A. Evaluation Prompt

<|start_header_id|>system<|end_header_id|>

You are a pattern-following assistant that can only answer with "Yes" or "No". Your goal is to determine whether a predicted caption conveys a similar enough meaning to the ground truth caption provided.<|eot_id|> <|start_header_id|>user<|end_header_id|>

Instruction:
Determine if the predicted caption conveys a similar meaning to the ground truth caption.

Ground truth caption:
A man is riding a bicycle through a park.

Predicted caption
A person is cycling along a path in a park.

Does the predicted caption convey a similar meaning to the ground truth caption (Yes or No)?<|eot_id|> <|start_header_id|>assistant<|end_header_id|>

Yes<|eot_id|><|start_header_id|>user<|end_header_id|>

Good job! Indeed, the predicted caption conveys a similar meaning to the ground truth. Both describe a person riding a bicycle in a park, even though different words are used. The core meaning is preserved.

Instruction:
Determine if the predicted caption conveys a similar meaning to the ground truth caption.

Ground truth caption:
A woman is sitting on a wooden bench in the park, reading a paperback novel under the shade of a tree.

Predicted caption:
A woman relaxes in a shaded area of the park, sitting on a bench while enjoying a book.

Does the predicted caption convey a similar meaning to the ground truth caption (Yes or No)?<|eot_id|> <|start_header_id|>assistant<|end_header_id|>

Yes<|eot_id|><|start_header_id|> user<|end_header_id|>

Great! Although the wording differs, the predicted caption captures the essence of the ground truth. Both describe a woman sitting on a bench in a shaded park area, reading a book. While the predicted caption simplifies certain details, such as omitting the specific mention of the "paperback novel" and "under the shade of a tree," it still conveys the same overall scene and activity, making the meaning similar.

Let's do one more. Remember to answer with one word either "Yes" or "No".

Instruction:
Determine if the predicted caption conveys a similar meaning to the ground truth caption.

Ground truth caption:
[...]

Predicted caption:
[...]

Does the predicted caption convey a similar meaning to the ground truth caption (Yes or No)?: <|eot_id|> <|start_header_id|> assistant <|end_header_id|>

Figure 6. Evaluation prompt used with Llama 3.1 70B to annotate the predicted descriptions. The prompt structure is adapted towards description prediction from Alpaca-Eval. The positions are indicated by placeholders [...], where the ground truth and predictions to be annotated are inserted.

B. Human Annotation Study

Study Design. Ninety examples were selected for evaluation, stratified by context length and sampled proportionally from the outputs of IMAGECHAIN, MLLM, and MLLM-FT. Three human annotators were recruited via Prolific (https://www.prolific.com/), ensuring a balanced gender distribution and a minimum education level of an undergraduate degree. Annotators provided informed consent and were compensated at a rate of £9 per hour, in line with fair pay guidelines. Each participant annotated 12 examples, which included three gold-standard control examples (labeled by the authors for quality control; annotators who failed these controls were excluded from the final analysis). For each example, annotators compared the ground truth next-scene description with a model-generated description, rating the semantic overlap on a five-point Likert scale (1 = "completely different meanings", 5 = "essentially identical meanings"). The instructions and the annotation interface are shown in Figure 7 and Figure 8 respectively.

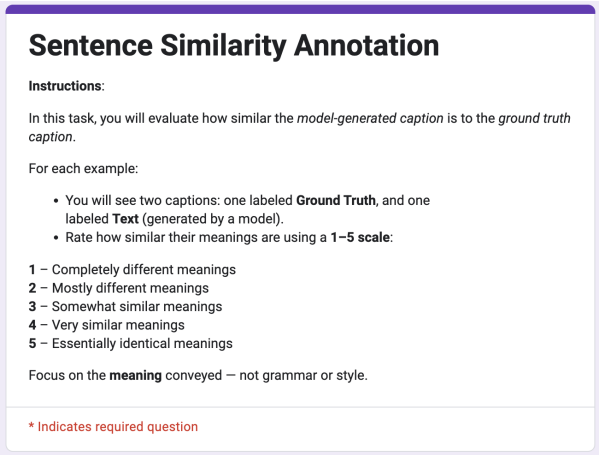

Figure 7. Instructions of human annotation study

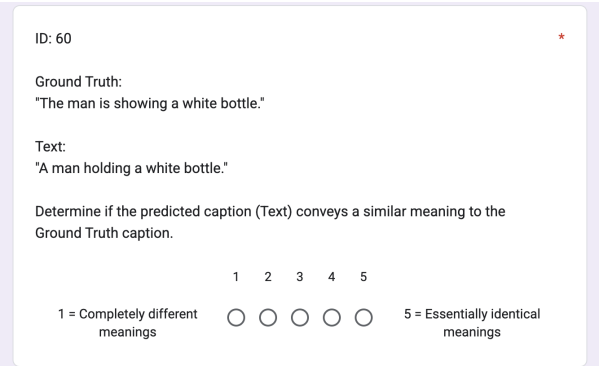

Figure 8. Example of the annotation interface

Data Analysis and Inter-Annotator Agreement. For subsequent quantitative analysis [17], the 5-point Likert scores were binarized: scores ≥ 3 were considered as positive class. This threshold resulted in a positive class prevalence of 23% in the human-annotated dataset. The reference label for each example was determined by a majority vote among the three annotators.

Inter-annotator agreement is fair [28], with Fleiss' $\kappa = 0.30$ [14] and Krippendorff's $\alpha_{\text{ordinal}} = 0.32$ [27] when computed on the original 1-to-5 ranks, indicating moderate consistency in how annotators judge semantic similarity, even if they differ on the exact rating level.

Alignment with Llama Judge. Compared to human annotations, the Llama 3.1 70B judge achieves 72% accuracy, Wilson [56] 95% CI: 62–80%, with balanced errors (10 false positives, 15 false negatives). McNemar's exact test for directional bias [39] produced a p-value of 0.42, indicating no significant systematic bias in one direction over the other. Overall, the SimRate produced by the Llama 3.1 70B judge differed from the human-aggregated SimRate by 5.6 percentage points (95% CI for the difference: –17.3 pp to +6.2 pp). Since this confidence interval includes zero, there is no statistically significant difference between the overall SimRate produced by the Llama judge and human evaluators.

Model-Specific Calibration Differences. While the overall alignment was accurate, the Llama 3.1 70B judge exhibited some model-specific calibration differences when compared to human ratings:

- **For IMAGECHAIN outputs:** The Llama judge was conservative. It achieved 87% accuracy on these samples but tended to underestimate positive cases. The human-rated SimRate for IMAGECHAIN was 13 percentage points higher than the Llama-judged SimRate.
- **For MLLM outputs:** The Llama judge was slightly liberal. It achieved 67% accuracy and tended to overestimate positive cases. The human-rated SimRate for MLLM was 7 percentage points lower.
- **For MLLM-FT outputs:** The Llama judge was again conservative. It achieved 67% accuracy, and the human-rated SimRate for MLLM-FT was 10 percentage points higher.

These differences suggest that if the main results were ranked purely based on human labels, IMAGECHAIN and MLLM-FT would likely see their scores increase relative to the Llama-judged SimRates, while MLLM's score would decrease.

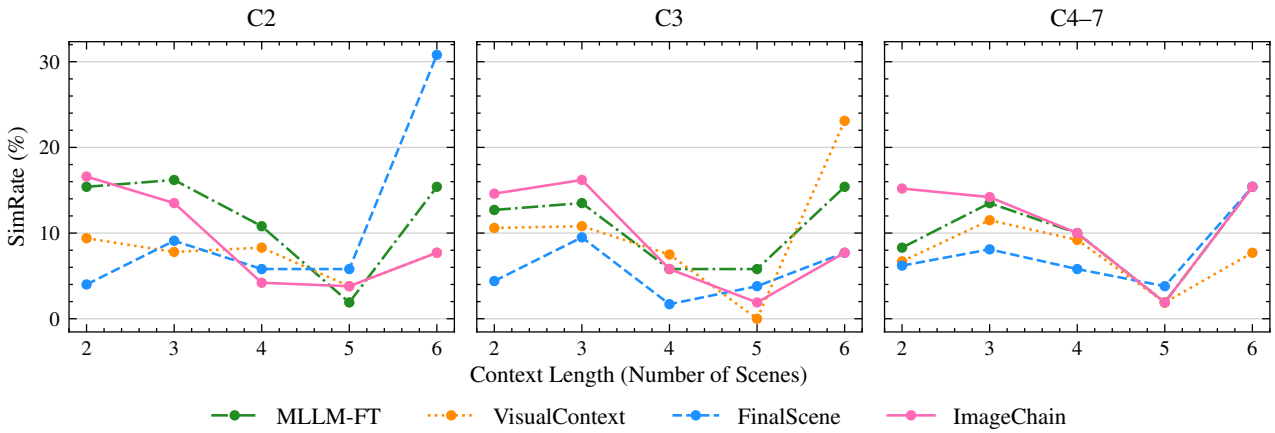

Figure 9. Model performance (SimRate) for fine-tuned models trained on different context lengths (C2, C3, C4-7). IMAGECHAIN achieves the highest overall SimRate when trained on long contexts (C4-7).

C. Context Length Ablation

Figure 9 shows the models performance for fine-tuned models trained on different context lengths. IMAGECHAIN achieves the highest overall SimRate when evaluated on C2-6, particularly when trained on longer contexts, reaching 13.6% when trained on C4-7. This surpasses other models, such as MLLM-FT, which achieves 9.8% under the same training conditions. MLLM-FT excels in short contexts but struggles with longer dependencies, suggesting limitations in handling extended sequences without explicit sequence modeling. VisualContext under-performs on longer sequences (7.7% trained on C4-7 and evaluated on C6), highlighting the benefit of including text descriptions for fine-tuning in long contexts.

We investigated the performance drop on C5 (see Table 1) for all models by examining the samples within this context, the evaluation pipeline, the context length, and the source of the samples in both training and evaluation sets. We did not identify any mistakes or notable differences across these factors. We therefore attribute the decline in performance to data variability, which we hypothesize could be mitigated by increasing the number of samples for this context length during fine-tuning. A closer investigation of this is left to future work.

D. Ethics Statement

This work uses publicly available licensed (CC BY 4.0) datasets consistent with their intended use (research) to ensure transparency and reproducibility. While IMAGECHAIN enhances sequential reasoning, it may inherit biases from pre-trained models and datasets. Our framework is designed for research and development purposes, and we encourage responsible use, particularly in applications involving decision-making in sensitive domains such as healthcare and robotics. We acknowledge the use of Microsoft Copilot during the development of the coding experiments. To ensure ethical data annotation, we recruited participants, balancing gender. Annotators provided informed consent and were compensated fairly at a rate of £9 per hour. Each participant annotated 12 examples (see Appendix B) using a task interface designed for clarity and ease of use. The task involved no sensitive or harmful content, and all data used were synthetic or publicly available, with no personally identifiable information involved.