SPECS: Specificity-Enhanced CLIP-Score for Long Image Caption Evaluation

Anonymous ACL submission

Abstract

As interest grows in generating long, detailed image captions, existing automatic evaluation metrics are increasingly strained. N-gram-based metrics though efficient, fail to capture semantic correctness, especially for longer outputs. Representational Similarity (RS) metrics, designed to address this, initially saw limited use due to high computational costs, while today, despite advances in hardware, they remain unpopular as they fall short even of weak baselines such as BLEU. Meanwhile, metrics based on large language models (LLMs) show strong correlation with human judgments, but remain too expensive for use in model development. We introduce SPECS (Specificity-Enhanced CLIP-Score), a reference-free RS metric tailored for long image captioning. SPECS modifies CLIP with a new objective that emphasizes specificity: rewarding correct details and penalizing errors. We show that SPECS matches the performance of leading LLM-based metrics in correlating with human judgments, while being far more efficient. This makes it a practical alternative for iterative checkpoint evaluation during image captioning model development.

1 Introduction

As the task of short image captioning approaches saturation, interest is growing toward the more challenging task of long, detailed image captioning (Johnson et al., 2016; Cho et al., 2022; Doveh et al., 2023; Li et al., 2023). This task requires strong visual grounding of the generated text and improved cross-modal alignment (Liu et al., 2024; Li et al., 2021). While this shift opens new frontiers for research in generative vision-language understanding, it also exacerbates the long-standing issue of reliable automatic evaluation.

Image captioning, like any natural language generation task, has long been a challenge when it comes to evaluation (Otani et al., 2023; Wang et al., 2023). Early metrics based on the n-gram overlap

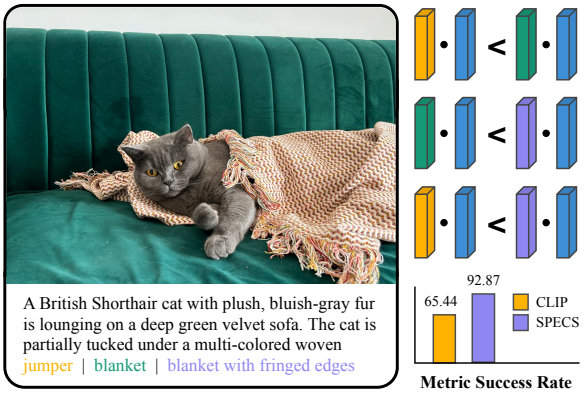

Figure 1: A model which exhibits good specificity, i.e. it ranks image-caption pairs in the correct way depicted above, makes for a good evaluation metric. SPECS succeeds in ranking minimal pairs correctly at a near-ceiling rate. The blue vector is the image representation.

of generated captions and references are fast but void of semantics and as such, highly inaccurate even as applied to short captions (Papineni et al., 2002; Banerjee and Lavie, 2005; Vedantam et al., 2015; Lin, 2004). Representational Similarity (RS) metrics were thus introduced to alleviate this issue through modeling the semantics of images and text, often operating in a reference-free fashion (Hessel et al., 2021; Sarto et al., 2023). Most recently, a proliferation has been seen of metrics that leverage the capabilities of large language models (LLMs) and prove superior to earlier ones, especially as the length of generated image captions increases (Chan et al., 2023; Yu et al., 2024; Ye et al., 2025).

Every new generation of evaluation metrics comes with magnitudes higher computational requirements, largely explained by the ever-growing capacity of modern hardware. Yet, there is typically a mismatch between what is *possible* at a given point in time, and what is *practical*. CLIPScore (Hessel et al., 2021), for example, was not as widely adopted as its apparent superiority over CIDEr (Vedantam et al., 2015) would call for, be-

cause for its time, it was a prohibitively expensive metric. Similarly now, the state-of-the-art LLM-based metrics are often used for final model evaluation, but remain prohibitively costly and impractical during iterative model development. In the context of long caption generation, however, there is hardly any viable alternative.

In a first effort towards benchmarking automatic evaluation metrics against human judgements on the task of long image captioning, Ye et al. (2025) report that the Sample-wise Kendall's Tau for CIDEr is a mere 0.06, and that the correlation caps at 0.29 for BLEU-4 (Papineni et al., 2002). RS metrics, which stand to be more practical today than they were in earlier, more computationally constrained times, all prove less effective than the simple, almost zero-cost BLEU-4 metric. CLIPScore, for example, scores at only 0.15. This calls for a new and effective RS metric.

In this work, we introduce **SPECS** (**S**pecificity-**E**nhanced **C**LIP-**S**core), a reference-free, RS evaluation metric which uses a CLIP model adapted to longer context (Zhang et al., 2024), and augmented with a novel objective that emphasizes *specificity*: the ability to reward correct details and penalize incorrect ones. In terms of human judgment correlation, SPECS matches the performance of the best open-source LLM metric (Lee et al., 2024), at a fraction of the computational cost. As such, SPECS stands to become the norm for fast-iteration evaluation in long image caption generation.

2 Related work

CLIP (Radford et al., 2021) is a 150-million parameters dual-encoder model trained on 400 million image-text pairs using a contrastive objective. It learns to predict which caption matches a given image, enabling the model to acquire broad visual-language representations that transfer well to downstream tasks without task-specific fine-tuning. CLIPScore (Hessel et al., 2021) was proposed as a reference-free image caption evaluation metric. It uses CLIP embeddings to compute cosine similarity between images and captions, yielding better correlation with human judgments than traditional n-gram-based metrics on short image captions.

To improve CLIP's capacity for fine-grained understanding and compositionality, several extensions have been introduced. NegCLIP improves compositional reasoning by fine-tuning CLIP with hard negative samples, as evaluated on the ARO benchmark (Yuksekgonul et al., 2022). LaCLIP enhances language diversity by rewriting image captions with large language models and training CLIP on both original and rewritten captions (Fan et al., 2023). TriletpCLIP introduces synthetic triplets with hard negatives for both modalities and optimizes a triplet loss to improve compositional generalization (Patel et al., 2024). Other methods such as DCI (Urbanek et al., 2024) and DAC (Doveh et al., 2023) improve compositionality through training on dense, region-aligned captions.

LongCLIP addresses the issue of CLIP's small input length window, extending it from 77 to 248 tokens by interpolating positional embeddings (Zhang et al., 2024), thus enabling the model to handle long captions. The model is trained on a large syntehtic dataset of long, detailed captions, ShareGPT-4V (Chen et al., 2024), with gradients blocked in the first 20 tokens, to preserve CLIP's strong zero-shot capabilities in this window.

Building on prior work, we introduce the notion of *specificity*—the ability of a vision-language model to consistently prefer more informative, visually grounded captions at varying caption lengths, and train a LongCLIP model with it to build an effective long image captioning evaluation metirc.

3 Specificity: A Fine-Grained VLM Metric

Let us consider the three caption variants depicted in Figure 1: saying that the cat is tucked under a *blanket* is correct and there is a lot of additional relevant detail mentioned before that, so a good evaluation metric should give this variant a high score. And if the caption was to mention that the blanked has *fringed edges*, the evaluation metric should reflect this detail too, in a slightly higher score. On the other hand, if instead of a *blanket* the caption said that the cat was laying under a *jumper*, that should result in a slightly lower score– all the other relevant detail is still present, but this particular object mention is incorrect. This simple example illustrates the notion of *specificity* which we adapt from Xu et al. (2024) to mean: the ability of a representation to encode every detail in a caption in a way that correctly reflects the relevance of this detail to a reference image. A metric based on a specificity-enhanced model, would thus favor captions that include more relevant details, and penalize those that omit important information or introduce hallucinated content. This aligns

closely with the notions of precision and recall, implemented implicitly in such a metric.

3.1 Detail Units

To concretely evaluate specificity, we begin by introducing the key concept of a **detail unit**. The abstract notion behind a detail unit refers to any minimal bit of information in a caption, such as the presence of a *blanket*, the *fringed edges* of the blanket, etc. For operational purposes, however, we define a detail unit to mean a phrase which contributes at least one new visual detail (and possibly more), and fits syntactically and semantically with preceding context. Under this definition, *a blanket with fringed edges* is a detail unit, but *a blanket with* is not, and neither is *The cat* mentioned in the middle of the caption in Figure 1, since it does not contribute new information.

Formally, we denote an image-caption pair as $\{i, c\}$, and decompose a caption as $c = \{d_1, d_2, \ldots, d_m\}$ where each d_i is a detail unit. Every subsequence of detail units, built cumulatively from left to right, constitutes a valid caption: $c_1 = \{d_1\}, c_2 = \{d_1{+}d_2\}, ..., \{c = d_1{+}\cdots{+}d_m\}$, each containing progressively more information.

Given a perfect caption, this ordered sequence should exhibit monotonically increasing representational similarity to its reference image, under a specificity-enhanced model. Conversely, if an incorrect detail unit is added at any point, this should be reflected in a decreased similarity. This decomposition provides a structured way to test and enhance model specificity to visual detail across any caption length.

Detail units that contain relevant information are referred to as **positive** (d_+), while detail units that introduce content not grounded in the image as referred to as **negative** (d_-). The expected behavior of a model with good specificity is then to assign higher similarity to the pair $\{i, c_j + \hat{d}_+\}$ than to the pair $\{i, c_j\}$, and a lower similarity to the pair $\{i, c_j + \hat{d}_-\}$ than to the pair $\{i, c_j\}$, where $j \in [1, ..., m]$ and the hat symbol denotes a new candidate detail. Each triplet, $\{i, c_j, c_j + d_+\}$ and $\{i, c_j, c_j + d_-\}$ constitutes a minimal pair of captions grounded in an image, the former being positive and the latter negative.

3.2 Specificity Rate

To aggregate specificity across a set of minimal pairs, we introduce the **Specificity Rate** (SR). We define two variants: SR_{pos} measures the proportion of cases in which adding an additional relevant detail (positive detail unit) increases the similarity score with the image, while SR_{neg} measures the proportion of cases in which adding an irrelevant detail (negative detail unit) decreases the similarity. Given a set of N positive or negative triplets, we compute the SR as follows:

$$\text{SR}_{\text{pos}} = \frac{1}{N} \sum_{j}^{N} \mathbb{I}[\theta(i, c_j + \hat{d}_+) > \theta(i, c_j)] \quad (1)$$

$$\text{SR}_{\text{neg}} = \frac{1}{N} \sum_{j}^{N} \mathbb{I}[\theta(i, c_j) < \theta(i, c_j + \hat{d}_-)] \quad (2)$$

where $\mathbb{I}[\cdot]$ is the indicator function that outputs 1 if the condition inside is true and 0 otherwise, and θ is the cosine similarity between the representations of image and text. This formulation captures the rate at which representational similarity increases with added positive details, or decreases with added negative ones, providing a robust method of measuring model specificity.

3.3 Specificity-Aware Leaning

Although specificity can be used purely for evaluation, we can also enforce it during training. To encourage the model to prefer captions that describe images with greater relevant detail, we introduce a training objective that rewards higher similarity scores for incrementally more informative captions, and lower similarity scores for less accurate ones. Given a dataset of size N of positive and negative triplets, we define the following hinge loss with a dynamic margin:

$$\mathcal{L}_{\text{pos}} = \frac{1}{N} \sum_{i}^{N} \max\left(0, \theta(i, c_j + \hat{d}_+) - \theta(i, c_j) + \epsilon\right), \quad (3)$$

where ϵ is a batch-wise average similarity difference between detailed and base captions, which is detached from gradient computation and clamped for numerical stability:

$$\epsilon = \text{detach}\left(\frac{1}{N} \sum_{i}^{N} \Big(\theta(i, c_j + \hat{d}_+) - \theta(i, c_j)\Big)\right),$$

The negative loss, $\mathcal{L}_{\text{pos}}$ is computed by analogy, from the negative triplets in the dataset.

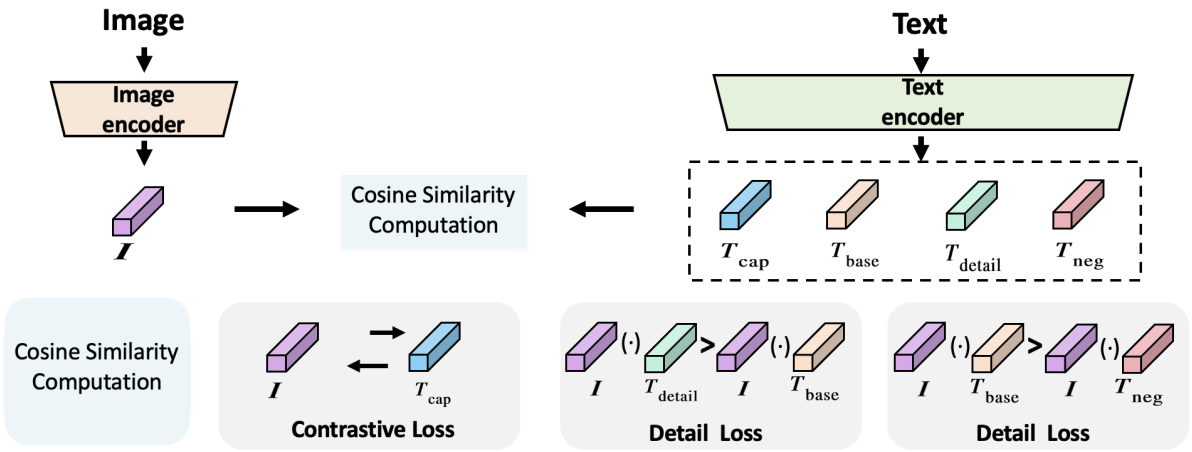

Figure 2: Training framework of model. Given an image and its caption, we generate a base caption, a more detailed caption, and a negative caption containing an incorrect detail. The model computes CLIP-style embeddings and is trained with three losses: contrastive loss on the full caption, a detail-aware loss to prefer more informative descriptions, and a negative detail loss to penalize misleading ones. This setup encourages sensitivity to fine-grained textual differences.

Our final training objective combines the contrastive loss, the detail-aware loss, and the negative detail loss:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{contrastive}} + \beta \mathcal{L}_{\text{detail}} + \gamma \mathcal{L}_{\text{neg}}, \quad (4)$$

where α, β, and γ are weighting hyperparameters tuned on a validation set. Figure 2 illustrates the overall training framework.

4 Experiment and Evaluation

4.1 Training and Validation Datasets

We train our model on the ShareGPT-4V dataset (Chen et al., 2024), which contains 1.2 million high-quality image-caption pairs synthetically generated by a strong captioning model, instructed to mention object attributes, spatial layouts, and aesthetic properties. The images in the dataset are sourced from COCO (Lin et al., 2014), SAM (Kirillov et al., 2023), and LAION (Schuhmann et al., 2022), and captions are 143 tokens long on average.

For intrinsic specificity evaluation, we use the sDCI dataset (Urbanek et al., 2024), consisting of 7805 images, each paired with 10 captions, which are synthetically desgined to fit in CLIP's context window of 77 tokens. This underuntilizes the full context window of our model, but enables controlled comparisons to other models, constrained by the 77-token context window, specifically the models introduced in Section 2.

4.2 Caption Segmentation

To prepare data for measuring specificity, we build a pipeline that segments captions into detail units.

Main Logic We considered various methods of implementing the logic described above, based on part-of-speech tagging, dependency parsing and coreference resolution: the results were either unsatisfactory, slow to obtain or even impossible to obtain due to technical challenges with the deployment of outdated libraries for coreference resolution, for example. The solution which proved best in terms of speed, implementational ease, and satisfactory results was obtained with the help of GPT-4 (Achiam et al., 2023). We presented the model with an example of a manually annotated caption and had it generate Python code that would follow the pattern of pipe symbol insertions. The resulting code is based on part-of-speech tagging and a rule-based grammar (see Appendix A).

This solution produced surprisingly good results as determined by the manual inspection of captions annotated with it. However, it somewhat oversegments the captions.

False Negatives vs. False Positives Considering the specific use of segmented data, we determined that introducing false negatives (i.e., missing splits) is less harmful than introducing false positives (i.e., incorrect extra splits). In other words, we prefer case (a), where a necessary split is missing, over case (b), where an incorrect split is inserted:

(a) *A front view of a statue on cement | in a park.*

(b) *A front | view of a statue | on cement | in a park.*

This preference is grounded in our design of both the metric and the training objective. We aim to ensure that every introduced segment corresponds to meaningful and novel information. Allowing fewer false negatives yields a finer segmentation, which aligns with our objective of evaluating incremental detail. On the other hand, false positives introduce noise that may corrupt the metric signal and compromise training, especially when such errors accumulate.

Given the above reasoning, we modify the segmentation code with several rules to avoid splitting off (1) sentence-initial noun phrases that begin with *The* as they are likely to introduce a previously mentioned entity, (2) prepositional phrases from the noun phrase preceding them as they are likely a modifier to the noun phrase, often referring back to previously mentioned objects (e.g. *The cat † is partially ...* in Figure 1),[1] (3) segments which start with a prepositional phrase from the context that follows, unless the segment contains a verb, as they are likely a location modifier to the following noun phrase (e.g. *To the left of the car † there is a box*).[2]

4.3 Experimental Setup

We train a base LongCLIP-B/32 model with a context window of 248, on the ShareGPT-4V dataset, for six epochs. The best checkpoint is then further finetuned for three epochs with our enhanced specificity objective, using ShareGPT-4V captions, segmented as described above. For every triplet $\{i, c_j, c_j + d_+\}$, we create a negative counterpart, $\{i, c_j, c_j + d_-\}$, by randomly sampling a detail unit from another image-caption pair in the batch. We use the Adam optimizer with a learning rate of 1×10^{-5}, weight decay of 1×10^{-2}, batch size of 100 per GPU, and gradient accumulation over 4 steps (yielding an effective batch size of 400). We set the loss weights to $\alpha = 1$, $\beta = 8$, and $\gamma = 0.8$ based on extensive hyperparameter tuning.

All experiments are conducted on four NVIDIA A40 GPUs. Training the model requires approximately one hour per epoch (4 GPU hours).

[1] † denotes a pipe symbol that the rule blocks.

[2] Sometimes, this rule would result in a false negative.

Model	Positive	Negative	Average
CLIP	62.61	68.28	65.44
LongCLIP	60.12	69.93	65.02
SigLIP	58.56	76.28	67.42
NegCLIP	54.96	78.84	66.90
DCI	55.68	63.63	59.66
DAC	46.84	66.88	56.86
La-CLIP	60.98	68.82	64.90
TripletCLIP	53.34	70.20	61.77
LongCLIP*	58.64	77.03	67.83
SPEC	**95.37**	**90.37**	**92.87**

Table 1: Specificity performance of various vision-language models on the sDCI dataset (**Specificity-Aware Learning**). Positive and Negative correspond to SR_{pos} and SR_{neg} as defined in Section 3.2. *LongCLIP* refers to the ViT-B/16 model as reported in the original LongCLIP paper, while *LongCLIP** is our own implementation using ViT-B/32.

4.4 Results

Intrinsic Evaluation To evaluate whether our training objective effectively enhances specificity, we specificity rate of various vision-language models (see Section 3.2.) Table 1 reports results on the sDCI benchmark. Our specificity-enhanced model, **SPEC** (Specificity-Enhanced CLIP), achieves the best performance across all VLM models, with SR_{pos} = 95.37 and SR_{neg} = 90.37, resulting in an average specificity score of 92.87. Compared to the LongCLIP* baseline of 67.83, our model yields a substantial improvement of +25.04 points. The largest gain appears in SR_{pos}, where SPEC outperforms LongCLIP* by +36.73, highlighting its superior ability to recognize and prefer more detailed captions, and the effectiveness of the custom training objective.

Interestingly, we observe that models with strong general-purpose performance do not necessarily achieve high specificity scores. For example, SigLIP, which has demonstrated strong results in standard vision-language benchmarks, underperforms in both SR_{pos} and average specificity compared to CLIP-based variants. This suggests that architectural strength alone is insufficient for capturing fine-grained alignment between image and caption. Models with enhanced compositionality, a property we expect should be highly relevant to specificity, show mixed performance: NegCLIP is a bit better than CLIP on average, but DCI, DAC show reduced specificity and La-CLIP shows no

Metric	PCC $\rho \uparrow$	1 - $R^2 \downarrow$	Kd $\tau \uparrow$	Sp $\tau \uparrow$	Base Model	Reference Free	TFLOPs
Rule-Based Evaluation							
BLEU-4	0.3439	62.78	0.2693	0.2931	-	✗	-
ROUGE	0.2509	156.05	0.1886	0.1893	-	✗	-
METEOR	0.3593	111.95	0.2417	0.2536	-	✗	-
CIDEr	0.0522	3.30E+07	0.0635	0.0601	-	✗	-
Representational Similarity Evaluation							
SPICE	0.2218	156.11	0.1731	0.1907	-	✓	-
CLIP-Score	0.2183	26.04	0.1724	0.1480	CLIP	✓	1.48×10^{-2}
PAC-Score	0.1525	20.93	0.1117	0.1260	CLIP	✓	1.48×10^{-2}
La-CLIP	0.1177	71.94	0.0911	0.1192	CLIP	✓	1.48×10^{-2}
TripletCLIP	0.1697	34.70	0.0852	0.1038	CLIP	✓	1.48×10^{-2}
NegCLIP	0.0872	131.57	0.0623	0.0256	CLIP	✓	1.48×10^{-2}
LongCLIP	0.2320	18.58	0.1769	0.2603	LongCLIP	✓	2.81×10^{-2}
LongCLIP*	0.1723	33.67	0.1484	0.1662	LongCLIP	✓	2.81×10^{-2}
SPECS (Ours)	0.5228	3.65	0.4078	0.5400	LongCLIP	✓	2.81×10^{-2}
LLM-Based Evaluation							
FaithScore	0.1937	3.22	0.1626	0.1115	LLaMA	✓	3.97
CLAIR	0.3815	1.98	0.3847	0.4552	LLaMA	✓	3.97
GPT4-Eval	0.3976	2.95	0.3447	0.3866	GPT-4	✓	-
RLAIF-V	0.3547	5.32	0.2774	0.2544	LLaVA	✓	3.97
CAPTURE	0.3521	7.62	0.2801	0.3449	LLaMA2	✗	7.74
FLEUR	0.4230	3.01	0.4246	0.5325	LLaVA	✓	7.74
DCSCORE	0.6605	1.54	0.5328	0.6166	GPT-4o	✗	-

Table 2: **Correlation of image captioning evaluation metrics and human judgements:** Pearson's ρ, $1 - R^2$, Kendall's τ (Kd τ), and Spearman's τ (Sp τ). Metrics are grouped into Rule-Based, Model-Based, and LLM-Based categories. For a fair comparison of computational cost, LLM-Based evaluations were computed using an input sequence length of 300 tokens, matching the setting used for Model-Based metrics. All p-values are less than 0.001.

change from the baseline model.

Extrinsic Evaluation To evaluate how well automatic caption metrics align with human preferences, we adopt the evaluation protocol from DECAPBENCH (Ye et al., 2025). Their human correlation benchmarks consists of 500 image caption pairs, with images sample from the ImageInWords (IIW) dataset (Garg et al., 2024). Human-annotated ratings are available for five captions per image, generated by different vision-language models. This setup enables a standardized comparison between automatic metrics and human judgments.

Here we see how SPECS, a version of CLIPScore which builds on our specificity-enhanced LongCLIP model, performs compared to a range of other metrics from different categories: rule-based, representational similarity-based and LLM-based. To quantify alignment with human ratings, we report four standard correlation metrics: Pearson correlation coefficient (PCC), coefficient of determination (R^2), Kendall's τ (Kd τ), and Sample-wise τ (Sp τ). Table 2 summarizes results across all evaluation methods previously considered in Ye et al. (2025), the CLIP variants most related to our work, and lastly, our proposed SPECS.

Among model-based metrics, SPECS achieves the highest correlation with human judgments. Compared to the standard CLIPScore baseline, it improves PCC from 0.2183 to 0.5228 and Kendall's τ from 0.1724 to 0.4078. It also outperforms most LLM-based alternatives. For instance, although FLEUR achieves strong performance among LLM methods, SPECS still surpasses it in both PCC and ranking consistency.

In terms of efficiency, SPECS requires only 2.81×10^{-2} TFLOPs per forward pass, making it significantly more efficient than LLM-based evaluators such as FLEUR (7.74 TFLOPs) or CLAIR (3.97 TFLOPs). Both CLIP and our SPECS model have approximately 0.15 billion parameters, further contributing to their lightweight and scalable design. This highlights SPECS as a scalable and

Model	Rel.	Attr.	C/O	F/O	SCPP
CLIP	59.84	63.96	47.28	58.54	53.33
LongCLIP	59.70	63.42	56.91	69.03	54.45
SigLip	46.52	56.24	32.95	40.86	20.88
NegCLIP	70.52	**81.08**	87.04	90.38	**63.79**
DCI	81.31	73.85	**94.53**	**95.68**	51.29
DAC	**76.18**	67.63	88.58	91.25	43.54
La-CLIP	45.48	58.72	34.97	40.54	54.99
TripletCLIP	54.94	63.07	23.53	27.58	55.71
LongCLIP*	52.96	65.81	63.97	70.20	56.74
SPEC	73.38	69.31	75.23	84.96	35.61

Table 3: **Performance of various models on the ARO and SCPP benchmarks. C/O** and **F/O** correspond to compositionality evaluation on COCO-Order and Flickr30k-Order, respectively.

human-aligned alternative for evaluating dense and detail-rich image captions.

5 Further Analysis

5.1 Compositionality Analysis

Although our primary focus is on improving specificity, we also examine whether this ability correlates with compositional reasoning skills. Intuitively, one might expect that models capable of handling compositional variations–such as changes in attribute order or relational details–would also be better at processing incrementally detailed descriptions. To explore this connection, we evaluate our models on two established compositionality benchmarks: the ARO benchmark (Yuksekgonul et al., 2022), which measures understanding of attribute-relation-object combinations and word order sensitivity, and the SugarCREPE++ (SCPP) benchmark (Dumpala et al., 2024), which tests sensitivity to semantic equivalence under lexical variation. Table 3 summarizes the results. Full SCPP results are provided in Appendix B.

On ARO, SPEC exhibits considerably higher performance than LongCLIP*, which suggests a direct relationship between specificity and compositionality. This finding does not hold on the SCPP benchmark, however. In fact among all compositionality-enhanced models, only NegCLIP shows a marked improvement on SCPP over the base CLIP model, all others either matching the base performance or showing a considerable degradation (for example, DAC.) . Our specificity objective includes hard negatives that do not explicitly encourage the model to handle semantic equivalence. This may partly explain the reduced performance on SCPP.

5.2 Hubness in Embedding Space

Although our specificity-enhanced model excels in fine-grained alignment, we observe a decline in performance on standard vision-language tasks such as retrieval and classification. We evaluate generalization on a diverse set of benchmarks, including Urban-1k (Zhang et al., 2024) and COCO (Lin et al., 2014) for text-image retrieval, and ImageNet (Russakovsky et al., 2015), CIFAR-10, and CIFAR-100 (Krizhevsky et al., 2009) for image classification. These datasets cover both multimodal and unimodal settings, providing a comprehensive view of how specificity-oriented training affects general-purpose representations.

Specifically, our training objective modifies the geometry of the embedding space by introducing additional constraints beyond contrastive similarity, in particular encouraging alignment with incrementally detailed captions. While this enhances specificity, it disrupts the isotropy of the space and leads to the emergence of hubness caption embeddings that are overly similar to many images ultimately degrading retrieval performance.

Overall, while our training strategy enhances specificity evaluation, it can distort the geometry of the embedding space, negatively affecting performance on related tasks. This does not devalue the SPECS metric, but sheds some light into the mechanism it adopts to provide reliable evaluation scores for long image captions.

5.3 Ablation Studies

We perform ablation studies to investigate the impact of four key factors: loss weight configuration (α, β, γ), learning rate, loss type, and dataset shuffle ratio. Table 5 summarizes our results.

The optimal setting $(\alpha = 1, \beta = 8, \gamma = 0.8)$ achieves the highest specificity score of 92.87. Alternative configurations such as $(1 : 9 : 0.8)$ and $(1 : 8 : 0.6)$ result in noticeably lower performance, underscoring the model's sensitivity to the precise relative weighting of different training objectives.

Interestingly, the optimal setting is highly imbalanced, placing much greater emphasis on the detail loss compared to the contrastive and negative loss components. We hypothesize that this imbalance arises from the nature of our specificity-focused training setup: since the contrastive loss is already well optimized from the pretrained CLIP checkpoint, and the negative detail examples are relatively noisy, the model benefits more from strong

Model	Urban-1k		COCO		Classification		
	Text-Image	Image-Text	Text-Image	Image-Text	ImageNet	CIFAR-10	CIFAR-100
CLIP	47.10	61.10	30.45	50.40	68.40	89.75	64.20
LongCLIP	79.30	79.20	40.40	57.63	66.80	90.69	69.30
SigLip	62.40	63.10	47.18	65.34	76.08	92.44	72.59
NegCLIP	52.80	55.60	41.56	56.86	55.84	85.90	60.90
DCI	43.00	29.70	21.44	20.55	53.34	87.38	57.96
DAC	23.60	11.40	37.53	33.49	52.36	89.86	64.04
LongCLIP*	77.00	75.80	35.50	52.44	59.91	90.38	66.36
SPEC	69.80	0.30	22.72	4.48	11.01	71.26	33.1

Table 4: **Evaluation across multiple benchmarks.**

Ablation	Config	Pos.	Neg.	Avg.
Loss Weight 1:8:0.8	1:8:0.6	85.85	82.33	84.09
	1:9:0.8	87.39	86.59	86.99
	1:8:0.8	95.37	90.37	92.87
Learning Rate 1e-5	1e-6	77.22	68.64	72.93
	5e-6	83.57	76.54	80.55
	1e-5	95.37	90.37	92.87
Loss Type hinge	esp1e-3	80.18	65.80	72.99
	hinge	95.37	90.37	92.87
Dataset Shuffle 90%	50%	83.12	87.33	85.22
	100%	87.27	83.11	85.19
	90%	95.37	90.37	92.87

Table 5: **Ablation study over four factors.**

and consistent supervision on the positive detail signal. The detail loss directly encourages the model to increase similarity for incremental, visually grounded additions—precisely the type of fine-grained distinction we aim to capture. Thus, assigning a large weight to this component reinforces the core objective of our method.

We also investigate the role of shuffle ratios when constructing negative captions. Our motivation for introducing shuffling is to avoid overly simplistic negative examples. Since each negative caption is created by appending a detail chunk—sourced from other images in the batch—to the current base caption, using unshuffled chunks may result in semantically coherent and fluent text that unintentionally resembles a valid caption. This risks introducing false negatives that confuse the model during training.

To address this, we introduce a shuffle ratio hyperparameter that controls the proportion of detail chunks that are randomly shuffled at the token level before being appended. A ratio of 90% means most negative chunks are shuffled to break semantic coherence, while a small portion (10%) remain in their original order to preserve some challenging cases. We find that shuffle(90%) yields the best performance. We interpret this as a balance between two extremes: fully shuffled chunks may become too disfluent and easy to reject, offering little learning signal, while mostly unshuffled chunks increase the chance of false negatives due to plausible but incorrect details. The optimal performance at shuffle(90%) suggests that introducing controlled noise into the negatives improves the model's ability to focus on genuine detail alignment without being misled by surface-level fluency.

6 Conclusion

We introduce specificity as a critical dimension for evaluating vision-language models, emphasizing their capability to distinguish and reward fine-grained visual details. By fine-tuning a CLIP model with a specificity-aware learning objective, we develop SPECS (SPecificity-Enhanced CLIP Score), an evaluation metric designed to better align with human sensitivity to visual details. Extensive experiments demonstrate that SpeCS significantly outperforms existing methods in specificity evaluation and human correlation, while remaining computationally efficient and scalable. We believe this work provides a foundational step toward more detailed and precise evaluation methodologies, facilitating future improvements in vision-language understanding.

Limitations

While SPECS offers strong alignment with human judgments and excels at evaluating fine-grained visual grounding, its performance on standard vision-language tasks is comparatively limited. As shown in compositionality benchmarks such as ARO and SCPP++, improvements in specificity do not directly translate into better reasoning over attribute structures or lexical variations. This indicates that the specificity-focused objective does not generalize well to tasks requiring structural or semantic invariance.

In addition, our hubness analysis reveals distortions in the embedding space caused by specificity-aware training. By encouraging sensitivity to incremental visual details, the model tends to over-align with frequent or stylistically similar captions, leading to degraded performance in retrieval and classification tasks. These findings highlight a trade-off between detail sensitivity and general-purpose utility.

Addressing this trade-off remains an open challenge. Future work may consider architectural modifications or auxiliary learning objectives that preserve fine-grained grounding while improving transferability to downstream tasks.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. 2023. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*.

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

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527, Seattle, United States. Association for Computational Linguistics.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, and 1 others. 2023. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36:76137–76150.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2009. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. *arXiv preprint arXiv:2406.06004*.

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *Preprint*, arXiv:2407.21771.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286.

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

Maitreya Patel, Naga Sai Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and 1 others. 2024. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in neural information processing systems*, 37:32731–32760.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and 1 others. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6914–6924.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

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

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

Yixuan Wang, Qingyan Chen, and Duygu Ataman. 2023. Delving into evaluation metrics for generation: A thorough assessment of how metrics generalize to rephrasing across languages. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 23–31.

Zhenlin Xu, Yi Zhu, Siqi Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. 2024. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1827–1836.

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

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and 1 others. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

A Segmentation Grammar

The following context-free grammar was used to define syntactic structures relevant to caption segmentation:

```
grammar = r"""
    NP: {<DT>?<JJ.*>*<NN.*>+} # Noun phrase with
        optional determiners and adjectives
    VP: {<VB.*><NP|PP|CLAUSE>+$} # Verb phrase
        with verb followed by noun phrases,
        prepositional phrases, or clauses
    PP: {<IN><NP>} # Prepositional phrase with
        preposition followed by noun phrase
    CLAUSE: {<NP><VP>} # Clause containing noun
        phrase followed by verb phrase
    CONJ: {<CC><NP|VP|PP|CLAUSE>} # Conjunction
        with conjoined structures
"""
```

B SCPP++ Result

Table 6 presents the full results on the SCPP++ benchmark, broken down across five compositional variation types: Swap Object, Swap Attribute, Replace Relation, Replace Object, and Replace Attribute. Each variation is evaluated under two settings: ITT (Image-to-Text retrieval) and TOT (Text-Only Transfer), reflecting different forms of generalization stress.

Overall, we observe that models like NegCLIP and TripletCLIP maintain relatively strong performance across both ITT and TOT settings, while our SPEC model, although competitive in overall specificity evaluation, exhibits lower compositional generalization performance. This is consistent with earlier analysis in Section 5, and supports the claim that specificity-oriented fine-tuning does not necessarily improve compositional reasoning.

Model	Swap Object		Swap Attribute		Replace Relation		Replace Object		Replace Attribute		Avg.
	ITT	TOT	ITT	TOT	ITT	TOT	ITT	TOT	ITT	TOT	
CLIP	45.18	19.74	45.21	33.03	56.26	38.62	86.80	83.72	65.61	59.14	53.31
Long-CLIP	42.85	15.10	49.39	31.98	55.68	40.54	90.19	87.71	71.31	59.77	54.45
Long-CLIP*	46.53	28.97	46.99	42.64	52.20	39.68	88.31	91.82	66.37	63.95	56.74
SigLIP	36.32	5.71	30.63	9.00	27.24	12.66	35.16	12.71	30.71	8.62	20.88
NegCLIP	55.25	34.65	57.99	56.47	52.27	51.57	89.53	94.55	69.41	76.27	63.79
DCI	44.10	31.80	45.60	38.00	43.20	35.70	80.20	81.20	60.90	52.20	51.29
DAC	27.80	11.40	33.50	25.40	48.60	48.60	64.30	75.80	44.00	56.00	43.54
La-CLIP	41.22	21.22	48.95	36.04	51.07	42.03	86.44	88.50	68.78	65.61	54.99
TripletCLIP	38.37	18.78	44.44	38.14	58.68	48.08	85.05	89.04	65.61	70.94	55.71
SPECS	30.61	16.73	28.37	24.02	25.96	24.25	48.36	73.91	38.57	45.30	35.61

Table 6: **Compositional Generalization Evaluation.** ITT and TOT denote image-to-text task and text-only task accuracy, respectively.