grass beneath.

Decomposing to visual components


```
"Visual Detail": [
    "brown dog",
    "working breed",
    "muscular build",
    "focused expression",
    "wide open mouth",
    "teeth",
    "textured agility ramp",
    "sandy material",
    "grassy field"
  ],
"Spatial Specificity": [
    "mid-leap over a textured agility ramp",
    "ramp positioned on a grassy field",
    "blurred figures of spectators in the background"
  ],
"Color & Texture": [
    "bright sunlight",
    "vivid colors",
    "shiny coat",
    "green grass"
  ],
"Interpretive Language": [
    "sense of excitement",
    "lively atmosphere"
  ]
```


Figure 7: An example of visual components decomposition from the original caption text of the image based on the **Visual Atomic Fact Decomposition Prompt**.



Original image

The image depicts a food cart situated on a city street, prominently displaying the sign \"DISABLED COMBAT VETERAN\" above a colorful array of food options. The cart is adorned with two yellow umbrellas, each featuring the words \"FRANKFURTERS & ROLLS\" and \"SABRET,\" indicating the type of food served. In the foreground, a woman with a black backpack and a pink strap stands facing the cart, while a man in a dark jacket appears to be ordering food. The cart showcases a variety of items, including hot dogs, drinks, and snacks, with glass containers filled with beverages visible behind the vendor. The background features a blurred view of a bus and trees, suggesting a bustling urban environment.

**Caption A SPECI** 0.04415755191692901



A street vendor is selling hot dogs and drinks from a small food cart, with two umbrellas providing shade for customers. A man and a woman are standing in front of the cart, waiting to place their orders. The street is lined with parked cars, and a blue car is visible in the background.

**Caption B SPECI** 0.01514376404038043



Figure 8: An example of the image reconstruction task where the two images are generated according to two captions generated by two models, in which Caption A is generated by GPT4o and Caption B is generated by InstructBLIP. The SPECI score is used to estimate the specificity of each caption. It can be seen that a higher SPECI score helps choose a better caption (Caption A) for generating a new image that is mostly similar to the original.

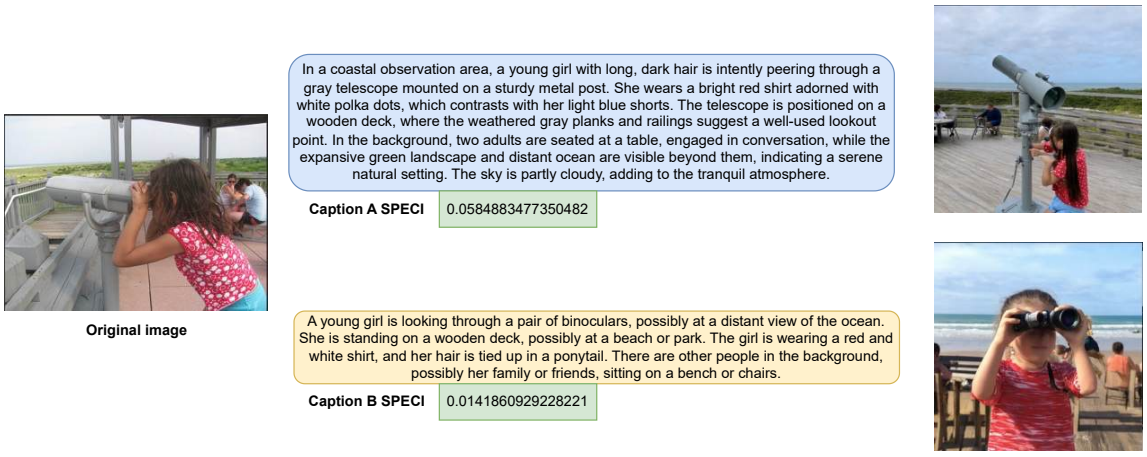



Figure 9: An example of the image reconstruction task where the two images are generated according to two captions generated by two models, in which Caption A is generated by GPT4o and Caption B is generated by InstructBLIP. The SPECI score is used to estimate the specificity of each caption. It can be seen that a higher SPECI score helps choose a better caption (Caption A) for generating a new image that is mostly similar to the original.

## E Prompt for determining specificity by LLMs

**Specificity Determnining Prompt**

You are given a caption that describes an image or scene at the paragraph level.
Assess its specificity on a scale from 1 to 5, where:

1 = extremely vague and generic (very few concrete details)
2 = somewhat vague (some details, but still general)
3 = moderately specific (clear but not richly detailed)
4 = specific (contains several concrete, precise details)
5 = highly specific (rich, precise, and concrete in all aspects)

Consider:
- Level of visual detail
- Spatial and compositional clarity
- Color and texture precision
- Linguistic focus (descriptive precision vs. generic/interpretive wording)

Return only the integer score from 1 to 5 with no explanation.

## F Visual SPECI components test

We randomly take 149 examples from the human preferences dataset and conduct the ANOVA (Table 7) and Turkey HSD (Table 8) tests to investigate the means among visual components in pairs. It can be seen that Visual Detail shows the highest difference from the other components, while spatial with color and linguistic with color indicate the indistinguishability.

| **ANOVA test** | |
|---|---|
| F-statistic | 6.00783 |
| P-value | 0.000490 |

Table 7: ANOVA test among visual components in Visual SPECI .

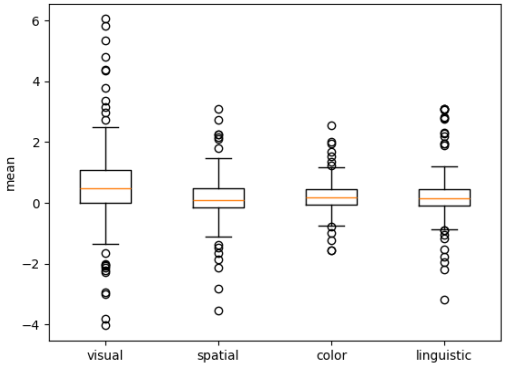



Figure 10: Illustrating the mean of four visual components in Visual SPECI.

| Comparison | Statistic | p-value | Lower CI | Upper CI |
|---|---|---|---|---|
| (visual - spatial) | 0.972 | 0.000 | 0.713 | 1.232 |
| (visual - color) | 1.028 | 0.000 | 0.768 | 1.287 |
| (visual - linguistic) | 1.123 | 0.000 | 0.863 | 1.382 |
| (spatial - color) | 0.056 | 0.946 | -0.204 | 0.315 |
| (spatial - linguistic) | 0.151 | 0.442 | -0.109 | 0.410 |
| (color - linguistic) | 0.095 | 0.782 | -0.165 | 0.354 |

Table 8: Turkey HSD test among visual components in pair in Visual SPECI .

## G Example of human evaluation on specificity

This section demonstrates several examples from the DCI dataset DCI (Urbanek et al., 2024) with human preferences for determining the specificity of two captions, A and B.

| Image | Caption A (Prefered) | Caption B |
|---|---|---|
|  | A horizontally-oriented high angle photo shows a large crowd of men, women, and children gathered on an open plaza. They wear winter coats (mostly the puffy style), winter hats, and gloves and stand with their backs facing the viewer. They face a tall open structure that extends from the left side of the image on a slight curve into the right side of the image. A tiered rooflike top connects a series of columns. The fronts of these roof-like sections are ornamented with abstract decorative motifs and sans-serif text in Cyrillic. Between the crowd and this structure, three large screens have been set up on a stage. In front of the center screen, a band performs on a variety of instruments including a keyboard, accordion, and acoustic guitar. In the background, the bare brown winter branches of the tops of a horizontal line of trees is partially visible. A clear, pale blue, perfectly cloudless sky is overhead. | People in their winter clothes enjoying a winter festival concert outside. |

| Image | Caption A (Prefered) | Caption B |
|---|---|---|
|  | A horizontally-oriented eye level image shows a residential concrete-like balcony destroyed by an explosion in Beirut. The image shows the balcony slightly slanted toward the background on the left side. Two parallel rectangular concrete-like sections have wrecked, twisted, and separated, missing metal safety grates. Torn and shredded white and red striped fabric hangs limply from partially separated metal grates still learning on the balcony. Behind the metal grates, window panes are missing and twisted strips of housing material are partially visible. Stacks of brown cardboard boxes are visible behind the leaning grates on the balcony on the right side. | This is the side of a building. It shows mostly one floor and a little of the floor above it. It has a balcony. It had a metal railing but it is all broken and has a dirty, ripped and striped cloth draped in front of it. There are boxes piled haphazardly behind it. Part of the railing is leaning out. The higher floor seems better. The building is tan and dirty. |

| Image | Caption A (Equal) | Caption B (Equal) |
|---|---|---|
|  | The RheinEnergieStadion in Cologne, Germany, is shown in a full shot on an overcast day. The bleachers are supported by many tall, thin, white pillars. The back of the bleachers is visible as are portions of the side bleachers and the far tall towers that support the white lights. The tall towers have suspension cables on both sides. Underneath the bleachers, facing the viewer, hang 3 vertical white rectangular banners, the first of which reads R̈hein, the second of which reads Ënergie, and the third of which reads S̈TADION. Under the middle banner is a horizontal white banner that reads Q̈UEEN. On a green soccer pitch in the foreground are many people playing soccer, some in white t-shirts and some in dark-red t-shirts. Some of the players also wear knee socks, some in white, some in black, and some in orange. A very short, thin, white fence separates the soccer pitch from the building. The sky is grey-blue with a great deal of cloud cover. | The soccer field in this image takes up the bottom 1/3 of the picture. It stretches from the left to the right of the image. From the center of the field to the left of center, there are around 21 soccer players in red and white jerseys walking on the field. There is one person in a yellow shirt on the left side of the field. Behind the field, there is a long white fence that stretches the entire picture. Behind the fence, there is a row of trees with no leaves. Behind the trees, we can see the back of a large stadium. The base of the stadium is about 1/3 the height of the image. It just about touches the left and right sides of the image. In the upper right and left corner of the stadium, there are metal towers. These towers are about as tall as the stadium's base. There is a long metal cord attached to both towers that droops down in the middle of the ceiling of the stadium. We can just make out some trees on the space left by the stadium on its sides. The sky in the background takes up the background of the top 1/2 of the image. The sky is above the stadium and can be seen on the lower sides of the stadium near the stairs. |

| Image | Caption A | Caption B (Marginal Prefered) |
|---|---|---|
|  | A Boeing 737 passenger airplane with the red word B̈ELAVIAön the side takes off to the left against a blue sky in a full eye-level outdoor shot on a sunny day. The airplane is white with a long, narrow, blue, horizontal stripe from under the cockpit to the back end. In this stripe is the entire long line of small, square, dark windows, as well as the doors at the front and back. The tail, also, is blue. It contains a white oval logo with a red design. The landing gear is still down, showing the wheels suspended below the airplane as it lifts upward. The jet engines are in front of the wings. The wings have dark shadows underneath them. Below the airplane is a long line of evergreen trees, stretching from left-to-right, packed closely together. | There is a white plane. It is mostly white, except for a blue strip that runs along the middle where the windows are and the tail. It has a nose. The windshield is visible on the front. There is a door behind that. It then has equally shaped and spaced windows going across. There is a red rectangle right above where the windows start. To the right of that, it has red writing that says, Ž01cBELAVIA.Ž01d There is another door on the right side. There is small writing to the left of that and then also to the right on the top. There is a little black wheel near the front. There is an engine about 1/3 from the front. There is the wing coming out behind that. There are larger wheels coming down from there. At the back there is a blue tail on the plane. It slants out and then up sharply. It has many grooves running along it, especially near the back. In the middle back, there is a white circle. It has a red curving design in it that has a thick line curving up and then three thinner curving lines going up. There is a straight line and then a line curving down. There is a tail fin coming out. There is a long line of trees in the distance. The trees have brown barks, green leaves and stems, as well as brown branches. They are full. The tops curve up. Above the sky is blue and clear. It is daytime. |

| Image | Caption A | Caption B (Prefered) |
|---|---|---|
|  | A partial exterior of a modern-style home is the focus of an eye-level, medium shot. What appears to be the back of the house is positioned along the right. The two-story house has a matte white first floor with a light gray door to the right of a window door. A white sconce hangs to the right of the gray door, and what appears to be another door is along the right. A large gray vent is on the bottom far left of the first floor, and a smaller vent is below the sconce near the ground. The second floor has a narrow balcony and is made of a light golden brown wood in horizontal planks. The wood is used in the walls and verada-like ceiling of the second floor. The side border around the second floor is made from a darker wood. A concrete patio is outside the door on the first floor, which leads to a light tan path in the distance. On the left is a lawn made of short light green grass. This stretches into the distance until a row of small bushes, followed by a green fence. Beyond the fence is an expansive green field. Beyond the field are the reddish brown roof tops of many houses that are white.The sky is clear blue that becomes lighter along the horizon. A few clouds dot the sky in the distance. Near the lower left corner is a patch of dirt. A gray cylindrical ground light is near the path leading away from the house. | A building sits next to a green field. The building is at least two stories tall, and it has a long balcony stretching across the upper floor. There are several doors on the lower floor. To the right side, there is a grassy area where at least one small sapling is growing. Behind the building, there is a large field which is enclosed by a fence. Several small plants grow at the outside edge of the fence. Beyond the field, there is a neighborhood of houses, and beyond the houses, there are some mountains far off in the distance. The sky is bright blue and mostly free of clouds. |