CAPEVAL: A DECOUPLED CAPTION EVALUATION ACROSS UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Captions serve as a primary supervision signal for both multimodal understanding and text-to-image generation. However, previous evaluations treat the caption quality as a single scalar objective, which conflates two distinct properties: (1) how much visual information a caption covers and (2) how reliably the image supports its stated claims. To this end, we design a *decoupled* caption evaluation benchmark, **CAPEval** (**C**overage **A**nd **P**recision **Eval**uation), with human-written ground-truth captions and human-verified atomic checklist items. Specifically, CAPEval decomposes caption quality into *Coverage* C and *Precision* P. The former quantifies how thoroughly a caption covers ground-truth factual content, while the latter reflects the factual correctness rate of all claims expressed in the caption. We select 10 captioners and further conduct controlled downstream *end-to-end* experiments with them from four model families, where the caption source is the only variable. Empirically, we find a consistent *task-dependent* dissociation: Coverage serves as the stronger correlate for understanding performance, whereas Precision acts as the dominant predictor for generation performance. This decoupled evaluation paradigm not only delivers a more fine-grained diagnosis of caption quality, but also offers actionable guidance for selecting and optimizing captioners tailored to different downstream tasks.

1 INTRODUCTION

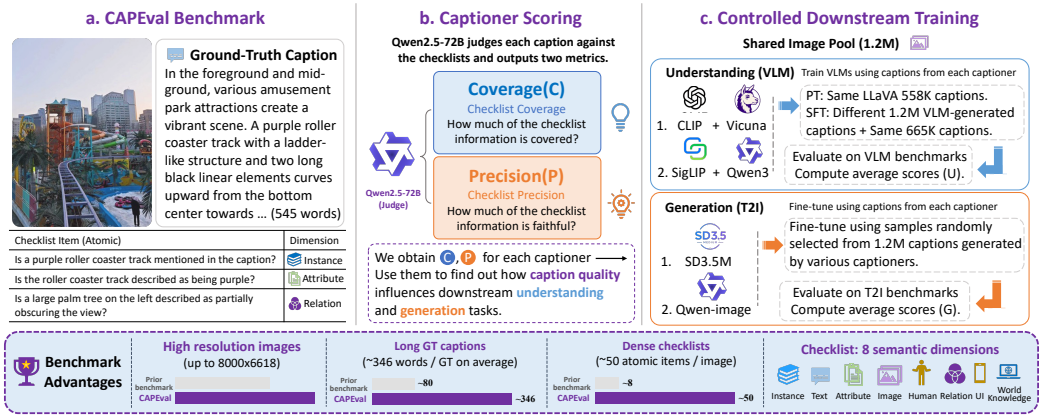

Figure 1: Overview of CAPEval. **(a)** CAPEval constructs fine-grained semantic checklists from high-resolution (up to 8K) images, long-form ground-truth captions, and dense atomic checklist items spanning eight semantic dimensions. **(b)** Each candidate caption is evaluated against the atomic checklist items by a judge model to obtain *Coverage* (C) and *Precision* (P). **(c)** Captions produced by different captioners are then used in controlled VLM training and T2I fine-tuning experiments, where the caption source is the only varying factor, yielding aggregate understanding score U and generation score G for each captioner.

To achieve comprehensive understanding and precise text-instruction following capabilities, large-scale image caption corpora (Schuhmann et al., 2022; Gadre et al., 2023; Zhang et al., 2025; Dong et al., 2025; Wei et al., 2025; Farina et al., 2026; Li et al., 2026) are widely used for training *both*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

vision-language models (VLMs) (Radford et al., 2021; Jia et al., 2021; Alayrac et al., 2022; Li et al., 2023b; Wang et al., 2025a;b; Lei et al., 2025; Wang et al., 2026a;b;c; Liu et al., 2026b; Zhang et al., 2026) and text-to-image (T2I) generators (Rombach et al., 2022; Wu et al., 2025; Team et al., 2025). Under such a setting, textual descriptions, *i.e.*, captions, serve as the primary supervision signal for aligning visual content with language. Recent studies reveal that caption quality has become an explicit data variable that can substantially affect downstream multimodal capabilities (Chen et al., 2024a; Zhang et al., 2025; Yang et al., 2026a).

However, how caption-quality properties affect multimodal learning remains insufficiently explored. More specifically, we ask *whether downstream multimodal models benefit more from captions that cover a broader range of visual content (i.e., coverage) or from captions that make fewer but more reliable claims (i.e., precision), and whether the conclusion differs between understanding and generation.* To systematically evaluate these questions, we design **CAPEval** (**C**overage **A**nd **P**recision **Eval**uation), based on the following two principles:

1. **Decoupled Evaluation of Coverage and Precision.** Both conventional single-scalar captioning metrics (Papineni et al., 2002; Vedantam et al., 2015; Hessel et al., 2021) and recent enhanced evaluation methods (Gao et al., 2026; Liu et al., 2026a) conflate coverage and reliability of descriptions into one unified score without disentanglement. We, on the contrary, aim to explicitly *decompose* caption quality into Coverage and Precision to systematically explore how the relative importance may *differ* across downstream scenarios.
2. **End-to-End Captioner Evaluation.** Prior benchmarks (Onoe et al., 2024; Dong et al., 2024) evaluate captions as standalone outputs, leaving unclear how caption quality propagates into downstream tasks. We, instead, adopt a fully *end-to-end* protocol: *each captioner directly generates training captions*, with the caption source as the only variable across identical pipelines. This design links intrinsic caption quality (Coverage and Precision) to downstream VLM and T2I performance, converting captioner evaluation into a direct measure of training data utility.

Concretely, CAPEval operates as a four-stage pipeline: (1) dense atomic checklists are constructed from selected images and long-form human-annotated ground-truth (GT) captions as shown in Figure 1a; (2) per-captioner Coverage and Precision scores are computed by a judge model as shown in Figure 1b; (3) captions from each captioner serve as the sole varying supervision signal in controlled VLM pretraining and T2I fine-tuning, yielding an aggregate understanding score U and generation score G per captioner as shown in Figure 1c; and (4) CAPEval fits ordinary least squares (OLS) regressions U or $G = \beta_0 + \beta_C C + \beta_P P$ across captioners, and systematically examines the *significance* (p-values) and *magnitude* (β_C and β_P) to characterize how Coverage and Precision contribute to downstream performance, as shown in Figure 2. Our experiments reveal three key findings:

1. **Caption quality profile matters more than model scale.** Smaller captioners can produce more useful supervision when their caption properties better align with the downstream objective. For instance, within the InternVL3.5 family, the smallest model (1B) achieves a higher understanding score than the 8B model (avg. $U = 58.5$ vs. 57.3) and also exhibits higher Coverage ($C = 48.3$ vs. 46.5). Similarly, the 4B model outperforms its larger 8B counterpart on generation (avg. $G = 71.6$ vs. 71.0) while attaining higher Precision ($P = 73.5$ vs. 72.6). These within-family observations suggest that understanding performance may be more closely associated with Coverage, whereas generation performance may be more closely associated with Precision. We next examine whether these relationships hold more broadly across captioners through systematic regression analysis.
2. **Understanding is dominated by Coverage.** For understanding, Coverage is the stronger predictor ($\hat{\beta}_C = +0.118$, p-value $= 0.026$ on SigLIP-Qwen3; $\hat{\beta}_C = +0.215$, p-value $= 0.047$ on CLIP-Vicuna), while Precision is not statistically significant in either pipeline (both p-values ≥ 0.274), indicating that broad semantic coverage is the dominant factor. We further find that unlike general understanding, hallucination performance is dominated by Precision (p-value $= 0.124$), while Coverage is less significant (p-value $= 0.914$).
3. **Generation is driven by Precision.** Across ten captioners, Precision is the only statistically significant predictor of generation quality ($\hat{\beta}_P = +0.189$, p-value < 0.001 on SD3.5M; $\hat{\beta}_P = +0.235$, p-value < 0.001 on Qwen-Image), while Coverage is not statisti-

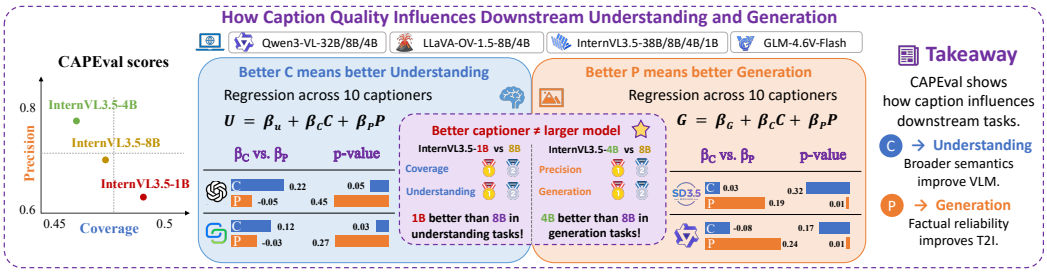

Figure 2: Relating CAPEval scores to downstream performance reveals a task-dependent asymmetry: *broader semantic Coverage drives VLM understanding*, whereas *higher factual Precision drives T2I generation*. This pattern is confirmed by both the InternVL3.5 case study and regression analysis across ten captioners in Section 3.

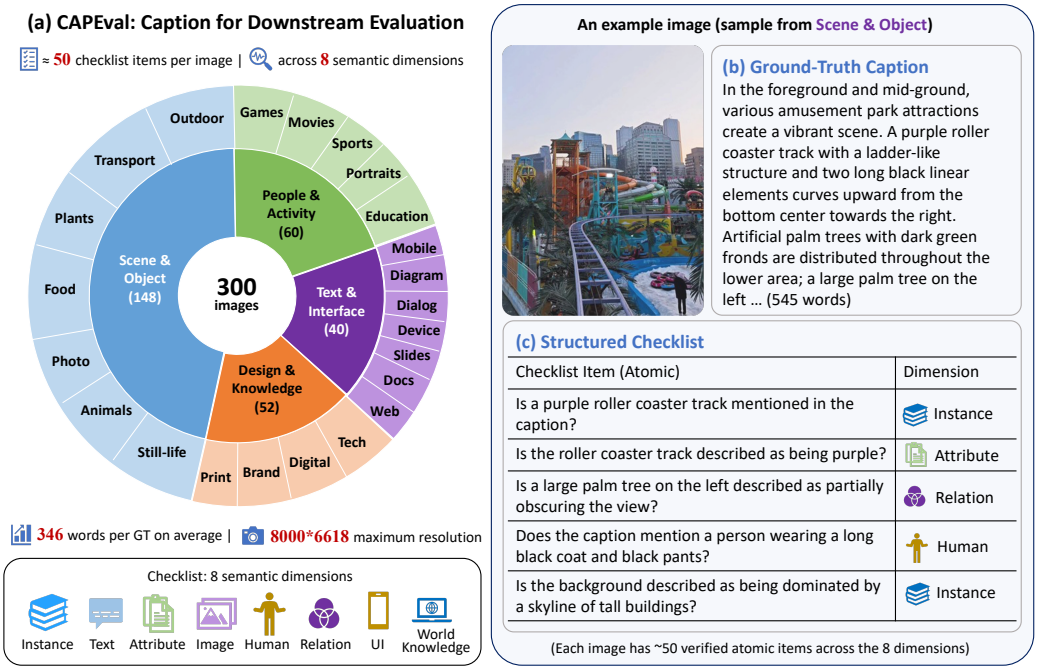

Figure 3: CAPEval image collection and annotation design. **(a)** Hierarchical taxonomy and distribution of 300 images across four domains and subcategories. **(b)** Example ground-truth caption (545 words) for a sample image. **(c)** Corresponding structured semantic checklist, showing atomic items and their assigned dimensions. The complete information of GT caption and checklists for this example is provided in Appendix B.1.

cally significant in either pipeline (both p-values ≥ 0.171), indicating that factual precision is the dominant factor for text-to-image generation.

These results reveal a task-relative view of caption quality: the dimension that matters most is determined by the downstream objective rather than by captioner scale or aggregate quality score. CAPEval provides a practical foundation for task-aware caption data curation: understanding models should be trained on captions with broad semantic coverage, while generation models benefit from captions with high factual precision.

2 CAPEval Design

2.1 Overview

We introduce **CAPEval** (**C**overage **A**nd **P**recision **Eval**uation), a unified framework for systematically connecting caption quality with downstream multimodal capability. CAPEval is grounded

Table 1: Per-dimension statistics of checklist annotations in CAPEval, including the total number of verified atomic items, average number of items per image, and image coverage for each semantic dimension.

Semantic Dimension	Total Count	Avg. per Image	Images Covered
Instance	2,681	8.94	300 (100.0%)
Attribute	3,660	12.20	300 (100.0%)
Relation	2,632	8.77	300 (100.0%)
Image	2,074	6.91	300 (100.0%)
Text	1,800	6.00	217 (72.3%)
Human	783	2.61	165 (55.0%)
UI	584	1.95	123 (41.0%)
World Knowledge	751	2.50	218 (72.7%)
Total	**14,965**	**49.88**	**300 (100.0%)**

in structured semantic checklists: each image is paired with a human-generated GT caption that is decomposed into atomic checklist items spanning eight dimensions (Figure 1a). Given a candidate caption, Qwen2.5-72B (Hui et al., 2024) judges each checklist item as *correct*, *wrong*, or *not mentioned*, based solely on the caption text without access to the image (Figure 1b). Based on these checklist annotations, CAPEval decomposes caption quality into two dimensions:

- **Coverage** C: the fraction of GT checklist items that are mentioned by the candidate caption, measuring semantic completeness.
- **Precision** P: the fraction of mentioned checklist items that are factually correct, measuring factual reliability.

Rather than evaluating captions in isolation, CAPEval connects intrinsic caption quality with downstream performance through controlled training experiments. As illustrated in Figure 1c, captions generated by each captioner serve as the sole varying supervision signal across two VLM pretraining pipelines and two T2I fine-tuning pipelines. By jointly analyzing Coverage, Precision, and downstream performance, CAPEval establishes an explicit bridge between caption evaluation and multimodal performance, enabling systematic identification of which caption quality dimension is most predictive for each downstream objective, as summarized in Figure 2.

2.2 Benchmark Construction

Image Collection. We select 300 images covering four visually diverse domains, each further divided into several subcategories, as shown in Figure 3. Images are sourced from publicly available web sources and real-world photographs, spanning a wide range of resolutions and aspect ratios. As shown in Figure 4a, CAPEval's maximum image resolution exceeds that of prior benchmarks (Lee et al., 2024; Cheng et al., 2025; Agrawal et al., 2019; Sidorov et al., 2020; Young et al., 2014; Petryk et al., 2024; Lin et al., 2014; Yang et al., 2026b; Guan et al., 2024; Fang et al., 2026).

GT Captions. Each image is paired with a GT caption written by a human annotator following explicit guidelines: captions must cover all present elements across five dimensions (visual subjects, in-image text verbatim, aesthetic and photographic attributes, portrait attributes, and special image properties), describe only objectively visible content without speculation or repetition, and pass review by a second annotator before finalization. These guidelines produce captions that are substantially longer and more factually dense than those in prior benchmarks, as shown in Figure 4b.

Semantic Checklists. Each GT caption is decomposed into a structured checklist of atomic statements spanning eight semantic dimensions. All items are manually verified for factual grounding and correct dimension assignment. As shown in Figure 4c, CAPEval provides dense checklists, with 14,965 verified atomic facts across 300 images and an average of 49.88 items per image. Instance, attribute, relation, and image-level dimensions are covered across all images, while the remaining dimensions capture specialized visual information when applicable. This annotation structure enables comprehensive evaluation beyond object recognition, including fine-grained attributes, interactions, textual content, and contextual knowledge. Table 1 summarizes the checklist statistics.

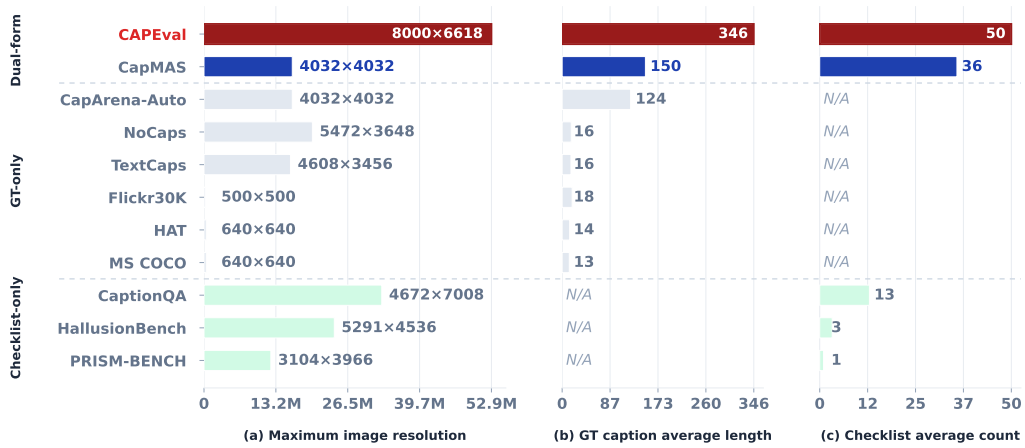

Figure 4: Comparison of CAPEval with existing benchmarks across **(a)** maximum image resolution (Additional details are provided in Appendix B.2), **(b)** average GT caption length, and **(c)** average checklist items per image. Benchmarks are grouped into Dual-form, GT-only, and Checklist-only. CAPEval leads on all three axes.

2.3 CAPEval Metrics

CAPEval decomposes caption quality into two axes: *Coverage* (C), measuring the fraction of GT facts the caption attempts to mention, and *Precision* (P), measuring the fraction of mentioned facts that are correct. For each image, the caption generated by each captioner is compared against the GT checklist using Qwen2.5-72B as the judge. Every checklist item is labeled as *yes*, *no*, or *not mentioned*, indicating whether the item is correctly described, incorrectly described, or omitted by the caption. Let N_{yes}, N_{no} and N_{nm} denote the numbers of *yes*, *no* and *not mentioned* items. Then Coverage is computed by

$$C = \frac{100 \times (N_{\text{yes}} + N_{\text{no}})}{N_{\text{yes}} + N_{\text{no}} + N_{\text{nm}}}. \tag{1}$$

Precision is computed by

$$P = \frac{100 \times N_{\text{yes}}}{N_{\text{yes}} + N_{\text{no}}}. \tag{2}$$

2.4 Controlled Training and Evaluation

To study how caption quality affects downstream vision-language understanding and text-to-image generation, we conduct controlled training experiments in which the caption source is the only variable. We evaluate ten captioning models from four model families: InternVL3.5 with 1B, 4B, 8B, and 38B variants (Wang et al., 2025c); Qwen3-VL with 4B, 8B, and 32B variants (Bai et al., 2025); LLaVA-OneVision-1.5 with 4B and 8B variants (An et al., 2025); and GLM-4.6V-Flash (Hong et al., 2025). For each captioner, we construct a corresponding training corpus by replacing only the caption supervision. This design allows performance differences to be attributed primarily to differences in caption quality rather than confounding factors in data composition or training configuration.

Vision-Language Understanding. We fine-tune two VLM pipelines to find out how caption quality influences vision-language understanding. The first pipeline uses CLIP-ViT-Large-Patch14-336 (Radford et al., 2021) as the visual encoder and Vicuna-7B-v1.5 (Touvron et al., 2023) as the language model. The second pipeline uses SigLIP-SO400M-Patch14-384 (Zhai et al., 2023) as the visual encoder and Qwen3-4B (Yang et al., 2025) as the language model. For both pipelines, we follow the standard two-stage training procedure. In Stage 1, we train the visual-language projector on the same LLaVA 558K pretraining set (Liu et al., 2023) with fixed captions, so this stage introduces no variation. In Stage 2, we vary only the caption supervision while following the ShareGPT4V data composition, sampling 1.2M images from COCO (Lin et al., 2014), SAM (Kirillov et al., 2023), and LLaVA/LCS (Liu et al., 2023). For each caption source, the generated captions are combined with the same 665K instruction-following mixture used in LLaVA-1.5 (Liu et al., 2024a). Across all

runs, we keep the model architecture, image pool, optimization settings, and training schedule fixed; only the caption source for the 1.2M image pool is changed. This controlled design isolates the downstream effect of caption-quality properties on VLM performance.

We evaluate the resulting VLMs on 16 benchmarks covering four capability dimensions: (1) general multimodal understanding on MME (Fu et al., 2026), MMBench (EN and CN) (Liu et al., 2024b), SEED-Bench (Li et al., 2023a), and MMMU (Yue et al., 2024); (2) multimodal reasoning on ScienceQA (Lu et al., 2022), AI2D (Kembhavi et al., 2016), and RealWorldQA (AI, 2024); (3) visual perception on MMStar (Chen et al., 2024b), MMVP (Tong et al., 2024b), CV-Bench (2D and 3D) (Tong et al., 2024a), and OCRBench (Liu et al., 2024c); and (4) hallucination robustness on POPE (Li et al., 2023c), HallusionBench (Guan et al., 2024), and AMBER (Wang et al., 2023).

Text-to-Image Generation. We fine-tune two diffusion-based T2I models to test whether caption-quality effects also transfer to image generation. Specifically, we fine-tune Stable Diffusion 3.5 Medium (SD3.5M) (Esser et al., 2024) on 50K images and Qwen-Image (Wu et al., 2025) on 100K images. The two training subsets are sampled from the same 1.2M image pool and the captions are generated by each captioner. Following the same controlled-variable principle, we construct one training set for each caption source by replacing only the caption supervision while keeping the sampled images unchanged. For each T2I backbone, all runs use the same model architecture, image subset, optimization settings, and training schedule. Thus, the only variable is the caption source which allows us to evaluate whether different caption-quality properties affect image generation.

We evaluate the fine-tuned T2I models on three benchmarks targeting complementary aspects of generation quality: (1) compositional generation on T2I-CompBench++ (Huang et al., 2025), covering attribute binding, spatial relations, and complex compositions; (2) object-level accuracy on GenEval (Ghosh et al., 2023), covering object presence, counting, and color; and (3) dense prompt following on DPG-Bench (Hu et al., 2024).

3 Experiments

3.1 CAPEval Metrics Better Explain Downstream Utility than Scale

CAPEval Reveals Decoupled Coverage and Precision Profiles. Figure 5 reports the Coverage and Precision scores of different captioners on CAPEval and reveals substantial variation along both dimensions. The proprietary models (Comanici et al., 2025; DeepMind, 2026a;b; OpenAI, 2026) generally attain high scores on both dimensions. Among the evaluated open-source captioners, Coverage ranges from 41.2 to 68.5, while Precision ranges from 45.1 to 86.1, indicating that captioners differ both in the breadth of visual information they describe and in the reliability with which that information is grounded in the image.

Importantly, Coverage and Precision exhibit *no* monotonic relationship across captioners. While some captioners achieve high scores on both dimensions, others show divergent quality profiles: some provide broader descriptions with lower factual reliability, *e.g.*, InternVL3.5-1B, whereas others produce more conservative captions with higher precision, *e.g.*, InternVL3.5-4B and InternVL3.5-8B. This divergence suggests that captioners *differ* in how much visual information they attempt to capture and how reliably they ground that information, leading to distinct trade-offs between semantic breadth and factual accuracy.

Downstream evaluation. To examine how these distinct quality profiles relate to downstream utility, we use the captions generated by each captioner as supervision for two understanding pipelines and two generation pipelines. The *Avg.* columns in Table 2 further average the two pipeline-level scores within each downstream objective. Raw benchmark scores are provided in Appendix C.2.

A counterintuitive phenomenon within InternVL3.5. We first consider InternVL3.5-8B/4B/1B as a controlled comparison within a single model family. A natural expectation is that increasing captioner scale would produce more useful supervision and therefore stronger downstream models. However, Table 2 shows that downstream performance does not improve monotonically with captioner scale. For understanding, the ranking is 1B (58.5) $>$ 4B (57.5) $>$ 8B (57.3), whereas for generation, the ranking is 4B (71.6) $>$ 8B (71.0) $>$ 1B (70.1). Thus, the largest captioner may *not* always be the most useful source of supervision for either downstream objective.

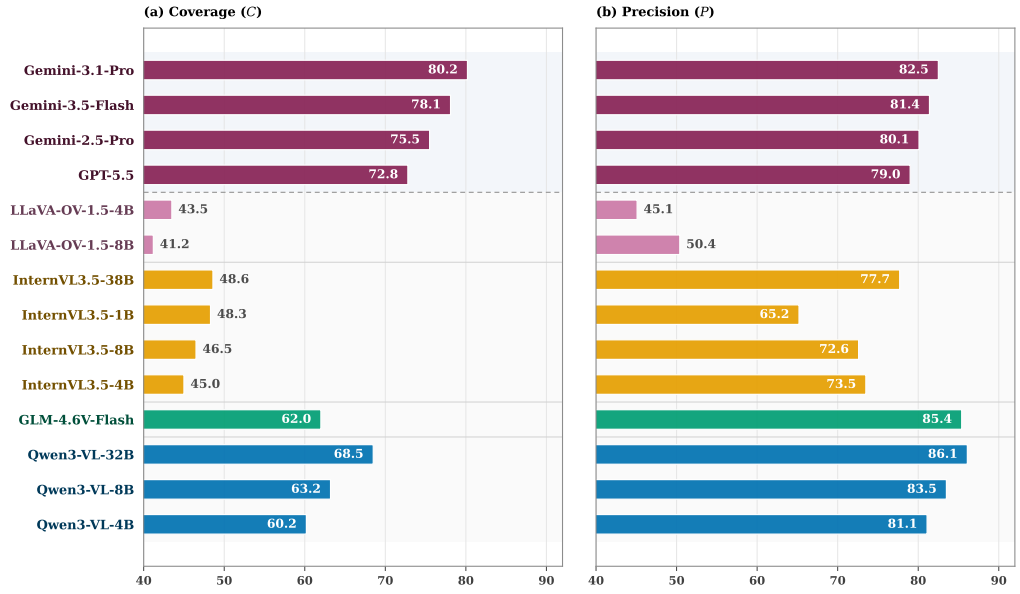

Figure 5: Coverage (C) and Precision (P) scores of multimodal models on CAPEval. The dashed horizontal line separates proprietary models from open-source models. Captioners exhibit diverse and non-monotonically aligned Coverage–Precision profiles, showing that the two dimensions capture distinct properties of caption quality.

Table 2: CAPEval Coverage (C) and Precision (P) together with aggregate downstream scores. Shaded rows correspond to the InternVL3.5 family members used for the within-family comparison. The results show that captioner scale alone does not determine downstream utility, while the preferred caption quality profile differs between understanding and generation. Detailed performance for each understanding and generation benchmark is provided in Appendix C.2.

Captioner	**CAPEval**		**Understanding**			**Generation**		
	C	P	SigLIP	CLIP	Avg.	SD3.5	QwenImg	Avg.
Qwen3-VL-32B	68.5	86.1	66.8	53.6	60.2	70.2	76.7	73.5
Qwen3-VL-8B	63.2	83.5	66.6	55.1	60.9	69.7	76.4	73.1
Qwen3-VL-4B	60.2	81.1	67.5	54.0	60.8	69.2	76.1	72.7
GLM-4.6V-Flash	62.0	85.4	66.8	55.1	61.0	69.3	75.5	72.4
InternVL3.5-38B	48.6	77.7	65.0	52.5	58.8	68.0	75.5	71.8
InternVL3.5-8B	46.5	72.6	65.2	49.3	57.3	66.5	75.4	71.0
InternVL3.5-4B	45.0	**73.5**	65.2	49.7	57.5	67.7	75.4	**71.6**
InternVL3.5-1B	**48.3**	65.2	64.2	52.7	**58.5**	65.9	74.3	70.1
LLaVA-OV-8B	41.2	50.4	65.5	52.9	59.2	62.1	69.0	65.6
LLaVA-OV-4B	43.5	45.1	66.0	50.6	58.3	61.9	68.9	65.4

These downstream rankings exhibit different alignments with C and P. Among the three models, InternVL3.5-1B has the highest Coverage ($C = 48.3$) and also achieves the strongest understanding performance. InternVL3.5-4B, by contrast, has the highest Precision ($P = 73.5$) and achieves the strongest generation performance. We next explore whether this pattern extends beyond a single model family through systematic regression analysis across all evaluated captioners.

3.2 Systematic Regression Analysis Across Captioners

Setup and interpretation. To examine whether the pattern observed within InternVL3.5 generalizes across captioners, we fit separate ordinary least squares (OLS) regressions for each of the four downstream pipelines. For captioner i and downstream pipeline t, we model

$$U_{i,t} \text{ or } G_{i,t} = \beta_{0,t} + \beta_{C,t} C_i + \beta_{P,t} P_i + \epsilon_{i,t}, \tag{3}$$

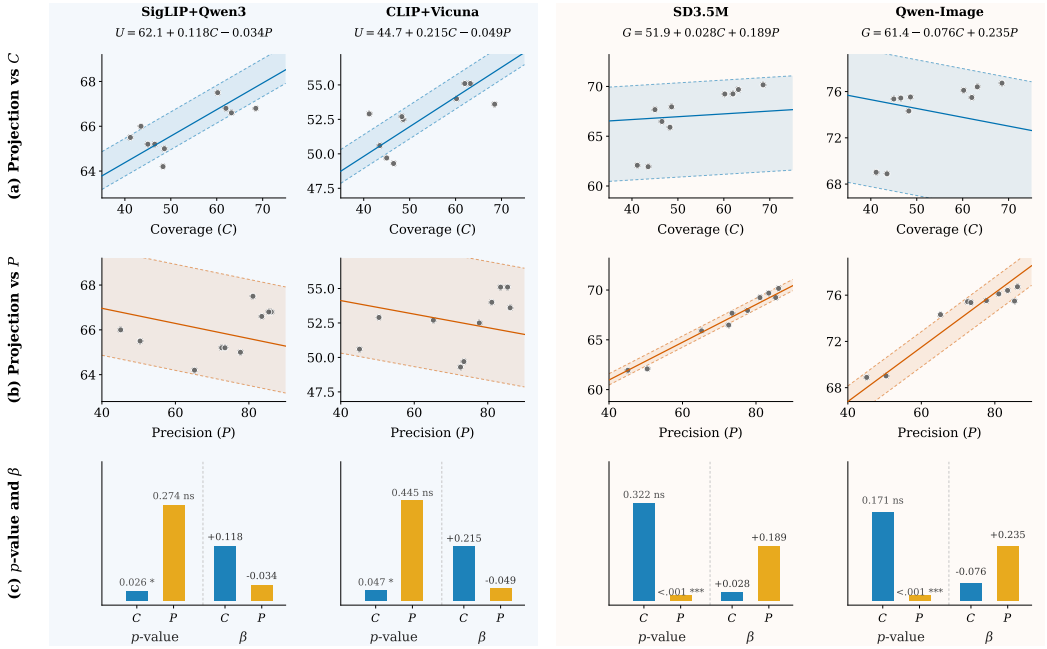

Figure 6: Regression analysis across four downstream pipelines. Row labels show the fitted equation for each pipeline. **(a)** Partial projection onto Coverage, with Precision fixed at its mean; the shaded region spans predictions obtained at the observed $P_{\min}$ and $P_{\max}$. **(b)** Partial projection onto Precision, with Coverage fixed at its mean; the shaded region analogously spans the observed $C_{\min}$ and $C_{\max}$. **(c)** OLS coefficient estimates and coefficient-level p-values for Coverage and Precision. For understanding, Coverage has a positive and significant coefficient in both SigLIP-Qwen3 and CLIP-Vicuna, whereas Precision is less significant. For generation, Precision is positive and significant in both SD3.5M and Qwen-Image, while Coverage is not.

where C_i and P_i denote the CAPEval Coverage and Precision scores of captioner i, respectively, and $U_{i,t}$ or $G_{i,t}$ denotes its downstream score for pipeline t. Each regression therefore estimates the association between each CAPEval dimension and downstream performance.

To jointly characterize the *strength* and *statistical reliability* of the link between caption quality and downstream performance, we center our analysis on two core statistics from the fitted models:

- **The fitted coefficients $\hat{\beta}_{C,t}$ and $\hat{\beta}_{P,t}$** denote the estimated change in downstream score corresponding to a one-unit increase in Coverage and Precision, respectively, for task t, which reflects the marginal contribution of each quality dimension to downstream utility.
- **The p-value for each coefficient** is computed under the null hypothesis $H_0 : \beta = 0$, and quantifies the statistical evidence that the observed conditional linear association differs from zero. A smaller p-value indicates higher confidence that the correlation is statistically meaningful, rather than arising from random variation.

Figure 6 visualizes the fitted models: (a) and (b) show partial projections onto Coverage and Precision, respectively, while (c) reports the estimated coefficients and coefficient-level p-values.

Understanding favors broader semantic coverage. Across both understanding pipelines, Coverage is the statistically significant predictor of downstream performance, indicating that captions with broader semantic coverage provide more effective supervision for VLM understanding. Specifically,

- For SigLIP-Qwen3, $\hat{\beta}_C = +0.118$ with p-value $= 0.026$, while the effect of Precision is not statistically significant (p-value $= 0.274$).
- For CLIP-Vicuna, $\hat{\beta}_C = +0.215$ with p-value $= 0.047$, while the effect of Precision is not statistically significant (p-value $= 0.445$).

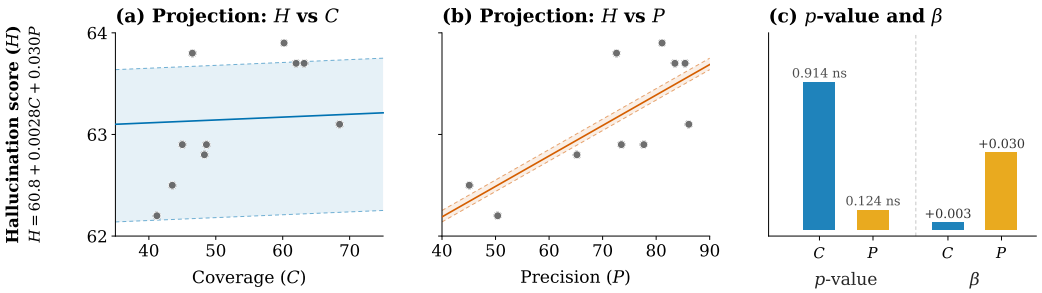

Figure 7: Regression analysis on *hallucination* score shows the fitted equation for the average score of CLIP-Vicuna and SigLIP-Qwen3. For hallucination, Precision is more significant than Coverage.

However, this pattern does not extend uniformly to hallucination performance. We further regress the average score of the two understanding pipelines across HallusionBench, POPE, and AMBER on Coverage (C) and Precision (P). As shown in Figure 7, Precision exhibits a stronger association with hallucination performance (p-value $= 0.124$) than Coverage (p-value $= 0.914$). Thus, while general understanding primarily benefits from broader semantic Coverage, hallucination robustness is more closely associated with caption Precision.

Generation favors factual precision. Across both generation pipelines, Precision is a positive and highly significant predictor of downstream performance, whereas Coverage provides no statistically significant additional explanatory power once Precision is controlled for. Specifically,

- For SD3.5M, $\hat{\beta}_P = +0.189$ with p-value < 0.001, while the effect of Coverage is not statistically significant (p-value $= 0.322$).
- For Qwen-Image, $\hat{\beta}_P = +0.235$ with p-value < 0.001, while the effect of Coverage is not statistically significant (p-value $= 0.171$).

Caption utility is downstream-objective dependent. Taken together, the regressions reveal a clear asymmetry which is consistent with the InternVL3.5 case study. For understanding, higher Coverage is associated with better performance, while Precision carries a weaker but negative association after controlling for Coverage. For generation, the relationship reverses: Precision is the consistent predictor, whereas Coverage contributes little once caption reliability is accounted for.

This agreement between the within-family comparison and the cross-captioner analysis provides converging evidence that caption quality cannot be reduced to a single scalar notion. Instead, different downstream objectives favor different caption quality profiles: *understanding benefits primarily from broader semantic Coverage, whereas generation depends more strongly on factual Precision.* CAPEval makes this distinction explicit by decoupling these two properties, thereby providing a more informative characterization of caption utility.

4 Conclusion

In this paper, we introduced CAPEval, a framework that decomposes caption quality into two distinct axes: (1) Coverage, measuring how much of the visual content a caption attempts to describe, and (2) Precision, measuring how reliably those described facts are correct. Moreover, CAPEval connects both axes to downstream multimodal performance through controlled training experiments. By keeping the image pool, model architecture, and training configuration fixed while varying only the caption source across ten captioners, CAPEval isolates the causal contribution of each caption quality dimension to vision-language understanding and text-to-image generation. Our experiments reveal a consistent and task-dependent dissociation between the two axes. For text-to-image generation, Precision is the only statistically significant predictor. For vision-language understanding, the pattern reverses: Coverage is the significant predictor. This *dissociation* surfaces a counterintuitive practical consequence: a smaller captioner can outperform a larger one when its Coverage–Precision profile better matches the target objective. Captioner scale alone is therefore an insufficient proxy for downstream utility. We provide a task-aware view of caption quality and offer practical guidance for caption generation, filtering, and large-scale multimodal dataset construction: moving beyond single-score caption evaluation toward a principled, objective-driven framework.

AI USE STATEMENT

We used GPT-5.6 Sol to edit the manuscript for readability, grammar, and linguistic clarity. The tool was not used to generate research ideas, methodologies, mathematical claims or proofs, experimental results. All AI-assisted edits were reviewed and approved by the authors to ensure that they preserved the intended technical meaning. The authors take full responsibility for the final content of this work, including all text edited with the assistance of generative AI.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8947–8956. IEEE, 2019.

X AI. Grok-1.5 vision preview. *https://x. ai/blog/grok-1.5 v*, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Google DeepMind. Gemini 3.1 pro model card. Technical report, Google DeepMind, 2026a. URL https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-1-pro. Accessed: 2026-07-26.

Google DeepMind. Gemini 3.5 flash model card. Technical report, Google DeepMind, 2026b. URL https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-5-flash. Accessed: 2026-07-26.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33272–33293, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 63157–63186, 2026.

Matteo Farina, Vishaal Udandarao, Thao Nguyen, Selim Kuzucu, Maximilian Böther, Andreas Hochlehnert, Adhiraj Ghosh, Marianna Nezhurina, Karsten Roth, Joschka Struber, et al. Datacomp-vlm: Improved open datasets for vision-language models. *arXiv preprint arXiv:2606.28551*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Zixian Gao, Atsushi Hashimoto, and Kuniaki Saito. Gavel: Grounded caption error verification and localization. *arXiv preprint arXiv:2606.26923*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

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

Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20758–20769. IEEE, 2025.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023b.

Tianle Li, Xuyang Shen, Yan Ma, Rongxin Guo, Shaoxiang Chen, Jiacheng Chen, Haochen Wang, Hongyang Tang, Yucong Zhou, and Yu Cheng. Claimdiff-rl: Fine-grained caption reinforcement learning through visual claim comparison. *arXiv preprint arXiv:2605.20278*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Jinyu Liu, Xincheng Shuai, Henghui Ding, and Yu-Gang Jiang. Unison: Benchmarking unified multimodal models via synergistic understanding and generation. *arXiv preprint arXiv:2606.26984*, 2026a.

Weisong Liu, Haochen Wang, Kuan Gao, Yuhao Wang, Yikang Zhou, Zhongwei Ren, Jacky Mai, Anna Wang, Yanwei Li, Jason Li, et al. Motionatlas: Detailed region captioning for motion-centric videos. *arXiv preprint arXiv:2606.29531*, 2026b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024c.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pp. 291–309. Springer, 2024.

OpenAI. Gpt-5.5 system card. Technical report, OpenAI, 2026. URL https://openai.com/index/gpt-5-5-system-card/. Accessed: 2026-07-26.

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

Suzanne Petryk, David M Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E Gonzalez, and Trevor Darrell. Aloha: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 342–357, 2024.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pp. 742–758. Springer, 2020.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024b.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9275–9286. IEEE, 2025a.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *International Conference on Learning Representations*, volume 2025, pp. 14374–14399, 2025b.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, volume 2026, pp. 148769–148794, 2026a.

Haochen Wang, Yuhao Wang, Tao Zhang, Yikang Zhou, Yanwei Li, Jiacong Wang, Ye Tian, Jiahao Meng, Zilong Huang, Guangcan Mai, et al. Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 2222–2250, 2026b.

Jiacong Wang, Zijian Kang, Haochen Wang, Xiao Liang, Ya Wang, Jiawen Li, Bohong Wu, Jiao Ran, Haiyong Jiang, Chao Feng, et al. Vgr: Visual grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 89506–89528, 2026c.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Yana Wei, Hongbo Peng, Yanlin Lai, Liang Zhao, Kangheng Lin, En Yu, Keyu Lv, Han Zhou, Yin Tang, Haodong Li, et al. Perceptionrubrics: Calibrating multimodal evaluation to human perception. *arXiv preprint arXiv:2606.28322*, 2026.

Zhixiang Wei, Guangting Wang, Xiaoxiao Ma, Ke Mei, Huaian Chen, Yi Jin, and Fengyun Rao. Hq-clip: Leveraging large vision-language models to create high-quality image-text datasets and clip models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22447–22456. IEEE, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Penghui Yang, Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Yibin Wang, Yujie Zhou, Jiazi Bu, Jianze Liang, Qidong Huang, et al. Caprl++: Unified reinforcement learning with verifiable rewards for dense image and video captioning. *arXiv preprint arXiv:2606.09393*, 2026a.

Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. Captionqa: Is your caption as useful as the image itself? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23741–23750, 2026b.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Xin Zhang, Haochen Wang, Yikang Zhou, Zhuochen Wang, Xiangtai Li, and Robby T Tan. Actor as its own critic: Unifying region understanding and localization via cyclegrpo. In *European Conference on Computer Vision*, pp. 586–604. Springer, 2026.

Xinsong Zhang, Yarong Zeng, Xinting Huang, Hu Hu, Runquan Xie, Han Hu, and Zhanhui Kang. Low-hallucination synthetic captions for large-scale vision-language model pre-training. *arXiv preprint arXiv:2504.13123*, 2025.

APPENDIX

A RELATED WORK

Captions serve as a foundational supervision signal across *both* multimodal understanding and text-to-image generation, and much work has consistently demonstrated that improving caption quality drives measurable gains in downstream model performance.

In multimodal understanding, captions act as the core supervision signal for cross-modal alignment and capability building. Qwen-VL (Bai et al., 2023) first establishes captions as a foundational training anchor, unifying image description, visual grounding, and text reading abilities by aligning image–caption–box tuples. ShareGPT4V (Chen et al., 2024a) further verifies that caption quality directly determines downstream gains: replacing coarse alt-text with detailed, information-rich captions consistently boosts modality alignment, instruction following, and visual reasoning across model backbones. Most recently, the GLM series (Hong et al., 2025) refines caption properties via factuality-centered recaptioning and data filtering, lifting caption density, image–text relevance, and factual accuracy to strengthen general multimodal reasoning.

In text-to-image generation, captions act as the primary conditioning interface that enables semantic control over visual outputs. State-of-the-art generation systems have further identified caption quality as a critical performance lever: Stable Diffusion 3 (Esser et al., 2024) mixes original captions with CogVLM-generated synthetic captions (Wang et al., 2024). DALL-E 3 (Betker et al., 2023) *recaptions* training images with a dedicated caption model to enhance prompt following. Qwen-Image (Wu et al., 2025) prioritizes high-quality image-text data curation to boost overall capability.

Despite this widespread recognition of caption quality, existing evaluation paradigms, from conventional reference-based lexical matching to recent fine-grained checklist- and rubric-based assessment, face two core limitations that block a fine-grained, actionable understanding of how caption properties impact downstream learning. Early benchmarks built on Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014) and MS COCO Captions (Chen et al., 2015) quantify quality via metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) and CIDEr (Vedantam et al., 2015), which measure lexical overlap between generated captions and human-written ground-truth references. As long-form detailed captioning has become the mainstream, newer works shifted to factuality-aware, fine-grained diagnosis free from rigid reference constraints: DOCCI (Onoe et al., 2024) and DetailCaps (Dong et al., 2024) adopt atomic checklists to evaluate dense visual

coverage. PerceptionRubrics (Wei et al., 2026) uses structured scoring rubrics. GAVEL (Gao et al., 2026) and Unison (Liu et al., 2026a) further extend the scope to grounded error localization and joint understanding-generation consistency.

While we have witnessed this methodological evolution, no existing benchmark disentangles caption quality into separate coverage and reliability dimensions. All frameworks conflate the two into a single scalar score *without explicit decomposition*. Moreover, nearly all evaluation paradigms treat captions as standalone generation outputs, rather than assessing their actual utility as training supervision signals through *end-to-end* downstream experiments. To fill these gaps, we present CAPEval, a decoupled caption evaluation benchmark that explicitly decomposes caption quality into Coverage and Precision, paired with a fully end-to-end evaluation protocol that directly measures the downstream utility of captions across both understanding and generation.

B Additional CAPEval Information

B.1 GT caption and Checklist of the Example

Ground-truth caption. In the foreground and mid-ground, various amusement park attractions create a vibrant scene. A purple roller coaster track with a ladder-like structure and two long black linear elements curves upward from the bottom center toward the right. Artificial palm trees with dark green fronds are distributed throughout the lower area; a large palm tree on the left partially obscures the view of the attractions behind it. To the far left, a low fence consists of vertical panels in alternating colors of blue, orange, teal, and magenta. Behind the palm tree on the far left, a yellow and blue ride structure with bucket-style seats is visible, resembling a small vertical Ferris wheel. In the center-left, a tall, intricate structure made of orange metal beams rises vertically, topped with a teal, pagoda-style roof with multiple internal staircases leading to a slide entrance on the right. Adjacent to it in the center is a teal framework supporting a platform where a group of people stands. Extending to the right, large tubular slides wind through the space; one section is orange, another is off-white, and another is green. In the lower right quadrant, a pool with a curved blue rim is filled with a white, snow-like substance rather than water. Several inflatable inner tubes in colors including pink and blue rest on the white surface. A person wearing a long black coat, black pants, white shoes, and a light pink headscarf stands near the pool's edge in the bottom right, facing away from the camera. A small portion of another person wearing a winter hat and dark clothing is visible at the very bottom edge near the fence. Behind the slides and tracks in the mid-ground, a low wall is decorated with a colorful mural depicting a tropical beach scene with blue waves and human figures. Real palm trees stand in front of the wall, distinct from the painted mural. To the right, a ride vehicle shaped like a vintage carriage with a pink curved roof, decorative elements, and black-and-white wheels runs on tracks. In the background among bare tree branches stands a vertical tower structure with a white spherical top decorated with pink dots, a flat platform below, and a multi-colored vertical body below that. A building with pinkish walls and windows is visible on the far right. The background is dominated by a skyline of tall buildings. A large building with a grid-like facade of numerous rectangular windows and a stepped roofline stands prominently on the right side of the skyline. A tall, solid, rectangular skyscraper with a light grey or metallic surface rises next to it. A modern building with a blue glass curtain wall is situated behind the orange tower structure. A distinctive building with a geometric, angular shape and a diagonal cross-braced facade is visible in the distance to the left of the orange tower. A blocky, beige residential-style building with rows of windows stands on the far left of the skyline. The sky occupies the upper portion of the image, displaying a smooth gradient from pale blue at the top to a soft, warm pinkish-orange hue near the horizon on the right side, suggesting the lighting of dawn or dusk. The lighting is soft and diffuse, creating a calm atmosphere with low contrast.

Instance checklist.

1. **[object]** Is a purple roller coaster track mentioned in the caption?
2. **[plant]** Does the caption mention the presence of artificial palm trees?
3. **[object]** Is a low fence with multi-colored panels described?
4. **[object]** Does the caption describe a ride structure with bucket-style seats?
5. **[object]** Are large tubular slides mentioned in the description?

6. **[object]** Does the caption mention a pool with a curved blue rim?
7. **[object]** Are inflatable inner tubes mentioned in the caption?
8. **[object]** Does the caption describe a mural depicting a tropical beach scene?
9. **[object]** Is a ride vehicle shaped like a vintage carriage mentioned?
10. **[building]** Does the caption describe a vertical tower structure in the background?
11. **[building]** Are there mentions of a skyline of tall buildings in the background?
12. **[characters]** Does the caption mention the presence of a person near a pool?

Attribute checklist.

13. **[color]** Is the roller coaster track described as being purple?
14. **[shape]** Does the caption state that the roller coaster track is curved?
15. **[color]** Are the fronds of the artificial palm trees described as dark green?
16. **[color]** Does the caption specify that the fence panels have alternating colors of blue, orange, teal, and magenta?
17. **[material]** Is the tall, intricate structure described as being made of orange metal beams?
18. **[shape]** Is the roof on the tall structure described as being pagoda-style?
19. **[color]** Does the caption mention that the tubular slides have orange, off-white, and green sections?
20. **[texture]** Is the substance in the pool described as white and snow-like?
21. **[color]** Are the inner tubes mentioned to be in colors including pink and blue?
22. **[color]** Is the person's coat described as being long and black?
23. **[color]** Is the person's headscarf described as light pink?
24. **[shape]** Is the roof of the vintage carriage ride described as curved?
25. **[shape]** Is the top of the background tower described as a white sphere?
26. **[color]** Are pink dots mentioned as a decoration on the tower's top?
27. **[texture]** Is one of the background buildings described as having a grid-like facade?
28. **[material]** Is a modern building described as having a blue glass curtain wall?
29. **[shape]** Is a distant building described as having a geometric, angular shape?
30. **[color]** Does the caption mention the sky has a gradient from pale blue to pinkish-orange?
31. **[orientation]** Is the person standing near the pool described as facing away from the camera?

Relation checklist.

32. **[spatial]** Does the caption state that the roller coaster track curves upward from the bottom center toward the right?
33. **[occlusion]** Is a large palm tree on the left described as partially obscuring the view?
34. **[spatial]** Is the yellow and blue ride described as being behind the palm tree on the far left?
35. **[spatial]** Is the teal framework described as being adjacent to the orange tower in the center?
36. **[composition]** Does the caption mention a group of people standing on a platform?
37. **[spatial]** Are the inflatable inner tubes described as resting on the white surface within the pool?
38. **[spatial]** Is a person described as standing near the pool's edge in the bottom right?
39. **[spatial]** Are real palm trees mentioned to be standing in front of the wall with the mural?
40. **[spatial]** Is the vertical tower structure located in the background among bare tree branches?

41. **[spatial]** Is the modern building with a blue glass curtain wall situated behind the orange tower structure?
42. **[spatial]** Does the caption state that the sky occupies the upper portion of the image?

Image checklist.

43. **[scene_information]** Does the caption describe the foreground and mid-ground as a vibrant scene with amusement park attractions?
44. **[scene_information]** Is the background described as being dominated by a skyline of tall buildings?
45. **[emotions_and_atmosphere]** Does the caption suggest a calm atmosphere?
46. **[hue_temperature]** Does the caption suggest the lighting is from dawn or dusk due to the warm pinkish-orange hue?
47. **[light_effect]** Is the lighting described as soft and diffuse?
48. **[saturation_contrast]** Does the caption mention that the image has low contrast?
49. **[picture_composition]** Does the caption detail the placement of elements across the foreground, mid-ground, and background?

Human checklist.

50. **[human_dressing]** Does the caption mention a person wearing a long black coat and black pants?
51. **[human_dressing]** Is a person described as wearing a light pink headscarf?
52. **[human_action_position]** Is a person described as standing near the pool's edge?
53. **[human_action_position]** Does the caption state that the person near the pool is facing away from the camera?
54. **[human_dressing]** Does the caption mention another person wearing a winter hat?
55. **[human_role]** Does the caption mention a group of people standing on a platform?
56. **[human_role]** Are human figures mentioned as being part of the colorful mural?

Text checklist. No text checklist items are provided for this example.

UI checklist. No UI checklist items are provided for this example.

World knowledge checklist. No world knowledge checklist items are provided for this example.

B.2 Image Resolution Distribution

Figure 8 shows the distribution of image resolutions in the CAPEval benchmark. The images span a wide range of aspect ratios and resolutions, reflecting the diversity of real-world visual content.

B.3 GT Caption and Checklist Annotation

Ground-Truth Caption Annotation. Each image is annotated by a trained human annotator who writes a comprehensive English caption as a single unbroken paragraph. Annotators are instructed to address the following dimensions *only when the corresponding elements are present in the image*:

- *Visual subjects.* Main subjects, quantity, position, interactions, and foreground/background scene context.
- *Text content.* All legible text transcribed verbatim without translation, including layout, font, color, and text type (title, price, slogan, watermark, disclaimer).
- *Aesthetic and photographic attributes.* Art style, color tone (saturation, contrast, temperature), lens type, shot size, light source, and shooting angle.

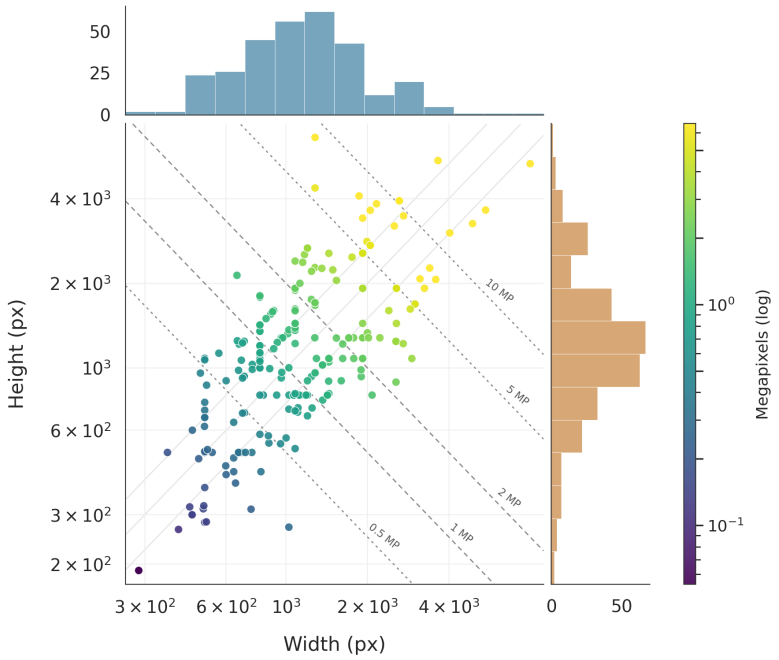

Figure 8: Distribution of image resolutions (height × width in pixels) across the 300 images in CAPEval. The benchmark includes images of various aspect ratios to ensure broad coverage of visual content types.

- *Portrait attributes.* Apparent age bracket, gender, body shape, hairstyle, clothing and accessories, posture, and expression. Celebrity or public figure identities are noted when unambiguously identifiable.
- *Special image properties.* Borders, watermarks, blur or distortion artifacts, composite or spliced images, and UI screenshot sources (app name noted where applicable).

The following constraints apply throughout:

- Describe only objectively visible elements; do not speculate, infer emotional states, or introduce subjective evaluations.
- Do not fabricate details for elements that are not clearly visible in the image.
- State each fact exactly once; do not repeat descriptions of the same entity.
- Avoid vague or generic phrasing; use precise descriptions.

Each completed caption is reviewed by a second annotator for factual accuracy and completeness; disagreements are resolved through discussion before finalization.

Checklist Annotation. Each GT caption is decomposed by human annotators into a structured checklist of atomic and verifiable statements. Every checklist item consists of two fixed fields:

- `Tags`: a categorical label string identifying the specific concept or entity the item refers to.
- `Question`: a binary yes/no question verifiable directly against the image.

Items are organized across eight mandatory semantic dimensions:

- *Instance.* Object and entity identity, category, and count.
- *Attribute.* Visual properties of entities: color, material, texture, shape, and size.
- *Relation.* Spatial and functional relationships betweenentities.

Table 3: CAPEval scores of captioners. Scores are reported on a percentage scale. For each semantic dimension, C denotes Coverage, and P denotes Precision. Overall scores are highlighted in gray.

Model	Overall		Scene & Object		People & Activity		Text & Interface		Design & Knowledge	
	C	P	C	P	C	P	C	P	C	P
Gemini-3.1-Pro	80.2	82.5	79.3	80.9	79.9	80.4	80.9	85.6	82.2	86.7
Gemini-3.5-Flash	78.1	81.4	77.5	79.2	76.1	81.4	78.8	82.9	81.4	85.8
Gemini-2.5-Pro	75.5	80.1	73.5	79.9	77.7	75.7	76.9	79.9	76.8	86.7
GPT-5.5	72.8	79.0	72.0	78.0	70.2	78.0	77.6	80.5	74.5	81.4
Qwen3-VL-32B	68.5	86.1	70.3	84.6	66.8	85.9	66.2	88.0	68.1	89.3
Qwen3-VL-8B	63.2	83.5	63.8	82.3	65.0	80.5	60.3	88.6	61.7	86.6
Qwen3-VL-4B	60.2	81.1	61.5	80.9	58.1	79.9	58.8	83.9	60.7	80.6
GLM-4.6V-Flash	62.0	85.4	62.7	85.0	58.3	86.0	62.6	86.6	64.1	84.7
InternVL3.5-38B	48.6	77.7	48.9	75.9	51.2	76.6	46.2	81.3	46.6	81.4
InternVL3.5-8B	46.5	72.6	49.1	68.5	45.4	73.8	45.3	79.8	41.9	77.3
InternVL3.5-4B	45.0	73.5	45.8	70.4	44.5	72.7	45.2	79.6	43.4	78.0
InternVL3.5-1B	48.3	65.2	47.0	64.6	53.3	60.8	47.2	74.1	46.4	65.7
LLaVA-OV-8B	41.2	50.4	45.4	46.3	38.2	55.7	30.2	59.7	42.6	50.7
LLaVA-OV-4B	43.5	45.1	44.2	44.9	45.9	42.4	35.5	48.9	45.3	46.6

- *Image.* Scene-level properties: overall composition, image quality, lighting, and global visual style.
- *Text.* All in-image text content, including signs, labels, watermarks, and interface strings.
- *Human.* Person-related attributes: apparent demographics, clothing, pose, gesture, and expression.
- *UI.* Interface elements such as buttons, menus, dialogs, icons, and layout components.
- *World Knowledge.* Named entities, landmarks, brands, and external factual associations that require knowledge beyond visual inspection alone.

The schema is exhaustive and fixed: all eight dimensions are mandatory and no custom dimensions may be added. Each item must be atomic (expressing exactly one verifiable fact), grounded in both the GT caption and the image, and assigned to exactly one dimension. Items requiring external knowledge are assigned to *World Knowledge*. A second annotator reviews all checklist items for atomicity, factual grounding, and correct dimension assignment before inclusion in the benchmark.

C CAPEval Scores and Per-captioner Downstream Scores

C.1 CAPEval scores

Table 3 reports Coverage and Precision scores for all 14 captioners across four semantic domains.

C.2 Scores of Downstream Understanding and Generation Benchmarks

To make heterogeneous benchmarks comparable, we map every raw benchmark score to a common 0–100 scale before aggregation. The normalized understanding and generation benchmark scores are shown in Table 4.

D Training Configuration

D.1 Vision-Language Understanding

Table 5 details the training setup for both VLM pipelines. Both use 8 GPUs, DeepSpeed ZeRO-3, bf16 mixed precision, AdamW optimizer (lr$= 2\times10^{-5}$, cosine schedule, warmup ratio 0.03), gradient checkpointing, and `model_max_length=2048`. The first stage (visual-language alignment) uses the fixed LLaVA 558K pretraining set for both pipelines. The second stage (supervised fine-tuning) combines the 665K instruction mixture with 1.24M captioner-specific captions, filtered by token length to ensure training stability.

Table 4: Normalized understanding and generation benchmark scores. The CAPEval scores (C and P), aggregate understanding score (U), and aggregate generation score (G) are highlighted in gray.

Metric	Qwen3-VL			GLM-4.6V	InternVL3.5				LLaVA-OV	
	4B	8B	32B	Flash	1B	4B	8B	38B	4B	8B
C	**60.2**	**63.2**	**68.5**	**62.0**	**48.3**	**45.0**	**46.5**	**48.6**	**43.5**	**41.2**
P	**81.1**	**83.5**	**86.1**	**85.4**	**65.2**	**73.5**	**72.6**	**77.7**	**45.1**	**50.4**
Understanding: CLIP - Vicuna										
POPE	87.4	86.7	85.7	87.5	87.5	86.9	86.6	85.4	81.1	85.7
Hall.	11.4	13.6	13.0	11.4	11.7	16.7	13.4	15.0	18.5	15.4
AMBER	83.4	81.9	81.5	83.6	80.7	77.3	79.9	79.4	74.1	76.6
MME	59.7	62.1	60.6	60.2	53.4	53.2	43.4	55.8	62.2	59.8
MMB-EN	83.3	80.6	83.3	75.0	72.2	61.1	66.7	75.0	75.0	80.6
MMB-CN	37.5	41.7	41.7	33.3	37.5	31.3	43.8	37.5	22.9	29.2
ScienceQA	56.4	53.9	46.2	59.0	46.2	48.7	41.0	53.9	43.6	41.0
MMStar	24.4	29.6	29.5	24.8	21.6	18.1	16.8	23.6	23.6	24.7
OCR	34.6	34.4	34.7	36.5	37.2	35.5	34.5	35.0	34.5	35.0
RWQA	42.8	47.5	41.2	53.9	49.7	43.1	33.5	46.8	42.4	50.6
MMVP	57.0	60.3	53.7	63.0	60.0	45.7	37.7	50.7	54.3	61.0
CV2D	59.8	62.8	61.5	60.4	60.0	55.5	61.8	59.9	57.0	63.1
CV3D	60.4	62.7	59.3	66.8	59.5	59.0	64.3	56.3	59.2	60.5
AI2D	56.6	56.3	57.1	56.4	56.0	56.6	55.4	55.2	54.2	56.4
SEED	70.6	69.6	70.1	70.5	70.2	69.4	70.3	69.3	68.5	68.3
MMMU	39.2	38.5	39.0	39.0	39.5	37.2	39.5	40.7	38.4	37.9
U	**54.0**	**55.1**	**53.6**	**55.1**	**52.7**	**49.7**	**49.3**	**52.5**	**50.6**	**52.9**
Understanding: SigLIP - Qwen3										
POPE	87.6	88.6	88.5	87.7	87.5	87.9	87.2	85.1	88.9	87.7
Hall.	28.1	30.8	24.6	25.7	27.0	28.1	31.4	27.5	27.7	27.7
AMBER	85.4	80.5	85.0	86.2	82.6	80.8	84.3	84.8	84.7	80.3
MME	68.9	68.2	66.8	67.7	67.4	69.0	66.2	66.2	68.5	69.6
MMB-EN	88.9	94.4	91.7	94.4	91.7	86.1	83.3	88.9	88.9	91.7
MMB-CN	85.4	81.3	79.2	75.0	68.8	68.8	79.2	68.8	83.3	77.1
ScienceQA	66.7	69.2	66.7	69.2	56.4	56.4	48.7	51.3	64.1	62.1
MMStar	42.9	41.2	42.5	40.4	42.4	42.7	43.6	46.4	43.2	41.8
OCR	45.6	45.0	44.3	45.4	44.1	45.1	43.5	45.1	44.4	44.4
RWQA	62.0	64.4	64.8	64.8	61.7	63.1	62.8	63.4	61.7	61.2
MMVP	75.8	77.3	78.7	79.7	74.6	75.3	75.7	74.3	74.0	74.3
CV2D	71.8	71.0	71.3	67.5	67.7	70.5	69.0	69.1	67.9	69.3
CV3D	75.4	72.7	71.1	73.0	67.9	76.2	76.6	76.2	68.1	69.1
AI2D	71.9	65.0	70.8	70.9	68.8	70.3	70.2	71.1	70.6	69.8
SEED	75.3	69.6	75.5	76.0	74.2	74.4	75.0	75.0	73.9	74.0
MMMU	48.5	46.1	47.9	44.9	44.9	47.7	46.7	46.8	46.5	46.8
U	**67.5**	**66.6**	**66.8**	**66.8**	**64.2**	**65.2**	**65.2**	**65.0**	**66.0**	**65.5**
Generation: SD3.5M										
GenEval	70.5	70.9	71.6	70.4	64.4	68.1	65.9	67.5	59.1	58.4
DPG	85.0	85.1	85.0	85.0	83.7	83.7	83.3	84.6	80.8	81.6
T2I	52.2	53.0	53.9	52.3	49.6	51.3	50.2	51.8	46.0	46.3
G	**69.2**	**69.7**	**70.2**	**69.3**	**65.9**	**67.7**	**66.5**	**68.0**	**61.9**	**62.1**
Generation: Qwen-Image										
GenEval	87.0	87.4	88.7	86.7	85.0	87.1	86.0	86.7	76.0	77.2
DPG	86.4	86.7	86.3	85.9	85.4	85.0	86.2	85.7	82.6	83.2
T2I	54.9	55.1	55.2	53.9	52.6	54.1	54.1	54.2	48.1	46.7
G	**76.1**	**76.4**	**76.7**	**75.5**	**74.3**	**75.4**	**75.4**	**75.5**	**68.9**	**69.0**

Vision encoders and LLMs: CLIP uses ViT-L/336px with Vicuna-7B; SigLIP uses SigLIP-SO400M/384px with Qwen3-4B. **Projector:** CLIP warm-starts from the pretrained LLaVA MLP adapter; SigLIP trains the projector from scratch. **ViT tuning:** layers ≥ 12 are unfrozen during alignment.

Table 5: VLM training configuration.

Pipeline	Stage	Data	Global BS	Per-dev	Accum	Steps
CLIP–Vicuna	1 (align)	LLaVA 558K	256	8	4	∼2,180
	2 (SFT)	665K+1.24M	128	4	4	∼14,900
SigLIP–Qwen3	1 (align)	LLaVA 558K	256	8	4	∼2,180
	2 (SFT)	665K+1.24M	64	2	4	∼29,800

Table 6: T2I training configuration.

Suite	Images	LR	Steps	Warmup	Loss Weight	DeepSpeed
SD3.5M	50K	5×10^{-6}	1600	100	logit-normal	ZeRO-2
QwenImage	100K	1×10^{-5}	3200	200	none	ZeRO-3

D.2 Text-to-Image Generation

Table 6 details the training setup for both T2I suites. Both use DeepSpeed (ZeRO-2 for SD3.5M, ZeRO-3 with optimizer offload for QwenImage), bf16 mixed precision, AdamW (β_1=0.9, β_2=0.999, weight decay 10^{-4}), cosine LR schedule, gradient clipping (max norm 1.0), resolution 1024, max sequence length 512, and global batch size 64. Only the diffusion transformer (DiT) parameters are trainable; text encoders and VAE are frozen, with embeddings or latents precomputed.

Preprocessing: SD3.5M precomputes T5 and CLIP text embeddings; QwenImage precomputes both VAE latents and text embeddings.