Controllable Image Captioning with Prompt-Conditioned Scene Rewards

Jongyeop Hyun[1]* **Taeyoung Kim**[1]* **Hyounghun Kim**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH
[2]Department of Computer Science and Engineering, POSTECH
{mldljyh,kty4119,h.kim}@postech.ac.kr

Abstract

Large Vision-Language Models produce fluent image descriptions but offer limited semantic control: users cannot reliably specify whether captions should emphasize attributes, relations, or particular image regions. We present **F**ine-grained Capti**o**ning **C**ontrol **U**sing **S**cene Rewards (**FoCUS**), a controllable image captioning method that lets users steer captions toward specific semantic emphases through natural-language control prompts. The core idea is a prompt-conditioned control objective based on scene-graph-aligned component scores. Generated captions are parsed and aligned to scene-graph components—objects, attributes, and relations—and these components are differentially weighted, including negative weights, according to the requested emphasis. We optimize this objective with GRPO and further improve its reliability through a stricter object validity threshold and reasoning-based verification for attribute and relation scoring. To evaluate controllability, we introduce **S**emantic **Co**ntrol and **P**recision **E**valuation (**SCoPE**), a benchmark with contrastive Include/Avoid constraints for measuring both target content coverage and out-of-scope suppression. Experiments on two VLM backbones show that FoCUS consistently improves controllability and fine-grained caption quality while largely maintaining general caption performance.[1]

1 Introduction

Image captioning has advanced with the emergence of Large Vision-Language Models (LVLMs) that couple strong visual encoders with powerful large language models, enabling fluent long-form descriptions and instruction-following behavior (Li et al., 2023; Liu et al., 2023; Bai et al., 2023; Chen et al., 2024b). In parallel, richer supervision—such as detailed caption datasets and scene-level annotations—has improved coverage and specificity (Chen et al., 2024a; Onoe et al., 2024; Pont-Tuset et al., 2020). Despite this progress, a practical limitation remains: users often need *semantic control* over what a caption emphasizes. For the same image, one may want an attribute-centric description (colors, materials), a relation-centric description (spatial/functional relationships), or a foreground/background-focused description. Current LVLM captioners produce a single "best overall" caption and offer limited mechanisms to steer *what to focus on*. Figure 1 shows that zero-shot prompting often defaults to a generic scene description rather than the requested semantic focus.

Figure 1: Comparison of Zero-shot and FoCUS generation. Zero-shot defaults to a generic scene description (red), whereas FoCUS emphasizes object attributes (green) when generating an Attribute-focused caption.

Existing controllable captioning methods usu-

*Equal contribution.
[1]The code and dataset are publicly available at https://focus-emnlp2026.github.io/.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 59590–59615
October 24-29, 2026 ©2026 Association for Computational Linguistics

ally require structured controls at inference time, such as length tokens (Deng et al., 2020), regions or boxes (Cornia et al., 2019), or formal semantic graphs (Basioti et al., 2024). Natural-language prompting is a more natural interface, but prompting alone is often unreliable for fine-grained emphasis and suppression, with models drifting toward generic high-probability content.

A natural alternative is to decompose captions into interpretable semantic units and let the prompt determine which units to reward or penalize. Scene graphs—objects, attributes, and relations—provide such a decomposition, and recent work has shown their value for fine-grained caption evaluation (Dong et al., 2024; Lu et al., 2025). However, they have mainly been used for post-hoc evaluation or prompt-agnostic training, not as a prompt-conditioned control signal that rewards requested content while suppressing off-scope content.

We propose **F**ine-grained Capti**o**ning **C**ontrol **U**sing **S**cene Rewards (**FoCUS**), a controllable captioning method built around a prompt-conditioned scene-graph objective. Given a natural-language prompt specifying a semantic emphasis (e.g., attribute-, relation-, foreground-, background-focused, or general), FoCUS parses a generated caption into scene-graph components and aggregates scores using prompt-specific signed weights: positive weights reward requested content, while negative weights penalize off-scope content. This turns scene-graph decomposition from an evaluation tool into an interpretable learning signal for caption generation, without architectural changes or structured side inputs. In our implementation, we optimize this objective with GRPO (Shao et al., 2024). To reduce reward noise, we use stricter object-validity checks and reasoning-based attribute/relation verification.

Evaluation is another challenge. Standard captioning metrics measure overall quality, but not *contrastive controllability*: covering requested content while avoiding off-scope content (Anderson et al., 2016). Existing fine-grained evaluators provide component-level feedback, but they do not explicitly test this include-vs.-avoid behavior (Dong et al., 2024; Lu et al., 2025). We therefore introduce **S**emantic **Co**ntrol and **P**recision **E**valuation (**SCoPE**), a benchmark with category-specific *Include* and *Avoid* lists derived from curated captions. SCoPE measures target coverage, off-scope suppression, and factual consistency.

Our main contributions are:

- We propose FoCUS, a controllable image captioning method based on a prompt-conditioned control objective over scene-graph components, with signed weights that both emphasize requested content and suppress off-scope content.
- We improve the reliability of scene-graph rewards for learning through stricter object matching and reasoning-based verification for attributes and relations, and show gains over holistic and fixed-weight reward baselines.
- We introduce SCoPE, a benchmark for controllable captioning with contrastive Include/Avoid constraints for measuring coverage, adherence, and contradiction-based faithfulness.

2 Related Work

Large Vision-Language Models and Detailed Captioning. Modern LVLMs couple strong visual encoders with large language models and produce fluent, detailed captions (Li et al., 2023; Liu et al., 2023; Bai et al., 2023; Chen et al., 2024b). This progress is supported by richer supervision sources, including ShareGPT4V (Chen et al., 2024a), DOCCI (Onoe et al., 2024), and Localized Narratives (Pont-Tuset et al., 2020). However, these models typically default to a single "best overall" description and offer limited, unreliable control over which semantics (e.g., attributes vs. relations, foreground vs. background) are emphasized.

Controllable Image Captioning. Prior controllable captioning methods usually rely on structured side inputs at inference time, such as explicit length signals (Deng et al., 2020), region-level inputs (e.g., boxes or masks) (Cornia et al., 2019), or formal semantic specifications such as AMR graphs (Basioti et al., 2024). Our setting instead uses natural-language control prompts and learns to realize them through a prompt-conditioned objective over scene-graph components, without additional structured inputs at inference time or architectural changes.

Training Objectives for Caption Alignment. Caption generation has been optimized beyond standard cross-entropy using metric-based objectives such as SCST with CIDEr (Rennie et al., 2017; Vedantam et al., 2015), embedding-based signals such as CLIPScore (Hessel et al., 2021), and preference-based objectives such as DPO (Rafailov et al., 2023). Scene-graph-based metrics such as SPICE, CAPTURE, and CompreCap provide finer object-, attribute-, and relation-level feed-

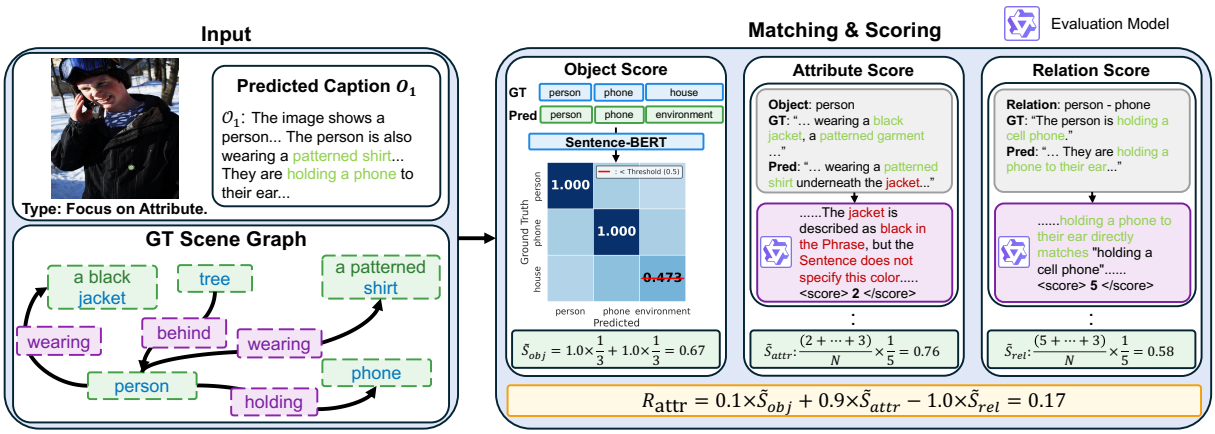


Figure 2: Overview of the FoCUS pipeline. During training, given an image and a natural-language control prompt, FoCUS aligns the generated caption with scene-graph annotations and computes component scores for objects, attributes, and relations. These scores are aggregated with prompt-conditioned signed weights to form the control objective, which emphasizes requested content and suppresses off-scope content.

back (Anderson et al., 2016; Dong et al., 2024; Lu et al., 2025), and recent work such as SC-Captioner (Zhang et al., 2025) uses scene-graph components as compositional supervision. Our work extends this line by making scene-graph rewards prompt-conditioned: object, attribute, and relation scores are combined with prompt-specific signed weights to produce captions with user-specified semantic emphases.

Captioning Benchmarks and Fine-grained Evaluation. Captioning benchmarks commonly measure overall quality with metrics such as CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). More recent evaluations, including CAPTURE (Dong et al., 2024) and CompreCap (Lu et al., 2025), provide component-wise analysis but do not explicitly test *contrastive controllability*—maximizing requested content while suppressing off-scope content. We address this gap with SCoPE, which introduces category-specific *Include* and *Avoid* constraints to measure both coverage (recall) and suppression (precision) under user-specified semantic emphases.

3 Method

We present FoCUS, a controllable image captioning method built around a prompt-conditioned, scene-graph-aligned objective. Given a natural-language control prompt, FoCUS decomposes a generated caption into objects, attributes, and relations, then aggregates their scores with prompt-specific signed weights to reward requested content and penalize off-scope content. We optimize this objective with a two-stage SFT+GRPO (Shao et al., 2024) pipeline. Figure 2 illustrates FoCUS.

3.1 Scene-Graph-Aligned Component Scores

Our scoring function is inspired by CompreCap (Lu et al., 2025), which decomposes captions into objects, attributes, and relations for component-wise scoring, but we adapt it for reward-based training. Directly optimizing such a pipeline in RL can be noisy: embedding-based object matching may drift without a validity threshold, and zero-shot LLM judges may over-reward generic descriptions. We therefore introduce stricter object validity checks and stronger Chain-of-Thought-based verification (Wei et al., 2022) for attribute and relation scoring with modern LLM judges. Additional background and motivation are in Appendix A.

We parse the generated caption into scene-graph elements using spaCy (Montani et al., 2023) for noun extraction and lemmatization, then compute scores at three semantic levels.

Object Score (S_{obj}). We compute the soft coverage of annotated objects mentioned in the generated caption. Let $\mathcal{C} = \{c_i\}_{i=1}^{n}$ be the candidate object nouns extracted from the generated caption, and let $\mathcal{G} = \{g_j\}_{j=1}^{m}$ be the ground-truth object categories from the annotated scene-graph. Using Sentence-BERT (Reimers and Gurevych, 2019), we compute pairwise similarities and construct a mutual-best match indicator matrix $\mathbf{S}' \in \{0,1\}^{n\times m}$ with a validity threshold $\tau = 0.5$ (details in Appendix B). The object score is defined as the fraction of ground-truth objects successfully matched:

$$S_{\text{obj}} = \frac{1}{m}\sum_{j=1}^{m} \max_{i} S'_{i,j}. \tag{1}$$

Attribute Score (S_{attr}). For objects matched at the object level, we evaluate the quality of attribute descriptions. Let $\mathcal{J}$ denote the index set of matched ground-truth objects. For each $j \in \mathcal{J}$, we employ `Qwen3-30B-A3B-Instruct` (Yang et al., 2025) with CoT prompting to produce an integer attribute alignment score $a_j \in \{0, 1, \ldots, 5\}$. The attribute score is the mean over all matched objects:

$$S_{\text{attr}} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} a_j. \tag{2}$$

If no objects are matched, we set $S_{\text{attr}} = 0$.

Relation Score (S_{rel}). For relation evaluation, we gather subcaptions containing all objects involved in each annotated directed relation and employ the same CoT-based judge to assess whether the generated caption correctly captures the directional relationship, producing a score $r_k \in \{0, 1, \ldots, 5\}$ for each relation k. The relation score is:

$$S_{\text{rel}} = \frac{1}{|\mathcal{E}|} \sum_{k \in \mathcal{E}} r_k, \tag{3}$$

where $\mathcal{E}$ denotes the set of annotated directed relations. If $|\mathcal{E}| = 0$, we set $S_{\text{rel}} = 0$. For stable optimization, we linearly scale all component scores to $[0, 1]$ (denoted as $\tilde{S}$), specifically setting $\tilde{S}_{\text{obj}} = S_{\text{obj}}$, $\tilde{S}_{\text{attr}} = S_{\text{attr}}/5$, and $\tilde{S}_{\text{rel}} = S_{\text{rel}}/5$. The exact judge prompts used for attribute and relation scoring are provided in Appendix L.1.

Foreground and Background Scores. To enable spatially aware controllable captioning, we classify objects as foreground (salient, central) or background (contextual, peripheral). We compute $\tilde{S}_{\text{fg}}$ and $\tilde{S}_{\text{bg}}$ by aggregating object, attribute, and relation scores over the corresponding subset, using the same weights as the general caption reward:

$$\tilde{S}_{\text{fg/bg}} = 0.25\tilde{S}_{\text{obj}}^{fg/bg} + 0.35\tilde{S}_{\text{attr}}^{fg/bg} + 0.40\tilde{S}_{\text{rel}}^{fg/bg}. \tag{4}$$

3.2 Prompt-Conditioned Control Objective

The central mechanism in FoCUS is a prompt-conditioned aggregation of scene-graph component scores. Given a generated caption y, a control prompt category p, and scene-graph annotations z^* for the image, we define

$$R(y \mid p, z^*) = \sum_{k \in \mathcal{K}} w_k(p)\, \tilde{S}_k(y; z^*), \tag{5}$$

where $\mathcal{K}$ denotes the relevant semantic components, $\tilde{S}_k(y; z^*) \in [0, 1]$ is the normalized score for component k, and $w_k(p)$ is a prompt-specific weight. Below, we omit the dependence of $\tilde{S}_k$ on (y, z^*) for brevity. Signed weights let the objective reward requested content, suppress off-scope content, and adapt its component emphasis to the user's request, unlike fixed or prompt-agnostic rewards.

We consider five prompt categories: *General*, *Attribute-*, *Relation-*, *Foreground-*, and *Background-Focused* (see Appendix C for descriptions). We use CompreCap's unified weighting scheme for the general prompt, which assigns larger weights to more challenging semantic components and was designed to align better with human judgments. The remaining prompts use signed weights to induce category-specific emphasis and suppression:

$$\begin{aligned} R_{\text{general}} &= 0.25\,\tilde{S}_{\text{obj}} + 0.35\,\tilde{S}_{\text{attr}} + 0.40\,\tilde{S}_{\text{rel}}, \\ R_{\text{attr}} &= 0.10\,\tilde{S}_{\text{obj}} + 0.90\,\tilde{S}_{\text{attr}} - 1.00\,\tilde{S}_{\text{rel}}, \\ R_{\text{rel}} &= 0.10\,\tilde{S}_{\text{obj}} - 1.00\,\tilde{S}_{\text{attr}} + 0.90\,\tilde{S}_{\text{rel}}, \end{aligned} \tag{6}$$

$$R_{\text{fg}} = \tilde{S}_{\text{fg}} - \tilde{S}_{\text{bg}}, \qquad R_{\text{bg}} = \tilde{S}_{\text{bg}} - \tilde{S}_{\text{fg}}. \tag{7}$$

3.3 Optimization with GRPO

We optimize the prompt-conditioned control objective above with GRPO. For an image–prompt–annotation triple (x, p, z^*) and generated caption $y \sim \pi_\theta(\cdot \mid x, p)$, we optimize

$$\begin{aligned} \max_{\theta}\, \mathbb{E}_{\substack{(x,p,z^*)\sim\mathcal{D} \\ y\sim\pi_\theta(\cdot|x,p)}} \Big[& R(y \mid p, z^*) \\ & - \beta \mathrm{D}_{\mathrm{KL}}\big(\pi_\theta(\cdot|x,p) \,\|\, \pi_{\text{ref}}(\cdot|x,p)\big)\Big], \end{aligned} \tag{8}$$

where $R(y \mid p, z^*)$ is computed from the scene-graph annotations for the training example. We sample multiple captions per example to compute group-relative advantages for stable critic-free updates, and train on all five prompt categories for robustness across control settings.

4 Benchmark: SCoPE

Standard captioning benchmarks measure overall quality, but not *contrastive controllability*: covering target content while suppressing off-scope content. We therefore introduce **S**emantic **Co**ntrol and **P**recision **E**valuation (**SCoPE**), a benchmark with category-specific *Include* and *Avoid* lists derived from curated captions. Figure 3 illustrates the construction and evaluation pipeline.

4.1 Benchmark Construction

SCoPE is constructed in three stages: category-specific caption generation, automated refinement

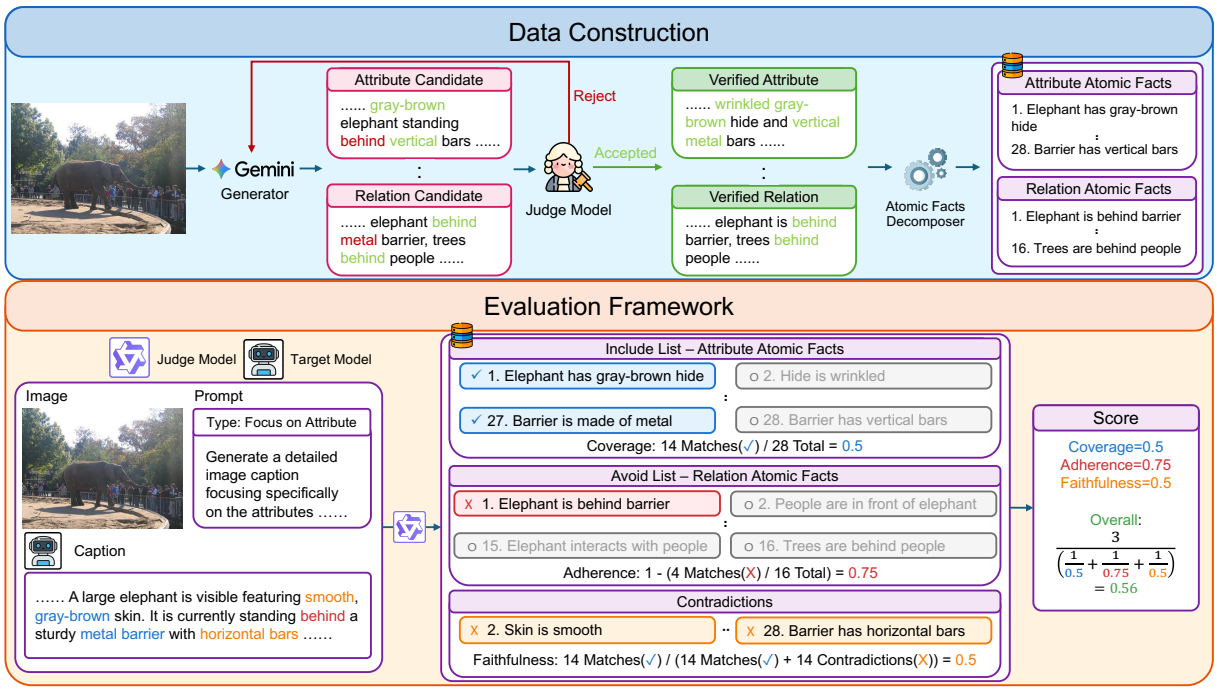


Figure 3: Overview of the SCoPE pipeline. The framework operates in two stages: Data Construction (top), which utilizes a generator-verifier loop to extract and decompose verified atomic facts; and the Evaluation Framework (bottom), which assesses model controllability by scoring generated captions against contrastive *Include* (target) and *Avoid* (out-of-scope) lists to derive Coverage, Adherence, and Faithfulness metrics.

to enforce category focus, and atomic fact extraction to form contrastive Include/Avoid lists.

Image curation. We manually curated 189 images from COCO, CompreCap, and DOCCI, ensuring clear category captions and no overlap with SFT or GRPO training data. See Appendix E.1 for details.

Caption generation and atomic fact extraction. For each image, we consider four categories: *Attribute*, *Relation*, *Foreground*, and *Background*. We use Gemini-3-Flash (Deepmind, 2025) to generate category-specific captions, iteratively refining them to enforce a strict focus. We then decompose each verified caption into atomic facts. For an image x and focus p, facts from the caption for p form the Include list $I_{x,p}$, while facts from the complementary focus form the Avoid list $A_{x,p}$ (Attribute $\leftrightarrow$ Relation, Foreground $\leftrightarrow$ Background). Thus, both the contents and sizes of the lists vary by image and focus. This contrastive design directly measures target coverage and off-scope suppression. Details and benchmark statistics are provided in Appendix E.2 and Appendix F.

4.2 Evaluation Metrics

To assess controllability on SCoPE, we employ an LLM judge (Qwen3-32B-Thinking (Yang et al., 2025)) to semantically verify generated captions C against atomic fact lists. SCoPE evaluates three complementary dimensions—Coverage, Adherence, and Faithfulness—and aggregates them into an overall Score. The inference prompts and LLM-judge prompts used in SCoPE are provided in Appendix L.2 and Appendix L.3, respectively.

Coverage. Coverage measures the recall of target information from the Include list. Let N_{match} be the number of facts in $I_{x,p}$ verified in caption C:

$$\text{Coverage} = \frac{N_{\text{match}}}{|I_{x,p}|}. \tag{9}$$

Adherence. Adherence measures the caption's suppression of off-scope content. Let $N_{\text{violation}}$ be the number of facts from $A_{x,p}$ detected in C:

$$\text{Adherence} = 1 - \frac{N_{\text{violation}}}{|A_{x,p}|}. \tag{10}$$

Faithfulness. Faithfulness measures factual consistency by penalizing cases where the caption explicitly states something incompatible with a target atomic fact. Let $N_{\text{contradiction}}$ denote the number of atomic facts in $I_{x,p}$ that are contradicted by caption C. We count a contradiction only when C asserts an incompatible value for the same property (e.g., mismatches in color, number, material, or spatial relations), as determined by the LLM verifier; mere

Base Model	Method	SCOPE				
		Attribute	Relation	Foreground	Background	Overall
Qwen2.5-VL-3B	Zero-shot	5.93	29.66	13.88	13.17	15.66
	SFT	25.01	32.84	15.02	17.58	22.61
	SFT+CLIP	25.35	41.01	20.15	19.10	26.40
	SFT+CompreCap	30.11	27.84	20.00	25.46	25.85
	SFT+FOCUS	**30.44** (+24.51)	**45.55** (+15.89)	**24.38** (+10.50)	**26.58** (+13.41)	**31.74** (+16.08)
InternVL3-2B	Zero-shot	8.38	47.46	22.48	24.19	25.63
	SFT	32.43	46.28	20.18	25.93	31.21
	SFT+CLIP	27.97	40.38	23.08	22.11	28.39
	SFT+CompreCap	34.71	33.55	19.57	29.22	29.26
	SFT+FOCUS	**36.57** (+28.19)	**54.21** (+6.75)	**25.29** (+2.81)	**31.37** (+7.18)	**36.86** (+11.23)

		CompreCap					
		General	Attribute	Relation	Foreground	Background	Overall
Qwen2.5-VL-3B	Zero-shot	41.62	37.17	47.93	49.90	38.77	43.08
	SFT	41.49	42.59	50.52	52.53	48.26	47.08
	SFT+CLIP	41.97	40.34	50.67	54.33	39.47	45.35
	SFT+CompreCap	45.21	43.34	45.20	49.86	50.19	46.76
	SFT+FOCUS	**47.51** (+5.89)	**45.05** (+7.88)	**54.68** (+6.75)	**55.66** (+5.76)	**51.20** (+12.43)	**50.82** (+7.74)
InternVL3-2B	Zero-shot	39.99	35.33	50.46	51.24	38.77	43.16
	SFT	44.68	43.79	52.50	51.95	48.22	48.23
	SFT+CLIP	30.30	33.26	37.24	43.94	33.99	35.75
	SFT+CompreCap	46.10	44.99	46.29	51.32	50.29	47.80
	SFT+FOCUS	**49.38** (+9.39)	**47.25** (+11.92)	**56.38** (+5.92)	**55.38** (+4.14)	**51.67** (+12.90)	**52.02** (+8.86)

Table 1: Performance comparison using SCOPE and CompreCap. SCOPE and CompreCap scores for two LVLM backbones under Zero-shot, SFT, and GRPO variants. SFT+FOCUS achieves the best performance and improves category-wise control without degrading general caption quality. Parentheses denote gains over zero-shot baseline.

omissions or lack of mention are not counted as contradictions. We compute:

$$\text{Faithfulness} = \frac{N_{\text{match}}}{N_{\text{match}} + N_{\text{contradiction}}}. \quad (11)$$

Overall Score. We aggregate the three metrics using the harmonic mean to ensure balanced performance across all dimensions:

$$\text{Score} = \frac{3}{\frac{1}{\text{Coverage}} + \frac{1}{\text{Adherence}} + \frac{1}{\text{Faithfulness}}}. \quad (12)$$

In Appendix G, we discuss the interpretation of the overall score with supporting empirical evidence.

Human Alignment. We validate SCOPE with an Amazon MTurk[2] study on 140 image–category pairs. For each pair, annotators compare three captions: either a within-backbone triplet (Zero-shot, SFT, and SFT+FOCUS) or an external-model triplet (LLaVA-NeXT-34B (Liu et al., 2024), Qwen3-VL-32B-Instruct (Bai et al., 2025a), and Gemini-3-Flash (Deepmind, 2025)). SCOPE correlates strongly with human preferences (Spearman's $\rho = 0.5704$), substantially outperforming a CompreCap-based controllability metric computed via category-specific reweighting ($\rho = 0.2291$). These conclusions are robust to judge choice: smaller same-family and backbone-mismatched judges show similar human alignment, and fixed-judge rescoring preserves the advantage of FOCUS. Details are provided in Appendix H.

[2]https://www.mturk.com/

5 Experiments

Experimental Setup. We experiment with two VLM backbones, **Qwen2.5-VL-3B-Instruct** (Bai et al., 2025b) and **InternVL3-2B** (Zhu et al., 2025). Each model is trained with a two-stage pipeline consisting of supervised fine-tuning (SFT) followed by GRPO using the proposed prompt-conditioned control objective. Training data construction for SFT and GRPO, together with implementation details and hyperparameters, are provided in Appendix I.1 and Appendix I.2. For SFT+FOCUS, results are averaged over three runs with different seeds. Per-seed results and standard deviations are reported in Appendix I.3.

Evaluation Setup. We use three complementary evaluation protocols: (i) SCOPE to directly measure controllability via Include/Avoid lists over four categories; (ii) a CompreCap-based compo-

Base Model	SCoPE				
	Zero-shot	SFT	SFT+CLIP	SFT+CompreCap	**SFT+FoCUS**
Qwen2.5-VL-3B	12.80	23.22	27.37	27.31	**32.56**
InternVL3-2B	20.46	31.01	26.90	28.50	**35.62**

Table 2: SCoPE Overall scores under stronger include and exclude inference prompts.

nent evaluation with category-specific reweighting (scaled 0–100) on the held-out CompreCap split, using stricter object validity and a stronger LLM judge (Appendix I.5); and (iii) a reference-based general-caption evaluation on 5,000 DOCCI test images using standard captioning metrics and fine-grained caption evaluators (Appendix I.6).

Baselines. We compare against: **Zero-shot** prompting of the pretrained backbone, **SFT** (cross-entropy fine-tuning only), **SFT+CLIP** (GRPO with CLIP similarity reward after adding CLIP-based focus indicators), and **SFT+CompreCap** (GRPO using the original CompreCap metric as reward). The two GRPO baselines represent alternative training objectives: a holistic similarity reward and a CompreCap-style scene-graph reward. Expanded baseline descriptions are in Appendix I.7.

6 Results and Analysis

6.1 Main Results

Table 1 summarizes results on the first two evaluation protocols: SCoPE for contrastive controllability (Include/Avoid) and the CompreCap-based component evaluation for fine-grained factual alignment. Across two LVLM backbones, SFT+FoCUS consistently achieves the best overall performance and the strongest category-wise controllability.

Controllability on SCoPE. Across both backbones, SFT+FoCUS achieves the best overall SCoPE score and improves consistently across all four control categories. Relative to zero-shot prompting, FoCUS yields large gains in overall controllability (approximately +16 points on Qwen2.5-VL-3B and +11 points on InternVL3-2B), indicating the learned policy follows the requested emphasis while better suppressing off-scope content. Qualitative examples in Appendix M illustrate this behavior. Notably, FoCUS also outperforms SFT and GRPO baselines (SFT+CLIP and SFT+CompreCap) on both backbones, showing that the proposed training objective yields more reliable semantic control than these alternatives. This advantage also holds when SCoPE is evaluated separately on CompreCap-origin and non-CompreCap images for both backbones; 95% image-bootstrap confidence intervals on the full benchmark further show clear separation from the strongest baselines (Appendix I.4). We also evaluate a stricter prompt-engineering setting with explicit inclusion/exclusion rules. As shown in Appendix J.1, stronger prompting alone does not close the gap to SFT+FoCUS on either backbone.

Fine-grained factual alignment on CompreCap. FoCUS also delivers the strongest overall CompreCap performance for both backbones (approximately +7–9 points over zero-shot overall), with improvements that are broadly distributed across general captions and the four fine-grained categories. This indicates that the gains on SCoPE are accompanied by stronger compositional grounding on objects, attributes, and relations, while maintaining or improving factual correctness. These gains are consistent with our design: additional ablations in Appendix J.2 show that replacing prompt-conditioned contrastive weighting with either a fixed global reward mixture or target-only up-weighting leads to weaker cross-category controllability. We additionally evaluate general captioning on 5,000 DOCCI test images. Across the two backbones, FoCUS largely maintains standard reference-based caption quality while improving general-prompt fine-grained alignment. Full results are reported in Appendix I.6.

Control under explicit prompts. Prompt specificity can itself affect controllability. We therefore re-evaluate all methods using explicit inference prompts that specify both the requested semantic focus and the content to avoid, while keeping each model checkpoint fixed. As shown in Table 2, their effect is mixed: they modestly improve several methods on Qwen2.5-VL-3B, but do not improve performance on InternVL3-2B, and zero-shot performance decreases on both backbones. Even with these engineered prompts, SFT+FoCUS remains

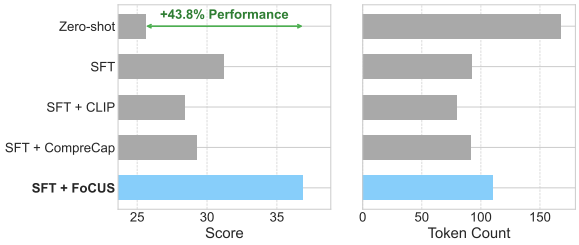


Figure 4: Token efficiency ablation: SCoPE Overall score (left) and average caption token count (right).

the best-performing method. This suggests that explicit prompt specification alone does not provide consistent semantic control, whereas the advantage of our learned control objective is robust to prompt formulation. Full category-wise results are provided in Appendix J.1.

6.2 Token Efficiency Ablation

Figure 4 compares caption length (measured in tokens) against SCoPE Overall score. Zero-shot generations are the most verbose (∼168 tokens) yet achieve the lowest SCoPE score, consistent with producing extraneous, off-scope details that degrade controllability and precision. In contrast, SFT markedly reduces output length (∼92 tokens) but tends toward excessive brevity, limiting coverage and capping the overall score.

Our framework learns a more effective length–content trade-off: it generates moderately long captions (∼110 tokens), reducing token count by approximately 34% relative to zero-shot while achieving the highest SCoPE Overall (a 43.8% improvement). This result also clarifies why simply optimizing for shorter captions (e.g., SFT+CLIP) does not guarantee improved SCoPE performance: effective controllable captioning requires allocating tokens to the *relevant* visual content rather than maximizing either verbosity or brevity.

6.3 Component-Scoring Ablation

Table 3 ablates the three key design choices in our scene reward pipeline by toggling each component during GRPO training. Starting from the baseline configuration (no thresholding, no CoT judging, and the original Llama3-8B judge), InternVL3-2B attains an overall score of 29.26. Introducing a validity threshold for object matching yields a modest but consistent gain (+0.53), indicating that stricter matching slightly reduces reward noise from spurious semantic alignments. In contrast, enabling CoT-based verification for attribute and relation scoring produces a substantially larger improvement (+4.79), supporting our claim that stronger, reasoning-based judging is crucial to prevent overly lenient rewards for generic descriptions. Replacing the judge with Qwen3-30B-A3B also yields a similarly large gain (+5.06), demonstrating that a more capable evaluator meaningfully improves the reward quality used for policy optimization.

Configuration			Overall	Δ from
Threshold	CoT	Qwen	Score	Baseline
			29.26	—
✓			29.79	+0.53 (+1.8%)
	✓		34.05	+4.79 (+16.4%)
		✓	34.32	+5.06 (+17.3%)
✓	✓		35.15	+5.89 (+20.1%)
✓		✓	34.61	+5.35 (+18.3%)
	✓	✓	36.67	+7.41 (+25.3%)
✓	✓	✓	**36.86**	**+7.60 (+26.0%)**

Table 3: Ablation of validity thresholding, CoT verification, and the Qwen judge during GRPO.

Combining components further increases performance: CoT combined with the Qwen judge reaches 36.67 (+7.41), and enabling all three achieves the best result of 36.86 (+7.60, representing a 26.0% relative improvement). Overall, the ablation confirms that our framework's main gains come from strengthening attribute and relation verification via CoT prompting and using a more capable judge model, while thresholded object matching plays a smaller but complementary role in stabilizing grounding.

6.4 Hyperparameter Analysis

Figure 5 shows the sensitivity of InternVL3-2B trained with FoCUS to two reward-computation hyperparameters: (i) Evaluation LLM used to score Attribute and Relation components, and (ii) object-matching threshold τ for the object score. We defer analysis of the negative penalty magnitude used to suppress off-scope content to Appendix J.3.

Evaluation LLM. Across a diverse set of judges, FoCUS is largely stable: changing the evaluator shifts absolute SCoPE scores moderately, with Overall scores spanning roughly 34.6–36.9. Nevertheless, evaluator capacity matters: stronger judges provide cleaner training signals and consistently improve performance. In our experiments, Gemma3-12B (Team et al., 2025a) and Qwen3-30B-A3B yield the highest Overall scores, while smaller judges (e.g., Qwen3-4B) underperform. The sensitivity is most pronounced for Relation, which varies the most across evaluators and largely

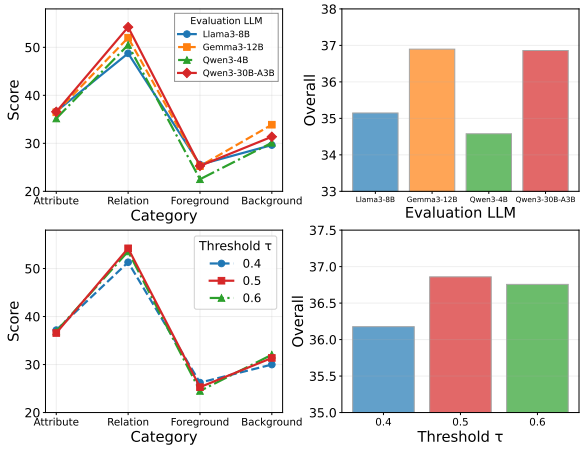


Figure 5: Hyperparameter sensitivity of FoCUS: SCoPE category-wise scores (left) and Overall (right) when varying the evaluation LLM (top) and object-matching threshold τ (bottom).

drives the Overall trend; Foreground remains the most challenging category under all judges.

Object-matching threshold τ. We vary the strictness of object matching via the validity threshold τ. Performance remains stable across a broad range of values, indicating that FoCUS does not require finely tuned matching to function well. Nonetheless, $\tau = 0.5$ achieves the best Overall result and offers the most balanced trade-off across categories. Lower thresholds slightly boost Attribute and Foreground scores but tend to hurt Relation (by admitting noisier object matches), whereas higher thresholds mildly improve Background at the expense of Foreground. These trends motivate our default configuration of Qwen3-30B-A3B with $\tau = 0.5$.

7 Conclusion

We presented FoCUS, a controllable image captioning method using a prompt-conditioned, scene-graph-aligned objective to emphasize requested content and suppress off-scope content by differentially weighting scene-graph component scores according to natural-language control prompts. We also introduced SCoPE, a benchmark with contrastive Include/Avoid constraints for evaluating semantic control. Experiments on two VLM backbones show that FoCUS improves controllability and fine-grained caption quality without requiring architectural modifications. Together, these results suggest that prompt-conditioned scene-graph control is a practical and effective direction for controllable image captioning.

Limitations

Our work has several limitations that suggest directions for future research. First, FoCUS relies on scene graph parsing and LLM-based evaluation during training, which introduces computational overhead and may propagate parsing errors into the reward signal. Second, while we demonstrate effectiveness on two VLM backbones, the scalability of our approach to significantly larger models remains to be validated. Third, the SCoPE benchmark employs LLM-based evaluation to assess Coverage, Adherence, and Faithfulness, which inherits potential biases and inconsistencies from the underlying language model judge; although we validate human alignment through an MTurk study showing reasonable correlation, LLM evaluators may still exhibit systematic blind spots or fail to capture subtle semantic distinctions that human annotators would identify.

Ethics Statement

This research adhered to the ACL Ethics Policy, utilizing only publicly available datasets and vision-language models for training and evaluation. The SCoPE benchmark was constructed using publicly available images and automated generation pipelines, with human validation conducted through MTurk following standard ethical guidelines for crowdsourced annotation. The goal of this work is to improve controllability in image captioning systems, and no negative ethical outcomes are anticipated.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)) and by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5A2A03052246). We thank the reviewers and the Action Editor for their valuable feedback.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors,

Computer Vision – ECCV 2016, volume 9909, pages 382–398. Springer International Publishing, Cham.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL Technical Report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-VL Technical Report. *Preprint*, arXiv:2502.13923.

Kalliopi Basioti, Mohamed A. Abdelsalam, Federico Fancellu, Vladimir Pavlovic, and Afsaneh Fazly. 2024. CIC-BART-SSA: Controllable Image Captioning with Structured Semantic Augmentation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXVI*, volume 15124 of *Lecture Notes in Computer Science*, pages 444–461. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. ShareGPT4V: Improving Large Multi-modal Models with Better Captions. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, volume 15075 of *Lecture Notes in Computer Science*, pages 370–387. Springer.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Intern VL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, Seattle, WA, USA. IEEE.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8299–8308, Long Beach, CA, USA. IEEE.

Google Deepmind. 2025. Gemini 3 Flash: Frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. 2020. Length-Controllable Image Captioning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12358, pages 712–729. Springer International Publishing, Cham.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and Improving Detail Image Caption. *Preprint*, arXiv:2405.19092.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing, Cham.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. 2025. Benchmarking Large Vision-Language Models via Directed Scene Graph for Comprehensive Image Captioning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19618–19627, Nashville, TN, USA. IEEE.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2023. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36, pages 72983–73007. Curran Associates, Inc.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spacy: v3.7.2: Fixes for apis and requirements.

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. 2024. DOCCI: Descriptions of Connected and Contrasting Images. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LX*, volume 15118 of *Lecture Notes in Computer Science*, pages 291–309. Springer.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting Vision and Language with Localized Narratives. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12350, pages 647–664. Springer International Publishing, Cham.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. arXiv.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. 2025. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, Honolulu, HI. IEEE.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *Preprint*, arXiv:2402.03300.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025a. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

GLM-4.5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025b. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models. *Preprint*, arXiv:2508.06471.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Boston, MA, USA. IEEE.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41

others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Lin Zhang, Xianfang Zeng, Kangcong Li, Gang Yu, and Tao Chen. 2025. SC-Captioner: Improving Image Captioning with Self-Correction by Reinforcement Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23145–23155.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *Preprint*, arXiv:2504.10479.

A Background on CompreCap and Reward Design Considerations

A.1 Overview of CompreCap

CompreCap (Lu et al., 2025) is a fine-grained evaluation framework for image captioning based on structured scene-graph analysis. Compared with holistic metrics such as CIDEr (Vedantam et al., 2015) and earlier scene-graph-based metrics such as SPICE (Anderson et al., 2016), CompreCap decomposes both the ground-truth annotations and generated captions into three semantic components: objects, attributes, and relations. The evaluation pipeline operates in two stages:

1. **Parsing:** It employs `spaCy` (Montani et al., 2023) to extract nouns (objects) and parses the caption to associate modifiers (attributes) and spatial prepositions (relations) with specific objects.

2. **Scoring:** It computes a weighted average of component scores. Object matching is performed using Sentence-BERT embeddings, while attributes and relations are evaluated using a Large Language Model (LLM) to judge semantic alignment.

Although CompreCap is effective for post-hoc evaluation, our preliminary experiments indicate that directly using its score as a reward signal introduces significant noise due to two structural limitations.

A.2 Practical Considerations When Using CompreCap as a Learning Signal

Semantic Drift in Object Scoring. CompreCap computes object scores from embedding similarity, which can assign moderate similarity to semantically related but incorrect pairs (e.g., matching `house` to `environment`). When such pairs are treated as matches, downstream attribute and relation scoring may be attached to the wrong entity, which adds noise to the learning signal. To reduce this effect, we use mutual-best matching together with a validity threshold ($\tau = 0.5$), as described in Section 3.

Leniency in Attribute and Relation Scoring. Standard zero-shot LLM judges (e.g., Llama3-8B (Grattafiori et al., 2024)) can be lenient for attribute and relation scoring, often assigning high scores to generic descriptions that omit requested specificity. For example, a caption containing only "refrigerator" may receive substantial credit against a ground-truth description such as "white fridge with side-by-side doors." As a training signal, this weakens pressure to produce precise descriptions. We therefore use Chain-of-Thought prompting with a stronger reasoning model (`Qwen3-30B-A3B-Instruct`) to obtain more discriminative attribute and relation scores.

B Object Matching Details

B.1 Scene Graph Parsing

Given a generated caption y and ground-truth scene graph annotations, we parse the caption into constituent scene graph elements. We employ the spaCy (Montani et al., 2023) parser for noun extraction and lemmatization, decomposing the generated caption into sentences and extracting candidate objects. For each identified object, we associate the relevant sub-captions that mention it, enabling object-bound attribute and relation evaluation.

B.2 Mutual-Best Matching with Validity Threshold

Let $\mathcal{C} = \{c_i\}_{i=1}^{n}$ be the candidate object nouns extracted from the generated caption, and let $\mathcal{G} = \{g_j\}_{j=1}^{m}$ be the ground-truth object categories from the annotated scene graph. We use Sentence-BERT (Reimers and Gurevych, 2019) with the all-MiniLM-L6-v2 pretrained model and compute a similarity matrix:

$$\mathbf{S} \in \mathbb{R}^{n\times m}, \quad S_{i,j} = \cos\big(\phi(c_i), \phi(g_j)\big) \tag{13}$$

where $\phi(\cdot)$ denotes the Sentence-BERT embedding.

To obtain discrete matches while preventing semantic drift, we construct a mutual-best match indicator matrix $\mathbf{S}' \in \{0,1\}^{n\times m}$:

$$S'_{i,j} = \mathbb{I}\big[j = \arg\max_{j'} S_{i,j'} \ \wedge \ i = \arg\max_{i'} S_{i',j} \ \wedge \ S_{i,j} \geq \tau\big], \tag{14}$$

where $\mathbb{I}[\cdot]$ is the indicator function and τ is a validity threshold. In our experiments, we set $\tau = 0.5$. The mutual-best constraint ensures each ground-truth object g_j matches at most one candidate c_i, so $\sum_i S'_{i,j} \in \{0,1\}$.

This formulation prevents partial credit for semantically distant matches caused by embedding-space artifacts, yielding a cleaner object-level signal for GRPO optimization.

C Prompt Category Descriptions

We use five control prompt categories for controllable caption generation:

- **General Detailed Caption** instructs the model to generate a comprehensive description covering all visual content.
- **Attribute-Focused Caption** emphasizes detailed attribute descriptions of objects (colors, textures, shapes, materials, states, etc.).
- **Relation-Focused Caption** prioritizes spatial and functional relationships between objects.
- **Foreground-Focused Caption** concentrates on main subjects and salient objects.
- **Background-Focused Caption** emphasizes contextual elements, settings, and peripheral content.

D Complete Reward Weight Configurations

For each prompt category, we use the following prompt-conditioned weights to align the control objective with the requested semantic emphasis:

General Caption.

$$R_{\text{general}} = 0.25\tilde{S}_{\text{obj}} + 0.35\tilde{S}_{\text{attr}} + 0.40\tilde{S}_{\text{rel}} \quad (15)$$

where $\tilde{S}$ denotes scores linearly scaled to $[0, 1]$.

Attribute-Focused Caption.

$$R_{\text{attr}} = 0.1 \cdot \tilde{S}_{\text{obj}} + 0.9 \cdot \tilde{S}_{\text{attr}} - 1.0 \cdot \tilde{S}_{\text{rel}} \quad (16)$$

The negative weight on relations discourages off-scope relational content while keeping the objective centered on object properties.

Relation-Focused Caption.

$$R_{\text{rel}} = 0.1 \cdot \tilde{S}_{\text{obj}} - 1.0 \cdot \tilde{S}_{\text{attr}} + 0.9 \cdot \tilde{S}_{\text{rel}} \quad (17)$$

The negative weight on attributes analogously discourages off-scope attribute content when the requested emphasis is relational.

Foreground-Focused Caption.

$$R_{\text{fg}} = \tilde{S}_{\text{fg}} - \tilde{S}_{\text{bg}} \quad (18)$$

This formulation rewards captions that emphasize main subjects while de-emphasizing background content.

Metric	Attribute	Relation	Foreground	Background	Overall
Pass rate	17.5%	69.3%	19.6%	46.0%	38.1%

Table 4: First-attempt pass rates for the Gemini-3-Flash generator–verifier loop used in SCoPE construction. Rates are computed before any regeneration.

Background-Focused Caption.

$$R_{\text{bg}} = \tilde{S}_{\text{bg}} - \tilde{S}_{\text{fg}} \quad (19)$$

E SCoPE Construction Details

E.1 Image Curation

For category-specific caption generation, we manually curated 189 images drawn from three established datasets: COCO (142 images), CompreCap (35 images), and DOCCI (12 images). All three datasets are licensed under Creative Commons Attribution 4.0. Care was taken to ensure that the selected images do not overlap with those used in the SFT and GRPO training sets.

E.2 Caption Generation and Fact Extraction

Category-specific candidate generation. We employ Gemini-3-Flash (Deepmind, 2025) as the caption generator. For each image, we define four semantic categories: Attribute, Relation, Foreground, and Background. For each category, we prompt the model to generate five diverse candidate captions with instructions to focus exclusively on the target aspect (e.g., "Describe only the foreground objects without mentioning the background context").

Automated quality assurance via self-refinement. To reduce out-of-scope content in category captions (e.g., background details appearing in a foreground-focused caption), we use Gemini-3-Flash as a judge within a generator–verifier loop with hard category-focus constraints. Each candidate is checked for strict adherence to the requested category. Captions containing out-of-scope information are rejected and regenerated for up to three attempts; if no candidate passes after exhausting this budget, the generation process for that image–category pair is restarted from scratch. We retain only verifier-approved captions in the final benchmark.

These pass rates therefore quantify the success of the iterative construction pipeline, rather than the model's single-shot controllability. Across the 756 image–category construction instances (189×4),

Statistic	Attribute	Relation	Foreground	Background	Avg. across categories
Mean #Facts ± Std	28.77 ± 6.23	16.53 ± 3.53	18.35 ± 4.33	13.73 ± 3.06	19.35 ± 2.90

Table 5: Average number of atomic facts per image in SCoPE by category. The final column reports the average over the four category-specific fact lists.

92.1% (696/756) obtained a verifier-approved caption within three attempts before a full restart was needed. The first-attempt pass rate is substantially lower and varies by category: 17.5% for Attribute, 69.3% for Relation, 19.6% for Foreground, and 46.0% for Background, or 38.1% overall (Table 4). This gap highlights that strict category adherence is achieved primarily through iterative verification and regeneration.

Atomic fact extraction and contrastive lists. We decompose verified captions into atomic facts using the same LLM. For a given category, its atomic facts form the Include list used during evaluation. We do not manually construct negative examples; instead, the Avoid list is derived from paired categories. For instance, the Include list for Attribute serves as the Avoid list when evaluating Relation, enabling direct measurement of the model's ability to emphasize target content while suppressing complementary information.

F SCoPE Statistics

F.1 Number of Atomic Facts by Category

SCoPE does not impose a fixed upper bound on the number of atomic facts per category or per image, so fact-list lengths vary across categories and images. Table 5 reports the mean number of atomic facts per category per image, together with standard deviations. Attribute-focused lists are the longest on average (28.77 ± 6.23 facts), followed by foreground (18.35 ± 4.33), relation (16.53 ± 3.53), and background (13.73 ± 3.06). Averaged across categories, SCoPE contains 19.35 ± 2.90 atomic facts per category per image. Equivalently, aggregating the four category-specific lists yields approximately 77.38 atomic facts per image on average. These statistics confirm that SCoPE typically contains dozens of atomic facts per image.

F.2 Object Frequency

We analyze the object vocabulary extracted from the curated atomic facts. Across the full benchmark, SCoPE contains 1,290 unique object types, indicating broad semantic coverage beyond a small set of recurring entities. The top-5 most frequent object types are trees (74), sky (65), building (46), grass (43), and sign (34). These statistics reflect broad object coverage and substantial diversity.

Object source	COCO	CompreCap	DOCCI
Original reference captions	48.89	83.36	74.09
SCoPE atomic facts	**83.21**	**90.32**	**84.66**

Table 6: Average Scene Coverage (%). Object mentions are grounded with OWLv2 and segmented with SAM2.1, with coverage computed from the union of object masks.

F.3 Scene Coverage

We measure how much of each image is covered by the objects mentioned in SCoPE. Following the benchmark construction in Section 4, we aggregate object mentions extracted from the curated atomic facts, ground them with OWLv2 (owlv2-large-patch14-ensemble) (Minderer et al., 2023), and segment them with SAM2.1 (sam2.1-hiera-large) (Ravi et al., 2025). We define Scene Coverage as the fraction of image area covered by the union of the resulting masks.

Table 6 shows that the object mentions in SCoPE cover most of the scene across all three source datasets. Applying the same grounding-and-segmentation pipeline to the original reference captions yields lower coverage overall, especially for COCO and DOCCI. This suggests that the category-specific captions and atomic facts used in SCoPE capture a broad portion of the visible scene, rather than only a small set of salient objects.

G Interpreting the Overall SCoPE Score

The overall SCoPE score is designed to reward balanced controllability. Coverage measures inclusion of requested content, Adherence measures suppression of complementary off-scope content, and Faithfulness measures factual consistency. By combining these three dimensions with the harmonic mean (Eq. 12), SCoPE favors captions that add relevant, grounded details while maintaining

Method	Caption	Tokens	SCoPE Overall
SFT + CompreCap	*A woman in a black sleeveless dress and large black hoop earrings stands holding a smartphone, her face illuminated by the phone's screen.*	27	15.00
SFT + FoCUS	*A young woman in a black sleeveless dress and large black hoop earrings stands intently looking at a smartphone she holds in both hands, wearing a headband and a glowing neon-purple waistband.*	39	28.57

Table 7: Example-level comparison on the foreground-focused image in Figure 9. The higher-scoring FoCUS caption uses additional tokens for on-scope, visually grounded details.

Metric	**Kendall's τ**	**p-value (τ)**	**Spearman's ρ**	**p-value (ρ)**
CompreCap	0.2048	1.23×10^{-5}	0.2291	1.10×10^{-5}
SCoPE	**0.5245**	$\mathbf{1.02\times10^{-7}}$	**0.5704**	$\mathbf{8.54\times10^{-8}}$

Table 8: Human alignment on SCoPE.

semantic focus.

This behavior is reflected in the token-length analysis in Figure 4. The shortest method, SFT+CLIP (79.74 tokens), does not achieve the best overall score, while the most verbose model, Zero-shot (167.57 tokens), performs worst. The strongest model, SFT+FoCUS, achieves the best overall result at an intermediate length (110.38 tokens), indicating that performance is driven by how effectively tokens are allocated to on-scope content rather than by brevity or verbosity alone.

Table 7 shows the same pattern on the foreground-focused example from Figure 9. Compared with SFT+CompreCap, the SFT+FoCUS caption is longer, but its additional details remain foreground-relevant and visually grounded, leading to a substantially higher SCoPE score. Together, these results suggest that higher SCoPE scores reflect more selective and semantically focused captioning.

H Human Alignment Study and Judge Robustness

H.1 Human Evaluation Protocol

We evaluate how well automatic controllability metrics align with human judgments on a subset of SCoPE. We construct 140 image–category pairs from the 35 images overlapping between SCoPE and CompreCap, spanning four control categories: *Attribute*, *Relation*, *Foreground*, and *Background*. For each image–category pair, annotators compare exactly three captions. The three captions are either the Zero-shot/SFT/SFT+FoCUS outputs from a single backbone (Qwen2.5-VL-3B-Instruct or InternVL3-2B) or the outputs of three external LVLMs (LLaVA-NeXT-34B (Liu et al., 2024), Qwen3-VL-32B-Instruct (Bai et al., 2025a), and Gemini-3-Flash (Deepmind, 2025)). We then collect human judgments on Amazon Mechanical Turk (MTurk). We restrict participation to MTurk workers with at least 10,000 approved HITs and a $\geq$98% approval rate, and we provide in-task instructions defining the evaluation criteria (Recall/-Precision/Accuracy) and the required 0–5 scoring plus forced ranking procedure (Figures 6–7). For each image–category pair, 14 independent annotators provide scores and rankings for the three captions.

H.2 Metrics Compared and Correlation Analysis

For each image–category pair, we compute a metric-induced ordering of the three candidate captions and compare it against the aggregated human preference ordering. We report rank-correlation using Kendall's τ and Spearman's ρ (with two-sided p-values). In addition to SCoPE, we also evaluate a CompreCap-based controllability score, where we assign different component weights per control category to measure controllability (i.e., different weightings for Attribute/Relation/Foreground/Background emphasis).

H.3 Human Alignment Results

Table 8 reports the correlations between automatic metrics and human judgments. SCoPE shows substantially stronger and statistically significant alignment with human preferences than the CompreCap controllability-weighting variant. In particular, SCoPE achieves Kendall's $\tau = 0.5245$ and Spearman's $\rho = 0.5704$, compared with 0.2048 and 0.2291 for CompreCap. All correlations are statistically significant ($p < 10^{-4}$). These results show that SCoPE aligns substantially better with human judgments than CompreCap in a candidate pool that includes the (Zero-shot, SFT, SFT+FoCUS) triplets for Qwen2.5-VL-3B-Instruct and InternVL3-2B, alongside LLaVA-NeXT-34B, Qwen3-VL-32B-Instruct, and Gemini-

Image Caption Evaluation Task

Instructions

You will be shown an image along with captions describing it in four different categories. For each category, you will see three different captions. Your task is to evaluate each caption based on how well it describes the image within the given category.

Evaluation Criteria

Please consider the following three aspects when evaluating each caption:

- **Recall:** How well does the caption include the content that should be covered for the given category?
- **Precision:** Does the caption avoid including unnecessary information or content that falls outside the scope of the given category?
- **Accuracy:** Is the caption factually accurate and consistent with what is actually shown in the image?

Your Tasks

- **Score each caption:** Rate each of the three captions on a scale from 0 to 5 (0 = very poor, 5 = excellent)
- **Rank the captions:** Assign a rank (1st, 2nd, 3rd) to each caption based on overall quality (1st = best)
- **Complete all 4 categories:** Evaluate Attribute, Relation, Background, and Foreground categories

⚠️ Important Ranking Rule

Each caption (A, B, or C) must be assigned to exactly ONE rank position.

- ❌ **INCORRECT:** Selecting Caption A for both 1st place AND 2nd place
- ❌ **INCORRECT:** Selecting Caption B for multiple rank positions
- ✓ **CORRECT:** Caption A = 1st place, Caption B = 2nd place, Caption C = 3rd place
- ✓ **CORRECT:** Caption C = 1st place, Caption A = 2nd place, Caption B = 3rd place

Each rank (1st, 2nd, 3rd) should have a different caption selected. Do not select the same caption multiple times.

Evaluation Image

Evaluation Image

Figure 6: MTurk task instructions shown to annotators for the Image Caption Evaluation Task, including the definitions of Recall, Precision, and Accuracy, and the required workflow (rate each caption from 0 to 5 and then rank the three captions with a unique 1st/2nd/3rd choice).

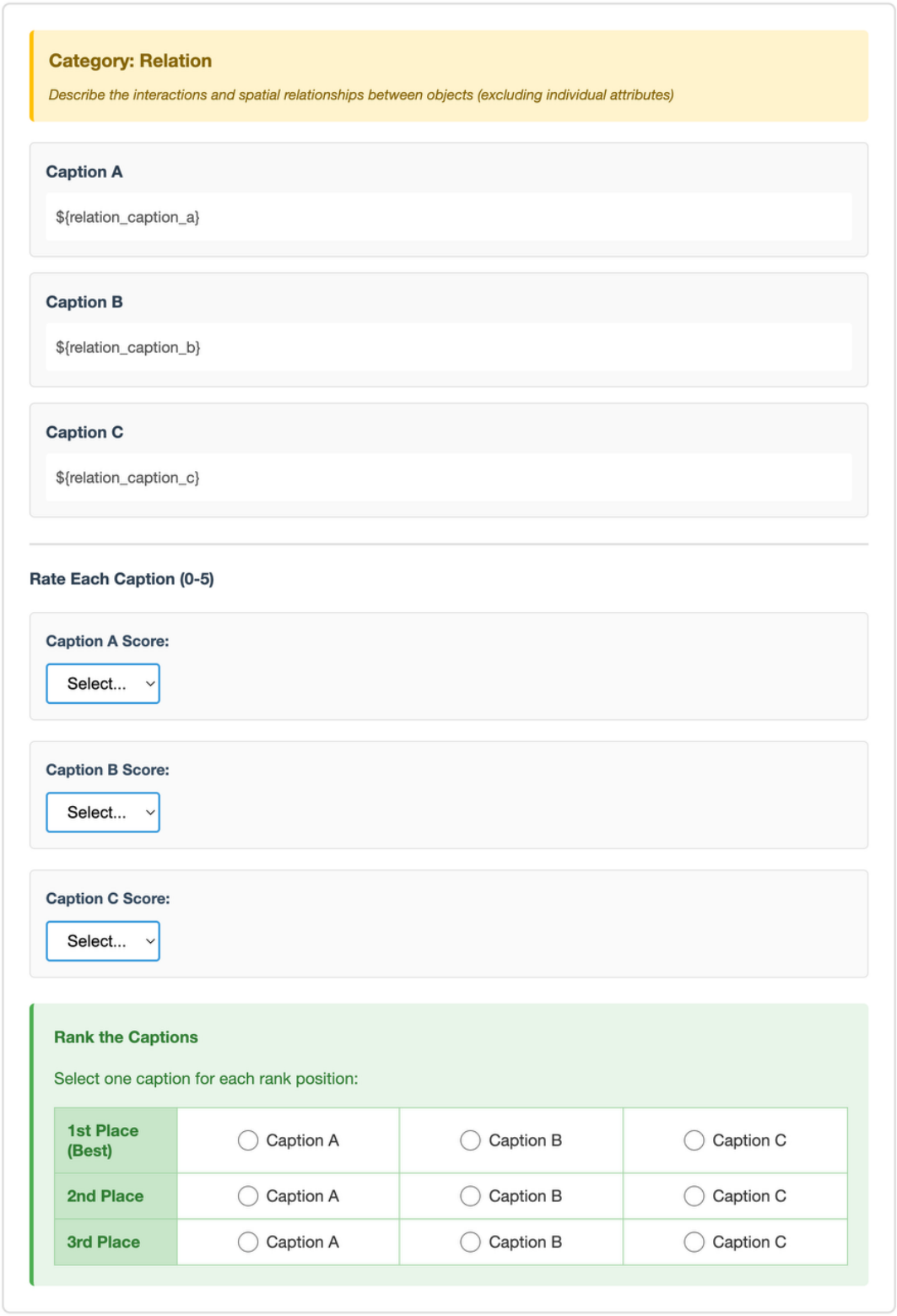


Figure 7: MTurk annotation interface for one category (Relation shown): annotators view three candidate captions (A/B/C), assign each a 0–5 score, and provide a forced ranking (1st/2nd/3rd) with no ties. The category header and guidance text change depending on whether the question targets Attribute, Relation, Foreground, or Background.

Model Name	Kendall's τ	p-value (τ)	Spearman's ρ	p-value (ρ)
Qwen3-32B-Thinking	**0.5245**	$\mathbf{1.02\times10^{-7}}$	**0.5704**	$\mathbf{8.54\times10^{-8}}$
Qwen3-4B-Thinking	0.5123	1.48×10^{-5}	0.5634	1.43×10^{-5}
Qwen3-30B-A3B-Instruct	0.4755	1.76×10^{-7}	0.5218	1.19×10^{-7}
GLM-Flash	0.4979	8.70×10^{-7}	0.5523	5.60×10^{-7}

Table 9: Judge–human alignment on the MTurk preference set.

Evaluation Judge	Method	SCoPE				
		Attribute	Relation	Foreground	Background	Overall
Qwen3-30B-A3B-Instruct	Zero-shot	16.68	61.92	33.23	39.74	37.89
	SFT	47.45	59.52	30.77	49.64	46.85
	SFT+CLIP	41.36	54.74	33.71	39.92	42.43
	SFT+CompreCap	48.63	44.53	29.00	52.27	43.61
	SFT+FoCUS	**50.30** (+33.62)	**66.10** (+4.18)	**37.88** (+4.65)	**54.08** (+14.34)	**52.09** (+14.20)
Qwen3-4B-Thinking	Zero-shot	7.66	47.58	20.28	24.00	24.88
	SFT	29.42	42.42	17.39	27.17	29.10
	SFT+CLIP	24.62	39.14	19.53	21.55	26.21
	SFT+CompreCap	32.20	30.04	17.40	29.18	27.21
	SFT+FoCUS	**33.02** (+25.36)	**50.82** (+3.24)	**22.29** (+2.01)	**30.23** (+6.23)	**34.09** (+9.21)
GLM-Flash	Zero-shot	10.44	41.41	20.19	24.32	24.09
	SFT	32.78	43.17	23.72	34.51	33.54
	SFT+CLIP	38.33	55.10	29.85	38.56	40.46
	SFT+CompreCap	43.22	40.12	28.83	47.57	39.94
	SFT+FoCUS	**43.35** (+32.91)	**58.86** (+17.45)	**34.02** (+13.83)	**48.58** (+24.26)	**46.20** (+22.11)

Table 10: Fixed-judge SCoPE rescoring on InternVL3-2B under three evaluation judges. Parentheses denote gains over Zero-shot.

3-Flash. This supports our claim that the contrastive Include/Avoid formulation better reflects the controllability judgments made by human evaluators than CompreCap-style component reweighting.

H.4 Robustness to Judge Choice

Because both SCoPE and our training reward rely on LLM judges, we additionally test whether the conclusions in the main text depend on a particular evaluator. We study two questions: (i) whether the human alignment reported in Table 8 is stable across judge families and model sizes, and (ii) whether the gains of SFT+FoCUS persist when SCoPE is rescored with fixed weaker or backbone-mismatched judges.

Judge–human alignment across families and sizes. Using the same MTurk preference set described above, we recompute rank correlation with three additional judges beyond the default Qwen3-32B-Thinking: a much smaller same-family model (Qwen3-4B-Thinking-2507), the exact training-time judge used in our reward pipeline (Qwen3-30B-A3B-Instruct-2507), and an unseen backbone (GLM-Flash (Team et al., 2025b)). Table 9 shows that all four judges remain strongly and statistically significantly correlated with human preferences, with only modest variation across evaluator family and size. Pairwise agreement among judges is also consistently high (Spearman's ρ between 0.64 and 0.73), indicating that SCoPE is not narrowly tied to one evaluator's preferences.

Choice of the default SCoPE judge. These results motivate our use of Qwen3-32B-Thinking as the default SCoPE evaluator. Among all judges tested, it achieves the highest agreement with MTurk preferences, making it the strongest empirical proxy for human judgments on SCoPE's Include/Avoid and contradiction-verification tasks. This evaluator is also distinct from the training-time reward judge, Qwen3-30B-A3B-Instruct-2507, reducing concern that evaluation simply reuses the optimization signal. The fixed-judge rescoring below further verifies that the main conclusions do not depend on this default choice.

Fixed-judge rescoring. We next rescore the InternVL3-2B results from Table 1 using three fixed evaluators: the training-time reward judge (Qwen3-30B-A3B-Instruct-2507), a substantially smaller same-family judge (Qwen3-4B-Thinking-2507), and a backbone-mismatched judge (GLM-Flash). As shown in Table 10, SFT+FoCUS achieves the best SCoPE Overall score under all three evaluators. Thus, the gains reported in the main text are not explained by matching a single judge's style; they persist under weaker and architecturally different judges, supporting the conclusion that FoCUS improves semantic controllability itself.

I Additional Experimental and Training Details

I.1 SFT and GRPO Data Construction

I.1.1 SFT data construction

For SFT, we randomly sample 2,000 images from the COCO 2017 training set (Lin et al., 2014). For each image, we use GPT-5 (Singh et al., 2026) to generate five category-specific captions aligned with our control prompt categories (General, Attribute, Relation, Foreground, Background). The generated captions are manually filtered for quality and category alignment, yielding 10,000 image–caption pairs in total.

I.1.2 GRPO data construction (CompreCap split and FG/BG labels)

For GRPO, we use CompreCap, which provides dense human-annotated scene graphs (objects, segmentation masks, bound attributes, directional relations). From 560 instances, we randomly sample 280 for training and reserve the remaining 280 for evaluation (no overlap). To support foreground/background-focused control, we augment CompreCap objects with *Foreground* vs. *Background* labels using GPT-5; location descriptors derived from visual cues are consistently categorized as background. These foreground/background classifications are subsequently manually filtered.

I.2 Implementation Details

We implement GRPO using the Hugging Face `trl` library (von Werra et al., 2020) on 8 NVIDIA RTX6000 Ada 48GB GPUs. For GRPO, we use 8 generations per prompt, gradient accumulation steps of 8, KL coefficient $\beta = 0.04$, and learning rate 2×10^{-6}. During training, we use a multi-prompt sampling strategy to balance optimization across the five prompt categories.

I.3 Random Seed Stability

To assess the stability of GRPO training, we repeated SFT+FoCUS with two additional random seeds for each backbone and report mean $\pm$ standard deviation over three runs. Overall, the results are stable across seeds on both SCoPE and the CompreCap evaluation: the standard deviation of the overall score remains small in all settings—below 0.5 for both backbones on both benchmarks.

We observe slightly larger variation in some harder fine-grained categories, particularly foreground/background-focused captioning, but these fluctuations are modest and do not affect the main conclusions of the paper. In particular, FoCUS shows consistent performance across runs for both Qwen2.5-VL-3B and InternVL3-2B, indicating that the reported improvements are not driven by a favorable single seed. Full per-seed results are provided in Table 11.

I.4 SCoPE Robustness to Source Composition and Sampling

Because SCoPE contains images from multiple source datasets, we examine whether the observed gains are concentrated on images originating from CompreCap, which is also used for GRPO training. Of the 189 SCoPE images, 35 originate from CompreCap and the remaining 154 from COCO and DOCCI; none overlap with the SFT or GRPO training images. We report Overall scores separately for these two source groups. To quantify uncertainty due to the finite evaluation set, we additionally compute 95% nonparametric image-bootstrap confidence intervals on the full SCoPE set using 20,000 repetitions.

As shown in Table 12, SFT+FoCUS achieves the best Overall score on both source subsets for both backbones. In particular, its performance is similar between CompreCap-origin and non-CompreCap images, despite GRPO training using CompreCap annotations. On the full benchmark, its 95% bootstrap interval is also separated from that of the strongest non-FoCUS baseline for each backbone. Together with the seed-stability results above, these results show that the observed gains are not concentrated in a particular SCoPE source subset or a favorable training run.

Base Model	Seed	SCOPE				
		Attribute	Relation	Foreground	Background	Overall
Qwen2.5-VL-3B	0	30.12	45.46	24.09	25.95	31.40
	1	30.31	46.12	23.54	26.76	31.68
	2	30.89	45.06	25.52	27.02	32.12
	Mean ± std	**30.44 ± 0.40**	**45.55 ± 0.54**	**24.38 ± 1.02**	**26.58 ± 0.56**	**31.74 ± 0.36**
InternVL3-2B	0	36.18	54.39	25.18	31.68	36.86
	1	36.39	54.35	25.46	29.97	36.54
	2	37.13	53.88	25.23	32.47	37.18
	Mean ± std	**36.57 ± 0.50**	**54.21 ± 0.28**	**25.29 ± 0.15**	**31.37 ± 1.28**	**36.86 ± 0.32**

		CompreCap					
		General	Attribute	Relation	Foreground	Background	Overall
Qwen2.5-VL-3B	0	46.86	44.29	54.16	55.26	50.88	50.29
	1	48.00	45.45	55.15	56.22	51.35	51.23
	2	47.68	45.40	54.73	55.50	51.38	50.94
	Mean ± std	**47.51 ± 0.59**	**45.05 ± 0.65**	**54.68 ± 0.50**	**55.66 ± 0.50**	**51.20 ± 0.28**	**50.82 ± 0.48**
InternVL3-2B	0	49.23	46.97	56.57	55.81	51.41	52.00
	1	49.78	47.69	55.59	55.55	52.37	52.20
	2	49.14	47.10	56.99	54.77	51.23	51.85
	Mean ± std	**49.38 ± 0.35**	**47.25 ± 0.38**	**56.38 ± 0.72**	**55.38 ± 0.54**	**51.67 ± 0.61**	**52.02 ± 0.18**

Table 11: Random seed analysis for Qwen2.5-VL-3B-Instruct and InternVL3-2B with SFT+FoCUS. We report results from three independent GRPO runs and the mean ± standard deviation for both SCOPE and CompreCap metrics.

Base Model	Method	SCOPE Overall		
		CompreCap-Origin	Non-CompreCap	All [95% CI]
Qwen2.5-VL-3B	Zero-shot	13.74	16.01	15.66 [14.21, 17.04]
	SFT	17.32	23.81	22.61 [20.61, 24.46]
	SFT+CLIP	27.90	26.03	26.40 [25.18, 27.62]
	SFT+CompreCap	25.81	25.86	25.85 [24.65, 27.06]
	SFT+FoCUS	**31.74**	**31.75**	**31.74 [30.55, 32.88]**
InternVL3-2B	Zero-shot	24.31	25.93	25.63 [24.60, 26.63]
	SFT	32.05	31.01	31.21 [29.97, 32.42]
	SFT+CLIP	27.94	28.49	28.39 [27.18, 29.57]
	SFT+CompreCap	30.44	28.99	29.26 [28.18, 30.31]
	SFT+FoCUS	**37.90**	**36.62**	**36.86 [35.79, 37.91]**

Table 12: Source-stratified evaluation on SCOPE. We report Overall scores separately for CompreCap-origin images and images originating from the other source datasets, together with 95% nonparametric image-bootstrap confidence intervals on the full benchmark using 20,000 repetitions.

I.5 Evaluation Variant on CompreCap

In addition to SCOPE, we evaluate on the held-out 280 CompreCap instances by reweighting CompreCap components for different control emphases and scaling scores to 0–100 for interpretability. To address the limitations noted in the main text, we apply a stricter object validity threshold ($\tau = 0.6$) and upgrade the scoring engine to `Qwen3-32B-Thinking` (Yang et al., 2025) for more rigorous reasoning-based assessment.

When performing category-specific reweighting for this *CompreCap-based evaluation variant*, we use a milder negative weight of -0.25 (instead of -1.0) on the off-target component (e.g., relations in attribute-focused scoring, and attributes in relation-focused scoring).

I.6 General Caption Evaluation

To assess whether improved controllability comes at the expense of general caption quality, we addi-

Base Model	Method	General Captioning on DOCCI				
		CIDEr	METEOR	ROUGE-L	CAPTURE	CompreCap (General)
Qwen2.5-VL-3B	Zero-shot	0.0040	0.1273	0.1694	0.5627	0.4162
	SFT+FoCUS	0.0879	0.1319	0.2062	0.5981	0.4751
InternVL3-2B	Zero-shot	0.1110	0.1415	0.2196	0.5989	0.3999
	SFT+FoCUS	0.0908	0.1490	0.2118	0.5990	0.4938

Table 13: General-caption evaluation on 5,000 images from the DOCCI test split. We report standard reference-based metrics together with CAPTURE and general-prompt CompreCap. All scores are scaled to $[0, 1]$.

tionally evaluate captions generated with the General prompt on 5,000 images from the DOCCI test split (Onoe et al., 2024). Using the human-written DOCCI captions as references, we report standard reference-based metrics—CIDEr (Vedantam et al., 2015), METEOR (Denkowski and Lavie, 2014), and ROUGE-L (Lin, 2004)—together with CAPTURE (Dong et al., 2024) and the general-prompt CompreCap (Lu et al., 2025) score. All scores are reported on a $[0, 1]$ scale.

As shown in Table 13, SFT+FoCUS improves all reported metrics for Qwen2.5-VL-3B. For InternVL3-2B, METEOR and general-prompt CompreCap improve, CAPTURE remains essentially unchanged, and CIDEr and ROUGE-L decrease slightly. Overall, FoCUS largely maintains general-caption quality across the two backbones while consistently improving fine-grained factual alignment under the general prompt.

I.7 Baselines (Expanded)

- **Zero-shot:** the pretrained backbone with only task-specific prompts (no parameter updates).
- **SFT:** cross-entropy fine-tuning on the COCO-based category-specific caption dataset (Appendix I.1).
- **SFT+CLIP:** after SFT, we add CLIP-based focus indicators to specify the target emphasis, then run GRPO using CLIP similarity as reward.
- **SFT+CompreCap:** GRPO using the original CompreCap scoring pipeline as the training reward, without the stricter object-matching threshold and reasoning-based verification for attribute and relation scoring used in FoCUS.

I.8 LLM Judges Used in Training vs. Evaluation

For the CompreCap-based evaluation variant, we use `Qwen3-32B-Thinking` (Yang et al., 2025) as the scoring engine for more rigorous reasoning-based assessment.

J Additional Controllability Analyses

J.1 Prompt-Engineering Ablation on SCoPE

We additionally evaluate all methods using stronger inference prompts that explicitly specify both desired and off-scope content. For example, the background-focused prompt instructs the model to describe only background elements and to exclude foreground subjects, body parts, actions, and interactions; analogous constraints are used for attribute-, relation-, and foreground-focused captioning. This ablation changes only the inference prompt and leaves the underlying model parameters unchanged.

As shown in Table 14, SFT+FoCUS achieves the highest Overall score on both backbones and performs best across all four control categories. These results complement the main evaluation by showing that prompt-conditioned training provides gains even when the desired behavior is specified with explicit inference-time instructions.

J.2 Ablating Prompt-Conditioned and Contrastive Reward Weighting

We further ablate the reward-weighting design introduced in Sec. 3 and evaluate on SCoPE (Sec. 4). To isolate the effect of weighting, for each backbone we keep the training and evaluation setup fixed and vary only the reward aggregation. We compare FoCUS against two alternatives: **Fixed Reward Weights**, which applies the same general-caption mixture to every prompt category, and **Simply Up-weighted**, which keeps category-specific target emphasis but removes negative weights on non-target components. Results are shown in Table 15.

Across both Qwen2.5-VL-3B and InternVL3-2B, both alternatives underperform FoCUS in

Base Model	Method	SCoPE				
		Attribute	Relation	Foreground	Background	Overall
Qwen2.5-VL-3B	Zero-shot	5.43	16.08	13.66	16.04	12.80
	SFT	20.97	33.72	19.94	18.25	23.22
	SFT+CLIP	22.15	40.31	25.35	21.69	27.37
	SFT+CompreCap	27.68	31.33	23.63	26.59	27.31
	SFT+FoCUS	**28.36** (+22.93)	**45.66** (+29.58)	**29.10** (+15.44)	**27.12** (+11.08)	**32.56** (+19.76)
InternVL3-2B	Zero-shot	9.02	28.32	21.64	22.87	20.46
	SFT	28.05	44.93	23.90	27.14	31.01
	SFT+CLIP	19.64	41.36	24.74	21.85	26.90
	SFT+CompreCap	29.78	32.83	22.15	29.22	28.50
	SFT+FoCUS	**32.26** (+23.24)	**50.79** (+22.47)	**29.50** (+7.86)	**29.95** (+7.08)	**35.62** (+15.16)

Table 14: SCoPE results under engineered prompts for Qwen2.5-VL-3B-Instruct and InternVL3-2B. Parentheses denote gains over Zero-shot.

Base Model	Method	SCoPE				
		Attribute	Relation	Foreground	Background	Overall
Qwen2.5-VL-3B	Fixed Reward Weights	9.40	36.77	**28.38**	18.13	23.17
	Simply Up-weighted	9.42	33.92	27.07	18.97	22.35
	FoCUS	**30.44**	**45.55**	24.38	**26.58**	**31.74**
InternVL3-2B	Fixed Reward Weights	12.68	40.96	**33.66**	27.09	28.60
	Simply Up-weighted	11.58	40.07	31.49	26.20	27.33
	FoCUS	**36.57**	**54.21**	25.29	**31.37**	**36.86**

Table 15: Ablation of reward-weighting design on SCoPE.

Overall score and show a less balanced category profile. The ablated variants can obtain higher Foreground scores than FoCUS, but they substantially lag on Attribute, Relation, and Background. This suggests that using either a prompt-agnostic mixture or target-only positive weighting tends to favor salient foreground or generic scene content across prompt types. With Fixed Reward Weights, the same aggregation is used for every prompt category; with Simply Up-weighted, target components receive more weight but off-scope components are not explicitly suppressed. Under SCoPE's Include/Avoid evaluation, these design choices lead to weaker cross-category controllability.

By contrast, FoCUS achieves the strongest Overall score on both backbones and the best Attribute, Relation, and Background scores, supporting the role of prompt-conditioned signed weights in jointly encouraging on-scope content and suppressing off-scope content.

J.3 Penalty Magnitude Sensitivity

Figure 8 shows how the magnitude of the negative reward penalty shapes behavior. As the penalty increases, behavior can shift sharply: very large penalties may produce refusal-like outputs, while removing the penalty (0.0) increases scope drift and invites off-scope background details. This suggests using a bounded, calibrated penalty rather than extreme values; accordingly, we adopt a moderate default that reduces off-scope content without

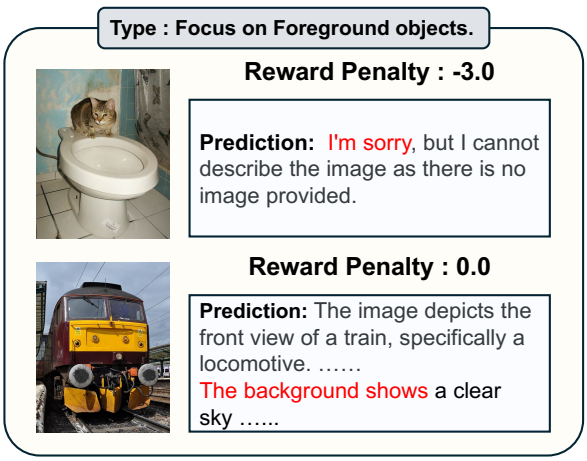


Figure 8: Penalty magnitude ablation for foreground-focused rewards. Large negative penalties (e.g., -3.0) can trigger refusal-like captions, while no penalty (0.0) causes scope drift by adding extra background details.

Base Model	Method	SCoPE				
		Attribute	Relation	Foreground	Background	Overall
GPT-5	nano	17.93	63.84	39.21	43.82	41.20
	mini	25.34	70.43	45.54	50.57	47.97
	-	**27.55**	**71.60**	**45.59**	**54.21**	**49.74**
Gemini 2.5	Flash Lite	18.52	60.15	41.96	36.73	39.34
	Flash	25.02	66.89	53.75	48.95	48.65
	Pro	**30.25**	**70.64**	**60.40**	**62.15**	**55.86**

		CompreCap					
		General	Attribute	Relation	Foreground	Background	Overall
GPT-5	nano	45.31	**40.49**	54.48	50.88	49.45	48.12
	mini	**47.96**	39.89	**59.27**	**51.40**	51.81	**50.07**
	-	45.68	40.15	56.95	49.45	**53.27**	49.10
Gemini 2.5	Flash Lite	45.32	39.23	52.91	51.33	47.11	47.18
	Flash	**52.14**	**43.43**	**58.30**	51.35	**51.91**	**51.42**
	Pro	51.11	42.13	56.65	**53.75**	51.33	50.99

Table 16: Performance of commercial LVLMs on SCoPE and CompreCap evaluation protocols. Best results within each model family are shown in **bold**.

encouraging refusals.

K SCoPE and CompreCap Benchmark for Commercial Models

Table 16 reports results for representative commercial LVLMs on both evaluation protocols used in the main paper: SCoPE for contrastive controllability via Include/Avoid constraints and CompreCap for component-wise factual alignment.

SCoPE: Controllability under Include/Avoid constraints. Across model families, larger variants achieve higher controllability scores. For GPT-5, performance increases from nano (Overall 41.20) to mini (47.97) and peaks at the largest setting (Overall 49.74), with particularly strong Relation control (up to 71.60). Gemini 2.5 exhibits a similar scaling trend and attains the best overall SCoPE performance among the tested commercial models: Flash Lite (39.34) $<$ Flash (48.65) $<$ Pro (55.86). Gemini 2.5 Pro also leads category-wise, especially on Foreground (60.40) and Background (62.15), indicating stronger spatial emphasis and out-of-scope suppression in the contrastive setting.

CompreCap: Fine-grained factual alignment. On CompreCap, scores exhibit lower variance across model variants than on SCoPE, suggesting that component-wise factual alignment is less sensitive to model size than strict controllability. For GPT-5, the best overall CompreCap score is achieved by mini (50.07), while nano performs best on Attribute (40.49) and the largest setting leads on Background (53.27). For Gemini 2.5, Flash achieves the best overall CompreCap score (51.42) and leads on General (52.14), Attribute (43.43), Relation (58.30), and Background (51.91), while Pro performs best on Foreground (53.75).

Connection to main findings. These results contextualize the gains of FoCUS: SCoPE imposes stricter, contrastive control requirements than component-wise evaluation alone, and models that perform well on CompreCap do not always dominate on SCoPE. This gap motivates our prompt-conditioned reward weighting (including negative weights) and the emphasis on suppressing off-scope content, which is explicitly measured by SCoPE but only indirectly reflected in CompreCap.

L Prompt List

This appendix lists the prompts used in FoCUS and for SCoPE evaluation. The Attribute Evaluation Prompt and Relation Evaluation Prompt are used during training to compute component-level scene rewards. The SCoPE inference prompts

are used to generate captions under each semantic control type. The Coverage-and-Adherence Evaluation Prompt and Faithfulness Evaluation Prompt are used by the LLM judge during SCoPE evaluation.

L.1 Training Reward Judge Prompts

Attribute Evaluation Prompt

System Prompt:
You are a strict evaluation judge. You will be given a 'Sentence' (prediction) and a 'Phrase' (ground truth description). Your task is to rate how accurately and completely the Sentence covers the detailed attributes (e.g., color, shape, material, design) described in the Phrase.

Rate on a scale from 0 to 5:
0: Irrelevant or contradictory.
1: Mentions the correct object but misses all specific attributes.
2: Captures the object and very minor details.
3: Captures the object and some key attributes.
4: Captures the object and most attributes.
5: Perfectly captures the object and all attributes described in the Phrase.

Explain the reason and then output the final score within <score></score>.

Input Template:
Sentence: {sentence}. Phrase: {phrase}.

Relation Evaluation Prompt

System Prompt:
You are a strict evaluation judge. You will be given a 'Sentence' (prediction) and a 'Phrase' (ground truth spatial relationship). Your task is to rate how accurately the Sentence describes the spatial relationship or interaction defined in the Phrase.

Rate on a scale from 0 to 5:
0: Relationship is missing, incorrect, or unrelated.
1: Mentions the objects but the relationship is wrong.
2: Mentions the objects and only weakly implies the relationship.
3: Captures the relationship vaguely or partially.
4: Correctly describes the relationship, but misses minor nuances or lacks full precision.
5: Perfectly captures the spatial relationship or interaction.

Explain the reason and then output the final score within <score></score>.

Input Template:
Sentence: {sentence}. Phrase: {phrase}.

L.2 SCoPE Inference Prompts

SCoPE Inference Prompts

Attribute: "Describe the image in detail, focusing on the attributes and characteristics of the objects."

Relation: "Describe the image in detail, focusing on the spatial relationships between objects."

Foreground: "Describe the image in detail, focusing on the foreground subject."

Background: "Describe the image in detail, focusing on the background elements and surrounding environment."

L.3 SCoPE LLM-Judge Prompts

Coverage-and-Adherence Evaluation Prompt

Prompt:
Does this caption mention the following?

Caption: "{caption}"
Mention: "{fact}"

Think carefully and answer Yes or No.

Faithfulness Evaluation Prompt

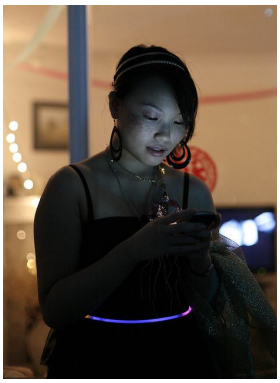
Prompt:
Does the caption describe something DIFFERENTLY than the fact?
(e.g., if fact says "red ball" but caption says "blue ball", that's a contradiction)

Caption: "{caption}"
Fact: "{fact}"

Think carefully and answer Yes or No.

M Qualitative Examples of FoCUS

Figure 9 shows a qualitative comparison between zero-shot and FoCUS generations across different control categories.



	Zero-shot	FoCUS
Attribute	The image shows a person standing indoors, likely in a dimly lit room. They are wearing a black dress with thin straps and a headband. The person has large hoop earrings and is holding a smartphone, seemingly focused on it. They are also wearing a necklace and a sparkly, possibly sequined, jacket draped over their shoulders. The background includes string lights and a television screen, suggesting a cozy, festive atmosphere. The person's expression is engaged, possibly looking at the phone screen.	A woman with dark brown hair pulled back and light skin shows small freckles on her face and wears a black sleeveless dress with thin straps and a smooth, matte fabric; she has large black circular hoop earrings and a thin gold-colored chain necklace. A thin silver headband with small shiny beads is visible. A dark, sparkly jacket with a metallic sheen and a narrow strap of sequined material is present. A glowing neon ring with a blue-purple light is visible. A small rectangular phone with a dark screen and reflective surface is visible. A vertical blue-painted metal post has a smooth painted finish. A string of small white bulbs produces small circular light spots. A framed picture has a dark mat and a light-colored frame. A television screen displays bright white and blue rectangular images. The wall is painted pale pink with a red circular printed motif.
Relation	The image shows a person standing indoors, likely in a dimly lit room. The individual is focused on a smartphone they are holding in their hands. They are wearing a black dress with thin straps and a headband. Around their neck, they have a necklace and large hoop earrings. A glowing neon ring is visible around their waist, adding a colorful accent to their outfit.the background, there is a wall decorated with string lights, creating a warm ambiance. A framed picture hangs on the wall, and a television screen is visible, displaying some content. The room has a cozy, festive feel, possibly decorated for an event or celebration.	A woman stands in the foreground with a vertical support pole behind her head, a string of lights to her left and a framed picture on the wall behind and slightly to her left; a television sits behind and to her right, while a jacket draped over her shoulder hangs beside her right side and a glowing neon band encircles her waist between her torso and hands, with her hands positioned on a small rectangular device held in front of her face.
Foreground	The image shows a person standing indoors, likely in a dimly lit room. The individual is focused on a smartphone they are holding in their hands. They are wearing a black dress with thin straps and a headband. Around their neck, they have a necklace and large hoop earrings. A glowing neon ring is visible around their waist, adding a colorful accent to their outfit. In the background, there is a wall decorated with string lights, creating a warm ambiance. A framed picture hangs on the wall, and a television screen is visible, displaying some content. The room has a cozy, festive feel, possibly decorated for an event or celebration.	A woman in a black sleeveless dress and large black hoop earrings stands holding a smartphone, her face illuminated by the phone's screen while a glowing purple ring accentuates her waist.
Background	The image shows a person standing indoors, likely in a dimly lit room. The individual is wearing a black dress and has large hoop earrings. They are focused on a smartphone, holding it with both hands. The person has a headband and is accessorized with a necklace and a bracelet. In the background, there are string lights creating a warm ambiance. A framed picture hangs on the wall, and a television is visible, displaying some content. The wall also features a decorative element with a red circular design. The overall setting suggests a casual, intimate environment, possibly a party or social gathering.	An indoor residential space with a painted wall featuring a circular red emblem and a framed picture, a blue-painted vertical support post, a string of small round fairy lights draped along the wall, a television screen showing bright images on a nearby surface, and a glimpse of a countertop and countertop lights, all illuminated by warm artificial lighting that casts soft shadows across the room.

Figure 9: Zero-shot vs. FoCUS qualitative example. Rows correspond to control categories (Attribute, Relation, Foreground, Background). Red highlights out-of-category content that should be avoided under the given control signal. Compared to zero-shot, FoCUS better concentrates on the requested semantics while suppressing off-scope details.