

# Trust the Model: Compact VLMs as In-Context Judges for Image-Text Data Quality

Supplementary Material

*Daulet Toibazar      Kesen Wang      Sherif Mohamed*  
*Abdulaziz Al-Badawi      Abdulrahman Alfulayt      Pedro J. Moreno*

Humain  
Riyadh, KSA

# 1 Appendix: A

This section illustrates the limitations of web-scale image-caption datasets. Emphasizing captions as the primary focus of this study, figure 1 demonstrates the quality spectrum of web-crawled image descriptions, ranging from inaccurate and noisy captions to accurate examples that accurately reflect the visual content.







CC 12M		<b>Caption:</b> From all of us at Heritage Iron, we wish you a very Merry Christmas! Photo by Super T. IH Farmall Gold Demos owned by <PERSON> of Sebring, FL. International Tractors, Very Merry Christmas, Puns, <PERSON>, Gold, Merry Christmas, Clean Puns, Word Games	
CC 12M		<b>Caption:</b> The benefits and harm of hazelnut	
CC 12M		<b>Caption:</b> Model <PERSON> sports a Giorgio Armani suit with a tee and sneakers.	

Figure 1: Spectrum of web-crawled image captions, demonstrating both low-quality (inaccurate or noisy) and good-quality (precise and descriptive) text-image relationships.

# 2 Appendix: B

To assess the quality of image-text relationships, we established a comprehensive evaluation framework using a 10-point scale, as following:

## STEP 1: IMAGE SAFETY and QUALITY CHECK

- If the image contains unsafe content (graphic violence, explicit content, disturbing imagery) → STOP and assign score 1.
- If the image is extremely low quality, blurry, or uninterpretable → maximum score 3.
- If the image contains significant visual noise or artifacts → maximum score 5.

## STEP 2: CAPTION CONTENT EVALUATION

- Specificity: Does the caption describe specific visible elements or only generic labels?
  - Generic/minimal (e.g., "Room," "Building," "Tree in park") → maximum score 3
  - Basic description with few details → maximum score 6
  - Specific, detailed description of visible elements → can score 7-10
- Accuracy: Are all mentioned elements actually present in the image?
  - Any hallucinated elements not in the image → maximum score 2
  - Minor inaccuracies → maximum score 5
  - Completely accurate → can score 6-10
- Comprehensiveness: Does the caption cover the main visual elements?

- Omits major visual elements → maximum score 4
- Covers most important elements → can score 5-8
- Thoroughly describes all significant elements → can score 9-10

#### STEP 3: SEMANTIC RELATIONSHIP ASSESSMENT

- Semantic alignment: Does the caption capture the meaning and context of the image?
- Caption quality: Is the caption well-written, clear, and informative?

#### STEP 4: FINAL SCORING

- The final score is the LOWEST score from any of the applicable criteria above.

### 3 Appendix: C

$$\bar{S}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} \cos(E_{\text{text}}(c_i), E_{\text{img}}(x_i)),$$

where  $N_d$  is the number of examples in split  $d$ ,  $c_i$  is the  $i$ th caption,  $x_i$  is the  $i$ th image, and  $E_{\text{text}}, E_{\text{img}}$  are the CLIP text and image encoders.

### 4 Appendix: D

$$\text{PPL}_d = \exp\left(\frac{1}{N_d} \sum_{i=1}^{N_d} \mathcal{L}(c_i)\right),$$

where  $\mathcal{L}(c_i)$  is the token-level cross-entropy loss of GPT-2 on caption  $c_i$  and  $N_d$  is the number of samples in dataset  $d$ .