Redemption Score: A Multi-Modal Evaluation Framework for Image Captioning via Distributional, Perceptual, and Linguistic Signal Triangulation

Anonymous WACV Algorithms Track submission

Paper ID *****

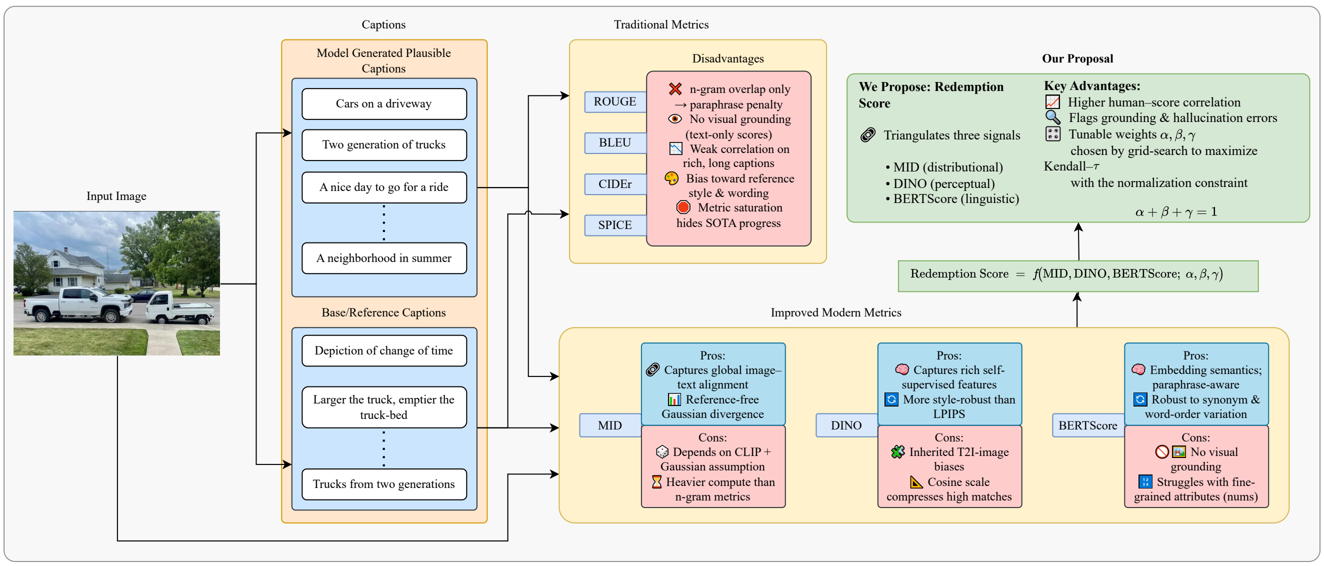


Figure 1. Abstract figure depicting the rationale for the study. Left hand side shows the inputs towards various metric systems whereas the proposed method is highlighted on green on the right hand side. Advantages and disadvantages of current metrics are highlighted on light blue and red respectively.

Abstract

Evaluating image captions requires cohesive assessment of both visual semantics and language pragmatics, which is often not entirely captured by most metrics. We introduce Redemption Score(RS), a novel hybrid framework that ranks image captions by triangulating three complementary signals: (1) Mutual Information Divergence (MID) for global image-text distributional alignment, (2) DINO-based perceptual similarity of cycle-generated images for visual grounding, and (3) LLM Text Embeddings for contextual text similarity against human references. A calibrated fusion of these signals allows RS to offer a more holistic assessment. On the Flickr8k benchmark, RS achieves a Kendall-τ of 58.42, outperforming most prior methods and demonstrating superior correlation with human judgments without requiring task-specific training. Our framework provides a more robust and nuanced evaluation by thoroughly examining both the visual accuracy and text quality together, with consistent performance across Conceptual Captions and MS COCO.

1. Introduction

It is most commonly acknowledged that a picture is worth a thousand words, but the pragmatics of image captioning focus on finding the most plausible set of 10-20 words that best justifies describing its complexities. This precipitates a unique problem on linguistics deeply rooted to subjectivity. It dictates that two highly plausible captions grounded to the same image may lead to drastic degradation of traditional scoring metrics like BLEU [23], BERTScore[36], METEOR [2] or ROUGE [19]. This necessitates an evaluation framework for caption evaluation which accounts for the semantic values of an image and the pragmatics of interpretability.

Recent advancements on vision-grounded NLP, shared embedding spaces, and strong multimodal systems have dawned new possibilities for vision-grounded language, alongside a proliferation of evaluation metrics to rank image captions. These metrics can be grouped into three broad families:

- **Surface-overlap scores** (BLEU-4 [23], METEOR [2], CIDEr [30] and SPICE [1]): emphasize n-gram or scene-graph agreement yet can penalise legitimate paraphrases.
- **Embedding-based or cross-modal measures** (BERTScore [36], CLIPScore [9] and Gaussian-assumed MID [12]): exploit pretrained representations but inherit modality biases and often struggle with fine-grained grounding.
- **Cycle-consistent approaches** (CAMScore) [5]: regenerate an image from the caption and compare it perceptually; sensitive to text-to-image artifacts and rely on a single synthetic perspective.

While each family illuminates part of the quality spectrum, none yields a fully reliable view of factual detail, linguistic nuance, and multi-reference visual grounding.

We introduce Redemption Score(RS), which recovers lost visual information by combining three complementary evaluation signals: distributional alignment (how well image-text pairs fit learned representations), perceptual grounding (visual consistency through image regeneration), and linguistic fidelity (contextual text similarity). By fusing these perspectives, our metric captures errors that any single signal alone would have otherwise been susceptible to miss. In summary, the key contributions of this paper are listed below:

- A training-free evaluation framework that addresses complementary failure modes in existing metrics, achieving 58.4% Kendall–τ correlation with human judgments.
- An efficient calibration procedure that optimizes fusion weights through constrained grid search to maximize human alignment while ensuring meaningful contribution from all modalities.
- Demonstration of cross-dataset generalization from Flickr8k to Conceptual Captions and MS-COCO without parameter retuning, indicating robust transferability.

The formal definition of each component and the fusion rule are detailed in Section 3.

2. Related Work

We organize existing image captioning evaluation metrics into three main categories based on their underlying approaches:

2.1. Historical Metrics

These metrics were based on surface overlap and are best suited and designed to evaluate machine translation and text summarization and hold no information regarding the "image" component of image captioning. Metrics like BLEU, METEOR, CIDEr and SPICE [1, 2, 23, 30] fall under this category. The limitations for these metrics include but not limited to penalize re-phrasings and paraphrasing, and lack of visual grounding.

2.2. Embedding and Cross-Modal Metrics

These are the second generation metrics based on multimodal or unimodal embeddings shift and similarities. Unimodal text-only variants include BERTScore, whereas multi-modal variants like CLIPScore, ViLBERTScore, UMIC, MID, and Polos [9, 13, 14, 31, 36] metric share a common latent space for texts and images. Although these metrics improve semantic awareness, they still cease to resist biases inherited from encoder, fine-grained object attributes and inherit a persistent modality gap in which spatial relation and numeracy are encoded differently even in the shared vision and language channels.

2.3. Cycle Consistent and Judge Models

These are the most recent works in vision grounded NLP metrics which attempt to circumvent the modality gap in its entirety. The idea of using label free metric by employing the diffusion models to complete the image to text to image cycle was first introduced by Huang et al [11]. Metrics like CAMScore [5] regenerates an image from candidate caption via a text-to-image (T2I) model and compares the synthetic image to the original image using perceptual and object-level criteria. Likewise, VLM/LLM-as-a-Judge [4, 15] frameworks employ the LLM foundation to assess caption quality directly through scalar scoring or pairwise preference. Furthermore, chain of thought reasoning was used by Tong et. al in [29] to better evaluate the image captions for large multimodal models. Although these metrics achieve the highest known correlation with human judgements on long and detailed captions, cycle consistency inherits the biases and failures of T2I models, while the LLM-as-a-Judge models introduce heavy computational cost and calibration drift.

2.4. Research Gap and Our Work

Across the three diverse group of metrics, two key gaps are persistent: (i) **inclusive coverage**: no single metric captures global image to text alignment, local visual grounding and linguistic fidelity jointly and (ii) **robustness**: metric often inherit systematic biases from their underlying models and reference corpora leading to misleading ranking in exception cases. To address these gaps, we propose Redemption Score - a training-free framework that mitigates individual metric biases through complementary signal integration. Our approach combines MID for distributional alignment, DINO similarity [3] for visual consistency through cycle-generated images, and general text embeddings for semantic understanding.

WACV
#*****

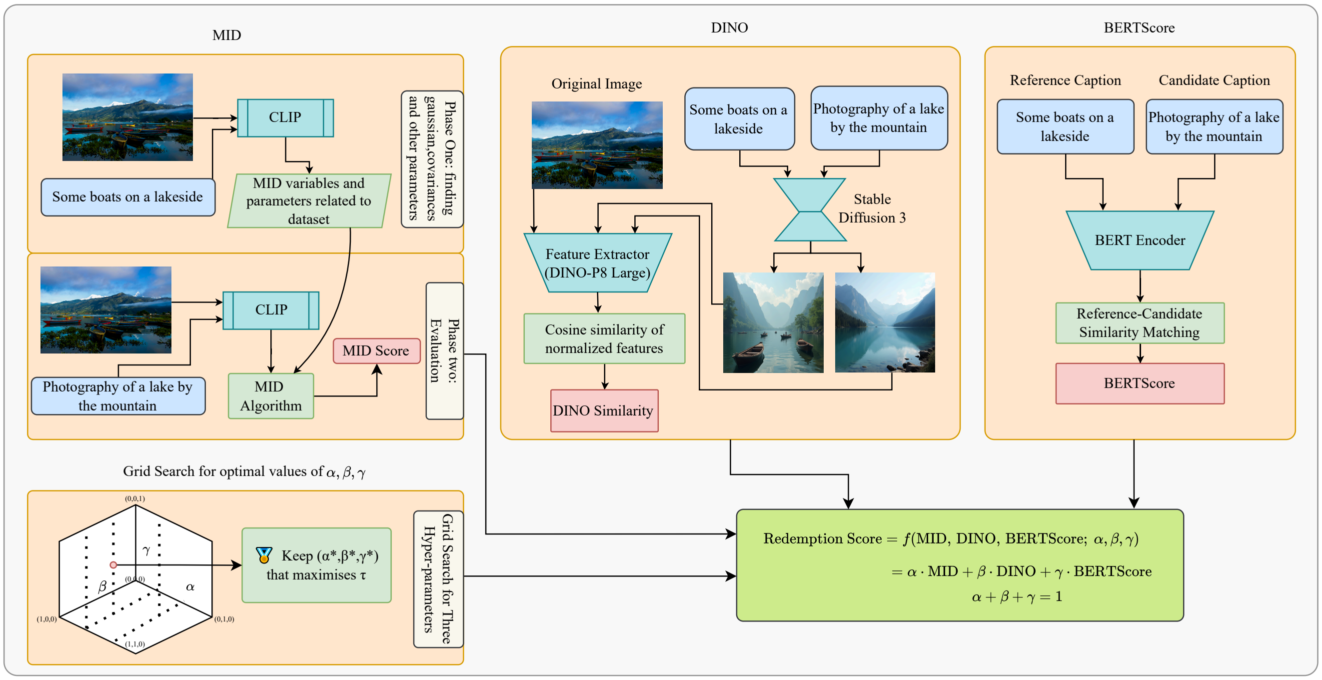


Figure 2. Overview of calculation of Redemption Score.

3. Redemption Score

An overview of calculation of Redemption Score is highlighted on Fig. 2 which is in-detail described in the following sections.

These three components address different failure modes in caption evaluation: MID captures whether image-caption pairs follow expected distributional patterns (detecting statistical outliers), DINO similarity identifies visual inconsistencies through cycle generation, and GTE embeddings catch semantic mismatches that surface-level metrics miss. No single component can reliably detect all caption quality issues.

3.1. Mutual Information Divergence (MID)

We adapt the Mutual Information Divergence (MID) from Kim et al. [12] to capture both image semantics and language pragmatics. Let $g(I)$ and $h(\hat{c})$ be the ℓ_2-normalized CLIP ViT-L/14 [25] embeddings of an image I and a candidate caption $\hat{c}$. Assuming these embeddings (random variables X for images, Y for captions) follow multivariate Gaussian distributions with means μ_x, μ_y and covariances Σ_x, Σ_y, the continuous mutual information is:

$$I(X;Y) = \tfrac{1}{2} \log \frac{\det(\Sigma_x)\det(\Sigma_y)}{\det(\Sigma_{xy})} \tag{1}$$

where Σ_{xy} is the joint covariance.

Kim et al. [12] define a point-wise mutual information (PMI) for individual pairs (x, y), which adjusts $I(X;Y)$ using Mahalanobis distances to reflect how well the pair aligns with the learned distributions. We redefine the PMI normalization on Eq. (7). The final MID score is the expectation of these PMI values:

$$\text{MID}(I, \hat{c}) = \mathbb{E}_{(x,y)\sim p}[\text{PMI}(x, y)] \tag{2}$$

The necessary distributional statistics $(\mu_x, \Sigma_x, \mu_y, \Sigma_y, \Sigma_{xy})$ are estimated once from the entire dataset (details in §4.1).

3.2. DINO Perceptual Similarity

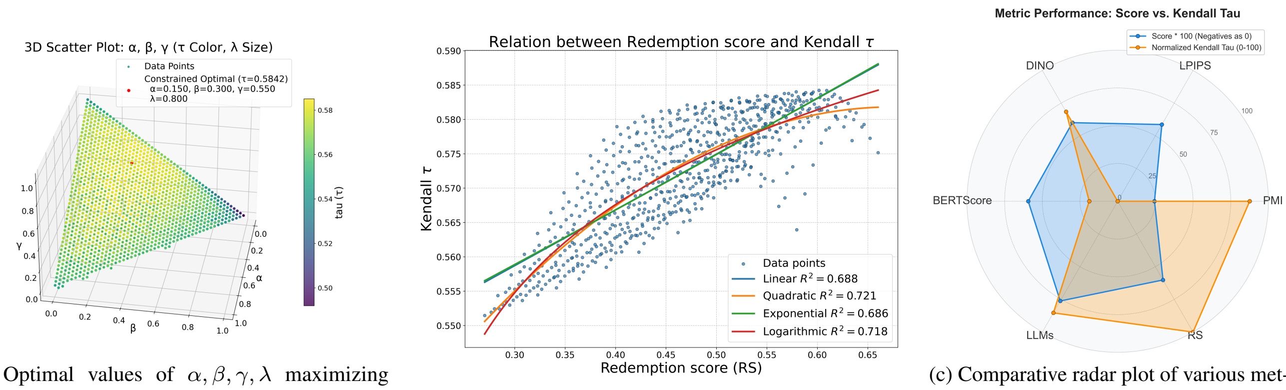
The perceptual component leverages the self-supervised DINO vision transformer [3]. This component captures visual consistency through cycle generation and comparison. For every example we consider the original image I, its human reference caption c^{ref}, and the model-generated candidate caption $\hat{c}$. Using the public Stable Diffusion 3 checkpoint [26], we synthesize two proxy images $\tilde{I}_{\text{ref}}$ and $\tilde{I}_{\text{cand}}$. All images are resized to 224×224 and ImageNet-normalised before encoding.

Feature extraction. A frozen `ViT-B/8` DINO encoder $E(\cdot) \in \mathbb{R}^{768}$ produces a `[CLS]` embedding for every image. Embeddings are ℓ_2-normalised: $\hat{e}(X) = E(X)/\|E(X)\|_2$.

Cosine similarities. We compute two edge scores

$$s_1 = \langle \hat{e}(I),\, \hat{e}(\tilde{I}_{\text{cand}}) \rangle, \tag{3}$$

$$s_2 = \langle \hat{e}(\tilde{I}_{\text{cand}}),\, \hat{e}(\tilde{I}_{\text{ref}}) \rangle. \tag{4}$$



(a) Optimal values of $\alpha, \beta, \gamma, \lambda$ maximizing Kendall-τ.

(b) RS vs. Kendall-τ relationship with fitted models.

(c) Comparative radar plot of various metrics' scores and Kendall-τ.

Figure 3. Comprehensive analysis of Redemption Score (RS) parameters, its relationship with Kendall-τ, and comparison with other metrics. Comparison with more metrics is shown on Tab. 3

Aggregate perceptual score. The final RS receives the mean of these two scores.

$$\mathrm{DINO}_{\mathrm{sim}} = \tfrac{1}{2}\big(s_1 + s_2\big). \tag{5}$$

3.3. LLM Embeddings

We adopt General Text Embeddings [18] with the thenlper/gte-large encoder to account for language pragmatics. Let $\mathbf{e}(x) \in \mathbb{R}^d$ denote the (L2-normalized) GTE embedding of text x produced by the thenlper/gte-large encoder. For a candidate caption $\hat{c}$ and its reference caption c^{ref}, we compute the GTEScore as the cosine similarity between their embeddings:

$$\mathrm{GTEScore} = \cos\big(\mathbf{e}(\hat{c}),\, \mathbf{e}(c^{\mathrm{ref}})\big) = \frac{\mathbf{e}(\hat{c})^{\top}\mathbf{e}(c^{\mathrm{ref}})}{\|\mathbf{e}(\hat{c})\|_2\,\|\mathbf{e}(c^{\mathrm{ref}})\|_2}. \tag{6}$$

3.4. Redemption Score Aggregation

Normalization. To ensure all three signals are directly comparable and compatible with the multiplicative component, we normalize all scores to the unit interval $[0, 1]$ using the transformation:

$$X_{normalized} = \frac{\frac{X}{1+|X|} + 1}{2} \tag{7}$$

where X represents the PMI component from MID (Equation 2), the raw DINO similarity score, or the GTE cosine similarity score. This transformation ensures all signals are positive and bounded in [0,1], with larger input values mapping to larger output values, enabling meaningful aggregation across the three heterogeneous modalities.

Aggregation. The final *Redemption Score (RS)* uses a hybrid aggregation that interpolates between linear combination and weighted geometric mean of the three normalized signals, with learnable weights α, β, γ and interpolation parameter λ. This hybrid approach combines the robustness of additive aggregation with the strict quality control of multiplicative aggregation, providing the benefits of both strategies.

$$RS[i] = \lambda \cdot L[i] + (1-\lambda) \cdot M[i] \tag{8}$$

where:

- $L[i] = \alpha Z_{mid}[i] + \beta Z_{dino}[i] + \gamma Z_{bert}[i]$ (linear component)
- $M[i] = Z_{mid}[i]^{\alpha} Z_{dino}[i]^{\beta} Z_{bert}[i]^{\gamma}$ (multiplicative component)

with constraints $\alpha + \beta + \gamma = 1$ and $\alpha, \beta, \gamma > 0$.

Parameter optimization. We employ a constrained optimization approach to determine optimal parameters for our proposed metric. Our methodology addresses two key objectives: (1) maximizing correlation with human judgments, and (2) ensuring meaningful contribution from all metric components.

We implement minimum weight constraints ($\alpha, \beta, \gamma \geq 0.15$) to prevent degenerate solutions where any single modality dominates the final score. This threshold ensures that each modality— distributional alignment(PMI), visual similarity (DINO), and contextual understanding (GTE embeddings)—contributes meaningfully to the evaluation while allowing data-driven optimization to determine their relative importance.

The constrained grid search explores the parameter space with step size $\Delta = 0.05$ for weights and $\Delta = 0.1$ for the interpolation parameter λ.

Parameter selection methodology. We select parameters using the following criteria:

Table 1. Metric values and Kendall–τ_c correlation on raw scores without normalization on Flickr8k dataset.

	BLEU-4		METEOR		MID		DINO		BERTScore		GTEScore		RS	
Dataset	Val	τ	Val	τ	Val	τ	Val	τ	Val	τ	Val	τ	Val	τ
Flickr8k	0.0465	33.5	0.2441	35.85	-17.55	**54.6**	0.268	48.76	0.59	38.05	0.76	53.93	0.48	**58.4**

Optimization objective: We identify the parameter combination that maximizes Kendall's τ correlation with human judgments while satisfying the minimum weight constraint ($\alpha, \beta, \gamma \geq 0.15$).

Statistical significance: All selected parameters must achieve statistical significance ($p < 0.05$) in their correlation with human ratings.

Robustness validation: We perform sensitivity analysis around the optimal parameters to ensure stable performance under small perturbations.

The resulting parameter configuration reflects the empirical evidence for optimal modality weighting as determined by human evaluation data, constrained by methodological requirements for multi-modal contribution.

4. Experiments

4.1. Dataset

We optimized our Redemption Score on the Flickr8k dataset [10], using its full validation set (5,822 images) and human preference data to maximize Kendall-τ correlation. We exclude 158 image-caption pairs where candidate captions were exactly the reference captions. This process yielded the score's weights $(\alpha, \beta, \gamma, \lambda)$ (see Eq. (8)). For evaluating the generalizability of the Redemption Score with these predetermined weights, we further evaluated on image-caption pairs from each Conceptual Captions [28] and MS-COCO [20]. This approach was adopted as Conceptual Captions lacks human preference data for direct Kendall-τ optimization, while in the case of MS-COCO we did not use available preference annotations to ensure a consistent, model-based evaluation across both datasets. The sample size also keeps DINO processing feasible (under one GPU-day).

Workload Summary. The DINO Similarity component required generating approximately 36,000 proxy images (1 reference + 5 candidates generated by models in Sec. 4.2 for $3,000$ samples across the 2 datasets). Adding the base image comparisons, DINO similarity processed around 42,000 images. In contrast, MID operated directly on image-caption pairs, resulting in roughly 60,000 total comparisons: (40,000 from the full Flickr8k split and 20,000 from the Conceptual Captions and COCO validation subsets). GTEScore operated on the same captions as DINO Similarity resulting to 36,000 total text-text comparisons.

4.2. Captioning Systems

The Conceptual Captions dataset was used to QLoRA [6] fine-tune 5 popular multimodal systems: (i) BLIP [16], (ii) BLIP2 2.7B [17], (iii) MS-GIT [32], (iv) ViT-GPT-2 [7, 24] and (v) Qwen 2.5-VL 7B [33]. We trained captioning models on the Conceptual Captions dataset, which is sourced from web alt-text and thus provides captions with greater stylistic diversity and less rigidly descriptive tendencies, offering broader generalization potential but also posing challenges for methods assuming literal image–text alignment [22, 28]. All models were fine-tuned for one epoch with an effective learning rate of 5e-5, QLoRA rank of 8 and quantization of 4-bits on their respective predefined-loss functions. We observed that larger LLM-style models such as Qwen tended to generate meta-statements (e.g., "I can think of a few different captions") and enumerate multiple caption options, which is undesirable for standard captioning tasks. To mitigate this behavior, we fine-tuned the larger models, and to ensure fairness across baselines, we applied the same fine-tuning procedure to all captioning systems.

4.3. Results

Finding $\alpha, \beta, \gamma, \lambda$. The optimal values of $\alpha, \beta, \gamma, \lambda$ were found to be 0.15, 0.30, 0.55, and 0.8 respectively via grid search as described in Sec. 3.4. The grid search and optimal value is highlighted in Fig. 3a.

Results on Flickr8k. Table 1, Tab. 3 and Fig. 3 summarizes the findings on the Flickr8k dataset. First of all, it is important to form our understanding that the Flickr8k dataset have some random mapping of captions to images which would lead to a lot of images being paired with unrelated caption thus achieving lower human rating. Figure 3c, Fig. 3b and Tab. 1 show this relationship clearly as the higher score value in the metrics relative scale have amounted to worse kendall-τ scores with human evaluators. Overall, we were able to achieve a kendall-τ score of 58.42 on the dataset which beats traditional and most of the recent methods in Tab. 1.

Qualitative Evaluation Figure 4 demonstrates how individual RS Score components fail in complementary ways, which our combined metric successfully addresses. MID assigns unreasonably low scores to valid captions due to

WACV #*****

Table 2. Metric scores per captioning model across Conceptual Captions and COCO. RS uses fixed $(\alpha, \beta, \gamma, \lambda) = (0.15, 0.35, 0.5, 0.8)$ (tuned on Flickr8k).

Model	METEOR	CIDEr	ROUGE-L	SPICE	CLIP-S	DINO	MID	BERTScore	GTEScore	RS
Conceptual Captions Dataset										
Qwen2-VL-7B	0.10	0.05	0.14	0.08	0.64	0.52	30.0	0.52	0.85	0.80
GIT (base)	0.07	0.36	0.20	0.12	0.60	0.38	17.0	0.47	0.83	0.77
ViT-GPT2	0.07	0.25	0.17	0.09	0.59	0.39	12.7	0.53	0.81	0.74
BLIP	0.12	0.98	0.28	0.20	0.62	0.48	30.5	0.60	0.86	0.81
BLIP-2	0.13	0.97	0.29	0.20	0.63	0.49	32.3	0.60	0.87	0.81
COCO Dataset										
Qwen2-VL-7B	0.16	0.01	0.21	0.11	0.64	0.60	30.2	0.60	0.88	0.82
GIT (base)	0.07	0.29	0.19	0.13	0.60	0.44	14.3	0.47	0.84	0.75
ViT-GPT2	0.15	0.82	0.34	0.21	0.61	0.55	26.8	0.65	0.87	0.81
BLIP	0.12	0.69	0.28	0.21	0.62	0.56	29.6	0.61	0.87	0.81
BLIP-2	0.12	0.59	0.26	0.18	0.63	0.56	29.9	0.57	0.87	0.81

Table 3. Flickr8K-Expert scores and qualitative properties. Δ indicates the difference from our Redemption score ($\tau = 58.4$). Positive values (green) indicate our score is higher. RT=Requires Training, IG=Image Grounded

Metric	Flickr8K (Expert)	RT	IG	Δ
ROUGE [19]	32.3	✗	✗	(+26.1)
CIDEr [30]	43.9	✗	✗	(+14.5)
SPICE [1]	44.9	✗	✗	(+13.5)
SPARCS [8]	48.1	✗	✗	(+10.3)
MoverScore [37]	46.7	✗	✗	(+11.7)
BARTScore [34]	37.8	✗	✗	(+20.6)
UMIC [14]	46.8	✓	✓	(+11.6)
ViLBERTScore [13]	50.1	✓	✓	(+8.3)
CLIP-S [9]	51.2	✗	✓	(+7.2)
RefCLIP-S [9]	53.0	✗	✓	(+5.4)
FLUER [15]	53.0	✗	✓	(+5.4)
PAC-S [27]	54.3	✓	✓	(+4.1)
RefPAC-S [27]	55.9	✓	✓	(+2.5)
Polos [31]	56.4	✗	✓	(+2.0)
DENEB [21]	56.8	✓	✓	(+1.6)
RefHICE-S [35]	57.7	✗	✓	(+0.7)
G-VEval-ref-free [29]	59.7	✗	✓	(-1.3)
RS (ours)	**58.4**	✗	✓	—

distributional modeling limitations. DINO penalizes captions too harshly for minor visual inaccuracies that humans find acceptable. LLM embeddings give high scores to linguistically fluent but visually misaligned captions. In each case, the integrated RS Score compensates for these individual weaknesses, producing scores that better match human judgments by leveraging the strengths of other components.

Statistical Robustness Analysis To assess the stability of our correlation results, we performed 1000-run bootstrap analysis on Flickr8k. The results are presented in Table 4, which shows that our proposed metric exhibits excellent stability ($\sigma = 0.43\%$), confirming the statistical robustness of our findings. The Redemption Score consistently achieves the highest correlation ($58.42 \pm 0.43\%$) with the most stable performance across bootstrap samples.

Table 4. Statistical Robustness Analysis: 1000-run Bootstrap Results on Flickr8k

Metric	Mean (%)	Std Dev (%)	95% CI
RS	58.4	0.43	[57.6, 59.2]
PMI	54.6	0.50	[53.7, 55.6]
GTEScore	53.9	0.48	[53.0, 54.8]
CLIP	51.1	0.48	[50.2, 52.1]
DINO	48.8	0.48	[47.8, 49.7]
BERTScore	33.6	0.58	[32.6, 34.8]

Results on Conceptual Captions. Since human evaluations are not available for the Conceptual Captions dataset, we evaluate our framework's generalizability by applying the optimal parameters $(\alpha, \beta, \gamma, \lambda)$ learned from Flickr8k without retuning. This cross-dataset evaluation protocol tests whether RS captures fundamental aspects of caption quality that transfer across different visual domains and annotation styles. The consistent parameter performance across datasets would demonstrate that our metric learns generalizable representations of image-text alignment rather than dataset-specific artifacts.

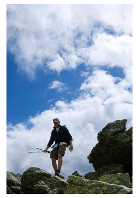
Figure 4. Complementary failure modes captured by different components. Each row shows cases where one component fails to detect caption quality issues that human raters identify, while our integrated RS Score successfully captures the problem through its multi-modal approach. Human ratings: 1-4 scale (higher = better quality), Metric scores: 0-1 scale (higher = better alignment).

PMI Score Failures

PMI gives unexpectedly low scores to captions that humans rate as reasonable, suggesting limitations in its distributional modeling approach

PMI	0.081
RS	0.649
Human	2.67

Caption: Man falling off a blue surfboard in the ocean.



PMI	0.114
RS	0.634
Human	3.00

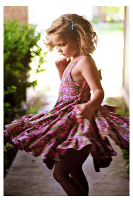
Caption: A rock climber climbs in between two very large rocks.

PMI	0.084
RS	0.593
Human	2.67

Caption: A man standing in front of a brick building.

DINO Score Failures

DINO penalizes captions too harshly for minor visual inaccuracies or missing details that humans consider acceptabley

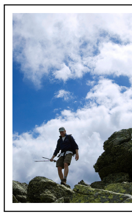


DINO	0.182
RS	0.746
Human	3.00

Caption: A young girl is wearing a purple shirt and pink headband.

DINO	0.177
RS	0.699
Human	3.00

Caption: A crowd of people at an outdoor event.

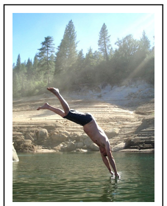


DINO	0.114
RS	0.697
Human	2.67

Caption: Two men holding their fishing poles.

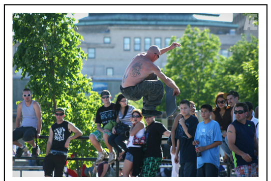
GTEScore Failures

General Text Embeddings overweight linguistic fluency, missing visual-semantic misalignments



GTEScore	0.664
RS	0.495
Human	1.67

Caption: Two dogs are jumping up at each other.



GTEScore	0.709
RS	0.500
Human	1.67

Caption: Three young adults look towards the camera in a school setting.

GTEScore	0.708
RS	0.490
Human	1.00

Caption: A little boy slides down a bright red corkscrew slide.

The results from this study is highlighted on Tab. 2. As noted previously on Sec. 4.2, all 5 image captioning multi-modal systems were fine tuned on the Conceptual Captions dataset thus leading to a higher Redemption Score, especially on those models which are known for their SOTA performance on image captioning.

Results on MS COCO. A similar framework of evaluation was employed to assess Redemption Score's capacity of consistently portraying the model's capacity to caption images. Since the models were fine-tuned over the Conceptual Captions dataset and the captioning style and distribution of the Conceptual Captions and COCO are slightly different, the exact numbers were not replicated but the general trend of Qwen2-VL and the BLIP-2 leading the pact was observed in Tab. 2. Only slight changes in the score's trend was observed on GIT and ViT-GPT2 but this could be attributed towards the comparatively smaller models overfitting to the fine-tuned dataset.

4.4. Ablation Study

Purely Multiplicative Strategy. To validate our hybrid aggregation approach in Eq. (8), we test a purely multiplicative strategy without the additive component ($\lambda = 0$):

$$\hat{RS}(I, \hat{c}; R) = MID^{\alpha} \cdot DINO_{sim}^{\beta} \cdot GTEScore^{\gamma} \quad (9)$$

$$\alpha + \beta + \gamma = 1; \alpha, \beta, \gamma \geq 0$$

Best Weights (α, β, γ)	Kendall's τ	Mean Score
(0.150, 0.200, 0.650)	0.5713	0.5141

Purely Additive Strategy. We also test a purely additive approach without the multiplicative component ($\lambda = 1$):

$$\hat{RS}(I, \hat{c}; R) = \alpha \cdot MID + \beta \cdot DINO_{sim} + \gamma \cdot GTEScore \tag{10}$$

$$\alpha + \beta + \gamma = 1; \alpha, \beta, \gamma \geq 0$$

Best Weights (α, β, γ)	Kendall's τ	Mean Score
(0.350, 0.250, 0.400)	0.5826	0.5405

Comparative Analysis. Table 5 presents a statistical comparison of the three approaches based on 1000-run bootstrap analysis. The column *Mean* reports the average τ across bootstrap runs, $\boldsymbol{\sigma}$ is the corresponding standard deviation, *95% CI* denotes the confidence interval, and *Width* is the span of that interval. *Imp.* indicates relative improvement over the baseline.

Table 5. Ablation Study: Robustness Comparison

Approach	**Mean**	$\boldsymbol{\sigma}$	**95% CI**	**Width**	**Imp.**
Hybrid	**0.5839**	**0.0043**	**[0.5750, 0.5923]**	**0.0174**	**Base**
Additive	0.5826	0.0046	[0.5734, 0.5915]	0.0181	-0.22%
Multiplicative	0.5713	0.0045	[0.5624, 0.5803]	0.0179	-2.21%

The hybrid method outperforms both additive and multiplicative approaches, achieving the highest mean τ (0.5839), the narrowest CI width (0.0174), and improvements of +0.22% and +2.21% respectively.

Metric Choice To ensure our metric selection is both principled and robust, we conducted an exhaustive ablation over all 20 possible three-metric combinations from PMI, LLM embeddings, DINO, BERT, LPIPS, and CLIP. For LPIPS, we use a normalized variant $LPIPS_{norm} = \frac{1}{1+LPIPS}$ to maintain comparability across scales. As shown in Table 6, our chosen combination (PMI + LLM + DINO) achieves the highest correlation with human judgments ($\tau = 0.584$) while maintaining a low standard deviation, demonstrating both accuracy and stability. Importantly, most of the top-performing combinations draw from different buckets of evaluation aspects: PMI provides distributional alignment, LLM embeddings and BERT capture semantic grounding, DINO and LPIPS measure visual fidelity, and CLIP captures cross-modal alignment. This pattern highlights that the strongest evaluators are not those relying on a single perspective, but those that integrate complementary dimensions of textual meaning, visual similarity, and multimodal consistency. By explicitly combining metrics across these buckets, our selected approach provides the most comprehensive and reliable evaluation signal.

Table 6. Comprehensive Results Table

Metric Combination	Weights (α, β, γ)	λ	Kendall τ	Std Dev
PMI + LLM + DINO	**0.15, 0.50, 0.35**	**0.80**	**0.584**	**0.106**
PMI + LLM + LPIPS	0.20, 0.50, 0.30	0.80	0.581	0.118
PMI + LLM + BERT	0.15, 0.65, 0.20	0.60	0.578	0.125
PMI + LLM + CLIP	0.15, 0.65, 0.20	0.00	0.574	0.386
LLM + DINO + CLIP	0.55, 0.25, 0.20	0.00	0.572	0.144
LLM + DINO + LPIPS	0.50, 0.30, 0.20	1.00	0.567	0.041
LLM + LPIPS + CLIP	0.65, 0.15, 0.20	0.00	0.567	0.140
PMI + DINO + CLIP	0.15, 0.60, 0.25	0.00	0.567	0.402
LLM + BERT + CLIP	0.65, 0.15, 0.20	0.00	0.565	0.159
PMI + BERT + DINO	0.40, 0.15, 0.45	0.70	0.564	0.196
LLM + BERT + DINO	0.50, 0.15, 0.35	1.00	0.563	0.047
PMI + BERT + CLIP	0.15, 0.40, 0.45	0.00	0.563	0.907
PMI + LPIPS + CLIP	0.15, 0.50, 0.35	0.00	0.562	0.556
PMI + DINO + LPIPS	0.50, 0.30, 0.20	0.80	0.561	0.216
PMI + BERT + LPIPS	0.55, 0.15, 0.30	0.70	0.556	0.237
LLM + BERT + LPIPS	0.60, 0.15, 0.25	1.00	0.549	0.043
BERT + DINO + CLIP	0.15, 0.60, 0.25	0.00	0.545	0.173
DINO + LPIPS + CLIP	0.60, 0.15, 0.25	0.00	0.543	0.155
BERT + LPIPS + CLIP	0.25, 0.55, 0.20	0.10	0.528	0.203
BERT + DINO + LPIPS	0.15, 0.65, 0.20	1.00	0.498	0.038

4.5. Efficiency Considerations

The DINO similarity component requires generating synthetic images via Stable Diffusion-3, which increases computational cost compared to embedding-only metrics. However, this generation step is performed offline during evaluation and can be parallelized across available GPUs. While computationally intensive, this approach enables more robust visual grounding assessment that purely embedding-based methods cannot provide.

5. Conclusion and Future Works

Conclusion. We introduce Redemption Score, a multimodal evaluation framework that achieves 58.42% Kendall-τ correlation with human judgments on Flickr8k, outperforming existing methods. RS addresses individual metric limitations through calibrated parameter optimization and demonstrates robust generalization across Conceptual Captions and MS-COCO datasets without requiring task-specific training.

Future Works. Key extensions include: (i) *Multilingual support* across diverse cultural contexts, (ii) *Computational efficiency* through model distillation and reduced image generation requirements, (iii) *Temporal Grounding:* expand Redemption Score from image-text to video-text domain.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 2, 6, 12

[2] 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*, pages 65–72, 2005. 1, 2, 12

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 2, 3

[4] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024. 2, 12

[5] Tianyu Cui, Jinbin Bai, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ye Shi. Evaluating image caption via cycle-consistent text-to-image generation. *arXiv preprint arXiv:2501.03567*, 2025. 2, 12

[6] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314, 2023. 5

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[8] Joshua Feinglass and Yezhou Yang. SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2250–2260, Online, 2021. Association for Computational Linguistics. 6

[9] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, pages 7514–7528, 2021. 2, 6, 12

[10] 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. 5

[11] Jia-Hong Huang, Hongyi Zhu, Yixian Shen, Stevan Rudinac, and Evangelos Kanoulas. Image2text2image: A novel framework for label-free evaluation of image-to-text generation with text-to-image diffusion models. In *International Conference on Multimedia Modeling*, pages 413–427. Springer, 2025. 2, 12

[12] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual information divergence: A unified metric for multimodal generative models. In *Advances in Neural Information Processing Systems*, pages 35072–35086. Curran Associates, Inc., 2022. 2, 3, 12

[13] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. ViLBERTScore: Evaluating image caption using vision-and-language BERT. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39, Online, 2020. Association for Computational Linguistics. 2, 6, 12

[14] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. Umic: An unreferenced metric for image captioning via contrastive learning. *ACL*, 2021. 2, 6, 12

[15] Yebin Lee, Imseong Park, and Myungjoo Kang. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. *arXiv preprint arXiv:2406.06004*, 2024. 2, 6, 12

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 5

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 5

[18] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 4

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1, 6

[20] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 5

[21] Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. Deneb: A hallucination-robust automatic evaluation metric for image captioning. In *Proceedings of the Asian Conference on Computer Vision*, pages 3570–3586, 2024. 6

[22] Edwin G Ng, Bo Pang, Piyush Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. *arXiv preprint arXiv:2012.02339*, 2020. 5

[23] 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*, pages 311–318, 2002. 1, 2, 12

[24] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Website*, 2019. 5

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3

[26] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 3

[27] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6914–6924, 2023. 6

[28] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[29] Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7419–7427, 2025. 2, 6

[30] 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*, pages 4566–4575, 2015. 2, 6, 12

[31] Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568, 2024. 2, 6, 12

[32] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *ArXiv*, abs/2205.14100, 2022. 5

[33] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5

[34] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, pages 27263–27277. Curran Associates, Inc., 2021. 6

[35] Zequn Zeng, Jianqiao Sun, Hao Zhang, Tiansheng Wen, Yudi Su, Yan Xie, Zhengjue Wang, and Bo Chen. Hicescore: A hierarchical metric for image captioning evaluation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 866–875, 2024. 6

[36] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 1, 2, 12

[37] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 6