# DECAR: EVIDENCE-GROUNDED CLUSTERED CALIBRATION FOR FINE-GRAINED VISION–LANGUAGE ALIGNMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Fine-grained image–text scoring becomes difficult when captions differ only in a local attribute, relation, count, or factual detail. A model may rank such captions correctly while still assigning scores that disagree with the reference scale. We introduce DECAR, a continuous vision–language scoring framework trained on image-centered groups of graded captions together with element-level support. DECAR combines a distributional global score head, element-level prediction, within-image ranking constraints, and soft evidence bounds that act on implausible global scores during training. The element branch is removed at inference, leaving only the image–caption branch and global score head. On the Qwen2.5-VL-7B instantiation, DECAR achieves 0.7583 SRCC and 0.5864 MAE. In a separate prompt-disjoint frozen-CLIP study with matched training settings, clustered labels increase mean SRCC from 0.4660 to 0.5601 across three seeds. Target-shuffle and gap-aware controls further show that caption-specific graded labels mainly improve correlation, whereas additional ranking constraints reduce numerical error. Additional analyses examine paraphrase stability, counterfactual ordering, residual errors, and performance across semantic subsets.

## 1 INTRODUCTION

Vision–language matching is a core component of image retrieval, multimodal ranking, and text-to-image evaluation. In fine-grained scoring, two captions may both describe the same image but differ in a count, relation, attribute, or other localized fact. Such cases cannot be represented well by a binary match decision alone. The predicted score should also reflect the degree of image–text agreement. Earlier image–text matching methods learned shared embedding spaces or explicit region–word correspondences (Yan & Mikolajczyk, 2015; Lee et al., 2018; Li et al., 2019; Liu et al., 2020). Large-scale vision–language pretraining later improved transferable alignment through models such as CLIP, BLIP, BLIP-2, and recent VLMs (Radford et al., 2021; Li et al., 2022; 2023; Bai et al., 2025).

Most existing systems are trained or evaluated on individual image–text pairs. Our setting instead contains multiple descriptions associated with the same image, including close paraphrases, weaker matches, and counterfactual captions that alter a visually relevant fact. This structure provides more information than isolated pairs because the captions form a local semantic neighborhood around the same visual content. Compositional and counterfactual benchmarks have shown that vision–language models remain sensitive to small changes in relations, attributes, and other localized semantics (Thrush et al., 2022; Le et al., 2023). At the same time, continuous evaluation introduces a second requirement: predicted scores should agree with the reference scale, not only preserve the correct ordering among captions.

We address these two aspects jointly with DECAR. During training, captions associated with the same image are used as a structured supervision unit. A global distributional head predicts the image–text alignment score, and an auxiliary element head estimates support for selected local phrases. Within-image ranking constraints encourage the expected ordering among anchor, positive, weak-positive, and counterfactual captions. Cross-level coupling connects the global score to aggregated local evidence, and conditional evidence bounds penalize implausible global predictions when the local



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.

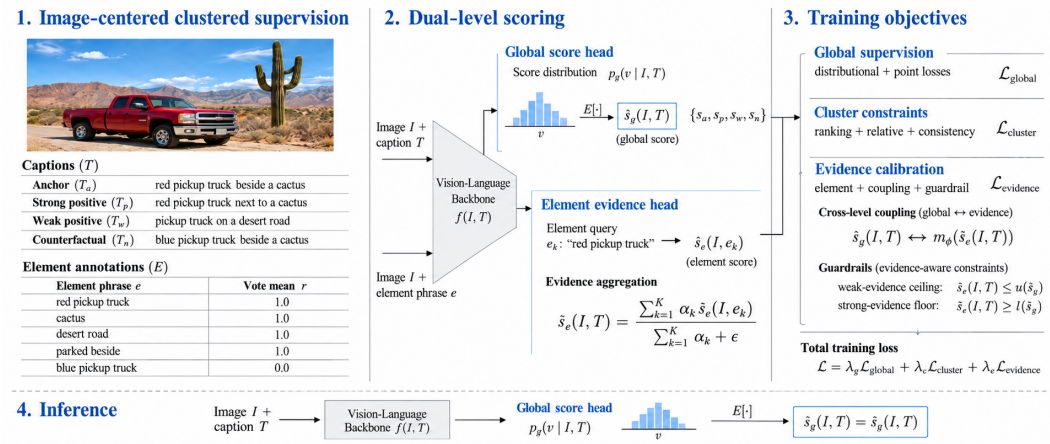


Figure 1: **Overview of DECAR.** Image-centered caption clusters supervise a global alignment branch and an auxiliary element-evidence branch. Training combines global score supervision, within-image cluster constraints, and evidence-based calibration. At inference, only the global image–caption scoring path is retained.

evidence is sufficiently confident. The element branch is used only during training; inference keeps the image–caption branch and global score head.

The main contributions are:

- We formulate fine-grained image–text scoring around image-centered groups of graded captions. This structure exposes local semantic differences among paraphrases, weaker matches, and counterfactual variants that are not available from isolated image–text pairs.
- We introduce DECAR, which combines distributional global scoring with element-level support, within-image ordering constraints, cross-level coupling, and conditional evidence bounds. The auxiliary element branch is removed at inference, so the final scorer remains an image–caption model with a global score head.
- We evaluate the method under both a Qwen2.5-VL instantiation and a prompt-disjoint frozen-CLIP control setting. Across three matched seeds, clustered labels increase mean SRCC from 0.4660 to 0.5601 over anchor-only training. Target-shuffle, bootstrap, ablation, and gap-aware controls further separate the effects of graded caption supervision, caption–score correspondence, and numerical score constraints.

## 2 Related Work

**Vision–language alignment and image–text matching.** Image–text matching evolved from shared latent spaces, hard-negative embeddings, and region–word attention (Yan & Mikolajczyk, 2015; Karpathy & Fei-Fei, 2015; Faghri et al., 2018; Lee et al., 2018; Chen et al., 2020) to relation-aware and structured reasoning, including VSRN, GSMN, SGRAF, NAAF, and CORA (Li et al., 2019; Liu et al., 2020; Diao et al., 2021; Zhang et al., 2022; Pham et al., 2024). Vision Transformers and large-scale vision–language pretraining further improved transferable representations through ViT, ALBEF, CLIP, BLIP, BLIP-2, and Qwen2.5-VL (Dosovitskiy et al., 2021; Li et al., 2021; Radford et al., 2021; Li et al., 2022; 2023; Bai et al., 2025). We focus on continuous alignment scoring while preserving the fine-grained distinctions studied in image–text matching.

**Fine-grained robustness, grounding, and evaluation.** Winoground and COCO-Counterfactuals expose failures under compositional or localized semantic changes (Thrush et al., 2022; Le et al., 2023). Flickr30K Entities and MS COCO provide region-level supervision for localized cross-modal reasoning (Plummer et al., 2015; Lin et al., 2014). Q-Align studies discrete-level visual scoring with large multimodal models (Wu et al., 2023); EvalMuse-40K provides fine-grained human judgments

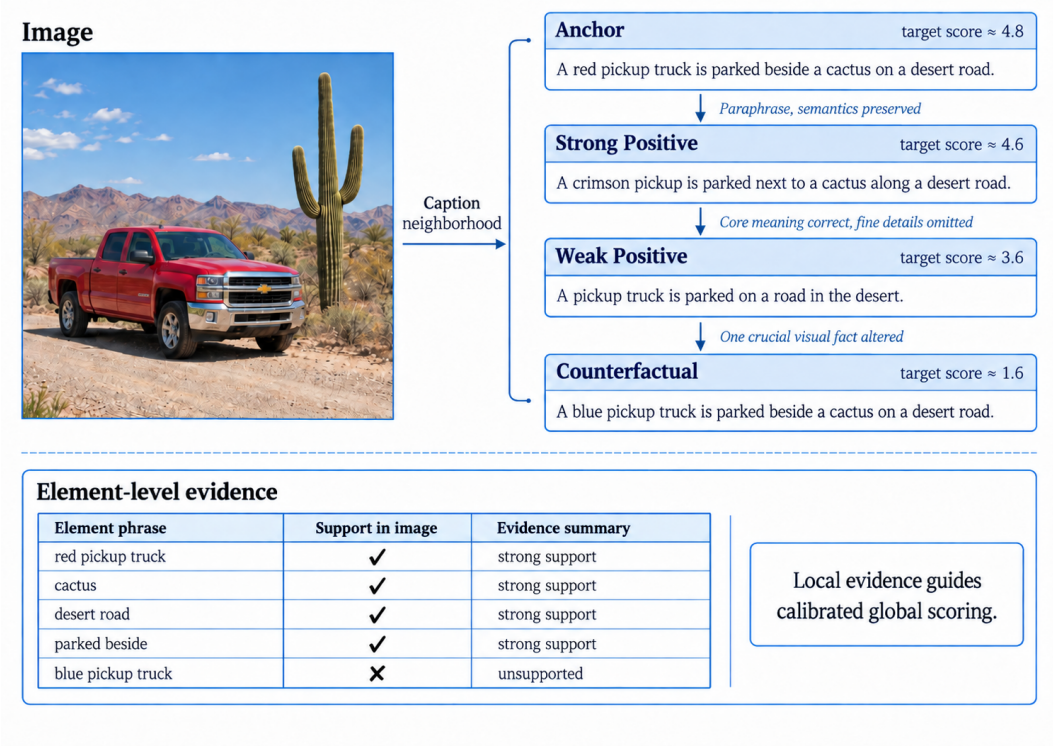


Figure 2: **Image-centered caption neighborhood.** Anchor, positive/weak-positive, counterfactual, and sampled element support jointly define the training supervision.

for text-to-image alignment evaluation (Han et al., 2024). Our training data combines graded caption scores with local element support.

**Ranking and calibration.** Retrieval losses are designed primarily for ordering. A continuous evaluator also has to match a reference score scale. Classical calibration work shows that confidence and ranking quality need not move together (Guo et al., 2017). In DECAR, graded captions provide the ordering constraints, and element support is used as a soft check on implausible score values.

# 3 Method

## 3.1 Problem setting and clustered supervision

For an image $I$ and caption $T$, let $y(T) \in [1, 5]$ denote the reference alignment score and $s_\theta(I, T) \in [1, 5]$ the model prediction. Each image defines

$$\mathcal{U}(I) = \{T_0, \mathcal{P}(I), \mathcal{N}(I), \mathcal{E}(I)\}, \tag{1}$$

where $T_0$ is the source caption, $\mathcal{P}(I)$ contains graded caption variants, $\mathcal{N}(I)$ contains counterfactuals, and $\mathcal{E}(I) = \{(e_j, r_j, c_j)\}$ contains element phrases, their vote means $r_j \in [0, 1]$, and confidence weights $c_j$. The Qwen instantiation samples four caption roles $\mathcal{B} = \{a, p, w, n\}$: anchor $T_a$, positive $T_p$, weak positive $T_w$, and counterfactual negative $T_n$. We write $y_b = y(T_b)$ and $s_b = \hat{s}^g(I, T_b)$. Exact caption sampling, element sampling, confidence computation, and hard-case weighting are detailed in Appendix A.

## 3.2 Dual-level distributional scoring

A pretrained vision–language backbone returns final-layer hidden states $H = \{h_1, \ldots, h_L\}$. We combine the attention-mask-weighted mean $\bar{h}$ and the last valid hidden state, followed by a shared

projection $g$,

$$z = g\big(\tfrac{1}{2}\bar{h} + \tfrac{1}{2}h_{\text{last}}\big)\,. \tag{2}$$

The global head predicts logits $o^g \in \mathbb{R}^{K_g}$ over $K_g = 15$ uniformly spaced centers $\{v_k^g\}_{k=1}^{K_g}$ on $[1, 5]$. Its scalar score is the expectation

$$\hat{s}^g(I, T) = \sum_{k=1}^{K_g} \mathrm{softmax}(o^g)_k v_k^g. \tag{3}$$

The element head is defined analogously with $K_e = 7$ centers on $[0, 1]$, producing $\hat{s}^e(I, e_j)$. Both heads use Gaussian-smoothed distributional targets in addition to Smooth-$L_1$ point supervision. The exact bin centers, Gaussian targets, and head architectures are listed in Appendix A.

### 3.3 Cluster and evidence constraints

The training objective follows the three groups summarized in Figure 1: global score supervision, within-image cluster constraints, and evidence-based calibration. The global branch is trained with distributional and pointwise score losses. Within each image-centered cluster, the main ordering constraint is

$$\mathcal{L}_{\text{rank}} = [1.0 - (s_a - s_n)]_+ + 0.5[0.25 - (s_p - s_w)]_+ \tag{4}$$

$$+ 0.3[0.45 - (s_w - s_n)]_+, \tag{5}$$

where $[x]_+ = \max(x, 0)$. Relative-score and anchor–positive consistency terms provide complementary supervision for score gaps and local stability; their exact forms are given in Appendix A.

For caption branch $b$ and sampled element $e_j$, let $\rho_{b,j}$ denote caption–element relevance and define $\alpha_{b,j} = \rho_{b,j} c_j$. Predicted element support is summarized by

$$\tilde{s}_b^e = \frac{\sum_j \alpha_{b,j} \hat{s}^e(I, e_j)}{\max(\sum_j \alpha_{b,j}, \epsilon)}, \qquad \epsilon = 10^{-6}. \tag{6}$$

A learned scalar map projects this aggregate to the global score range, and cross-level coupling encourages agreement between the global prediction and the resulting evidence summary. The element branch therefore acts as an auxiliary calibration signal rather than replacing the global scorer.

Target-side element evidence is aggregated in the same manner to define confidence-aware score bounds. Strong local support provides a soft lower bound on the global score, whereas weak local support provides a soft upper bound. These constraints are applied only when the aggregated evidence is sufficiently confident. Exact thresholds, branch weights, and penalty definitions are provided in Appendix A.

We group the complete objective as

$$\mathcal{L} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{cluster}} + \mathcal{L}_{\text{evidence}}, \tag{7}$$

where $\mathcal{L}_{\text{global}}$ supervises global score prediction, $\mathcal{L}_{\text{cluster}}$ preserves within-image caption structure, and $\mathcal{L}_{\text{evidence}}$ combines element supervision, cross-level coupling, and confidence-aware bounds. Appendix A gives the exact normalization, sample weighting, architectures, and loss coefficients. At inference, only the image–caption branch and global score head are retained, with $s_\theta(I, T) = \hat{s}^g(I, T)$.

## 4 Data and Evaluation Protocol

**Corpus and caption construction.** The corpus contains 32,717 image records from 2,991 prompt groups and 20 text-to-image generators. For each image, $T_0$ is the caption provided by the source dataset. An LLM generates semantically related auxiliary captions. Counterfactuals come from manual rewriting and LLM-assisted generation, preserving most of the original context while modifying a visually relevant fact. The sampler later assigns the positive and weak-positive branches from the positive pool according to the rules in Section 3 and Appendix A; the positive branch may include $T_0$. $T_n$ is drawn from the existing counterfactual pool, with the documented fallback used only when that pool is empty.

Table 1: Annotation statistics. About 10% of each caption category is scored directly by humans; GLM-4.5V assigns the remaining scores before human review. The 3/6-vote statistics apply only to the directly human-scored subset.

| Statistic | Value |
|---|---|
| Image-level records | 32,717 |
| Direct human-scored subset | $\sim$10% per caption category |
| Prompt groups / generators | 2,991 / 20 |
| Element vote groups | 165,224 |
| Element-label mean $\pm$ std. | $0.5724 \pm 0.3680$ |
| Labels in $\{0, .33, .50, .67, 1\}$ | 92.47% |
| Unanimous element groups | 48.36% |
| Human-subset vote length 3 / 6 | 76.43% / 23.57% |

Table 2: Fixed prompt-disjoint split used by the controlled frozen-backbone study. Caption counts are after within-image text deduplication.

| Split | Images | Prompt groups | Captions |
|---|---|---|---|
| Train | 26,088 | 2,392 | 684,785 |
| Validation | 3,298 | 299 | 84,518 |
| Test | 3,331 | 300 | 88,238 |

**Hybrid score annotation and human quality control.** Approximately 10% of each caption category is scored directly by human annotators. Each item in this subset receives three or six ratings, which are aggregated into its reference score. We place category-matched human examples in the prompt of the open-source GLM-4.5V model from Zhipu AI and use them as in-context demonstrations for the remaining items. GLM-4.5V is not fine-tuned for this task. It is also separate from the Qwen2.5-VL-7B-Instruct backbone used by DECAR. Human annotators review the propagated scores and correct clear image–caption mismatches in score. Items that pass this review are retained without a second three- or six-rater annotation round. We refer to the resulting targets as *human-reviewed reference scores*. Appendix B gives the construction and review details.

For the directly human-scored subset, 79.14% of within-item vote pairs differ by at most one point and the mean within-item standard deviation is 0.496. The corpus also contains 165,224 element vote groups. Exact pairwise agreement is 68.16%, and 48.36% are unanimous. These agreement statistics apply to the human-scored and element-vote subsets, not to the GLM-propagated scores. Element votes are used as auxiliary support, not as a complete symbolic description of the image.

**Evaluation protocols.** The Qwen experiment uses the original 90/10 prompt-grouped validation protocol. A second study freezes the backbone and uses a fixed prompt-disjoint train/validation/test split; all images and caption variants from one prompt stay in the same partition. This second protocol isolates the effect of the training labels under a shared representation. We report the two sets of results separately because their splits and training setups differ.

**Metrics and statistical reporting.** We report SRCC, PLCC, MAE, and RMSE. Frozen zero-shot baselines and one-dimensional BLIP score mapping are deterministic. Trained CLIP controls use seeds 42, 43, and 44, with checkpoint selection based on validation-anchor MAE. For seed 42, a prompt-group paired bootstrap resamples the 300 held-out prompt groups 5,000 times. These intervals describe test-group sampling uncertainty; the three-seed mean and standard deviation report training-run variation.

## 5 Experiments

We evaluate the complete Qwen2.5-VL model and use a separate prompt-disjoint frozen-backbone study to isolate the contribution of clustered training labels. As specified in Section 4, the two protocols use different splits and are reported separately.

Table 3: Qwen2.5-VL results for the DECAR instantiation. The full model adds evidence-grounded calibration to the dual-level model.

| Model | SRCC↑ | PLCC↑ | MAE↓ | RMSE↓ |
|---|---|---|---|---|
| Qwen2.5-VL-7B (zero-shot) | 0.4447 | 0.3960 | 1.4584 | 1.7098 |
| DECAR (dual-level) | 0.7526 | 0.7431 | 0.6512 | 0.7824 |
| DECAR (full) | **0.7583** | **0.7542** | **0.5864** | **0.7234** |

Table 4: Frozen-backbone reference results on the 3,331 anchor-caption test items with human-reviewed reference scores. Trained CLIP rows report mean ± sample standard deviation over seeds 42/43/44.

| Method | SRCC↑ | PLCC↑ | MAE↓ | RMSE↓ |
|---|---|---|---|---|
| CLIP zero-shot | 0.3621 | 0.3746 | – | – |
| BLIP-ITM zero-shot | 0.3957 | 0.3876 | – | – |
| BLIP-ITM + linear score map | 0.3957 | 0.4106 | 0.7927 | 0.9586 |
| Frozen CLIP + T0 Huber[†] | $0.4976 \pm 0.0003$ | $0.4892 \pm 0.0004$ | $0.7395 \pm 0.0002$ | $\mathbf{0.9242} \pm 0.0002$ |
| Frozen CLIP + cluster soft | $\mathbf{0.5601} \pm 0.0016$ | $\mathbf{0.5550} \pm 0.0014$ | $0.7398 \pm 0.0047$ | $0.9743 \pm 0.0067$ |
| + gap-aware ranking | $0.5592 \pm 0.0016$ | $0.5542 \pm 0.0012$ | $\mathbf{0.7242} \pm 0.0043$ | $0.9494 \pm 0.0066$ |

[†]The T0 Huber model is a separately implemented anchor-only reference. The exactly matched anchor-only control is reported in Table 5.

### 5.1 QWEN2.5-VL INSTANTIATION

The full model uses Qwen2.5-VL-7B-Instruct with LoRA on the attention projections and the dual-level heads described above. Training uses the 90/10 prompt-grouped protocol, and checkpoint selection is based on validation SRCC. Exact optimization, LoRA, evaluation, and sampling settings are listed in Appendix A.

The dual-level model already reaches 0.7526 SRCC. Adding the evidence bound raises SRCC to 0.7583 and reduces MAE from 0.6512 to 0.5864; RMSE falls by 0.0590. The larger change in MAE than in SRCC indicates that the guardrail primarily affects score magnitude.

SRCC rises from 0.4811 at step 100 to 0.7146 at step 800 and 0.7513 at step 2000. Prediction variance does not collapse during this period. Appendix E reports the checkpoint trace.

### 5.2 EXTERNAL BASELINES AND CONTROLLED SUPERVISION

CLIP ViT-B/32 provides a zero-shot cosine-similarity reference. BLIP-ITM uses `Salesforce/blip-itm-base-coco` with all backbone parameters frozen. Its score-mapped variant fits a one-dimensional linear map from the frozen ITM logit margin to the $[1, 5]$ reference scale using training anchors only. The controlled CLIP models reuse the same frozen ViT-B/32 representation, where $i$ and $t$ denote the frozen image and text embeddings, respectively, and the regression feature is $[i, t, i \odot t, |i - t|]$. "Cluster soft" trains on the supplied caption scores. For two within-image captions indexed by $u$ and $v$ with reference scores $y_u$ and $y_v$, "+ gap-aware ranking" adds a pairwise logistic loss when $|y_u - y_v| \geq 0.5$, weighted by $\min(|y_u - y_v|/2, 1)$ with coefficient 0.10.

BLIP-ITM has higher zero-shot SRCC than CLIP. Fitting a scalar map raises BLIP PLCC to 0.4106 but does not change its ranking. The T0 Huber row uses a separate training implementation, so we use it only as a reference. Adding gap-aware ranking to cluster-soft training changes correlation little and lowers both MAE and RMSE.

### 5.3 MATCHED CONTROLS AND TARGET-SHUFFLE ANALYSIS

To isolate the supervision effect, anchor-only, target-shuffled, and full cluster-soft models share the same frozen CLIP features, linear PyTorch head, AdamW optimizer, Smooth-$L_1$ objective, 10 training epochs, and validation selection rule. The target-shuffle condition preserves the same images,

Table 5: Exactly matched frozen-CLIP controls on the same 3,331 anchor-caption test items. Results report mean ± sample standard deviation over seeds 42, 43, and 44.

| Method | SRCC↑ | PLCC↑ | MAE↓ | RMSE↓ |
|---|---|---|---|---|
| Matched anchor-only | $0.4660 \pm 0.0013$ | $0.4503 \pm 0.0014$ | $0.7501 \pm 0.0009$ | $\mathbf{0.9421} \pm 0.0006$ |
| Within-image shuffled | $0.5422 \pm 0.0014$ | $0.5362 \pm 0.0016$ | $0.7764 \pm 0.0067$ | $1.0348 \pm 0.0090$ |
| Full cluster-soft | $\mathbf{0.5601} \pm 0.0016$ | $\mathbf{0.5550} \pm 0.0014$ | $\mathbf{0.7398} \pm 0.0047$ | $0.9743 \pm 0.0067$ |

Table 6: Progressive ablation in the Qwen instantiation. Dist.: global distributional head; Elem.: element head; Cpl.: cross-level coupling; Guard.: evidence-bound regularization.

| Dist. | Elem. | Cpl. | Guard. | SRCC↑ | MAE↓ |
|---|---|---|---|---|---|
| ✓ | | | | 0.7214 | 0.6842 |
| ✓ | ✓ | | | 0.7415 | 0.6691 |
| ✓ | ✓ | ✓ | | 0.7526 | 0.6512 |
| ✓ | ✓ | ✓ | ✓ | **0.7583** | **0.5864** |

captions, example count, and per-image auxiliary-target multiset as full cluster-soft while breaking caption-specific target assignments.

With matched training settings, cluster-soft raises mean SRCC from 0.4660 to 0.5601 and PLCC from 0.4503 to 0.5550. MAE changes from 0.7501 to 0.7398, but RMSE increases from 0.9421 to 0.9743, indicating a remaining large-error tail. Restoring the correct caption-specific targets after within-image shuffling improves all four metrics. Extra auxiliary captions alone do not account for this gain.

### 5.4 Matched paired-bootstrap analysis

Prompt-group bootstrap on seed 42 supports the same interpretation. Full cluster-soft improves SRCC over matched anchor-only by 0.0898 (95% CI $[0.0416, 0.1375]$) and PLCC by 0.1013 ($[0.0546, 0.1469]$), both $p < 0.001$. Against within-image shuffling, full cluster-soft significantly lowers MAE by 0.0355 and RMSE by 0.0582 (both $p < 0.001$); the seed-42 correlation differences are positive but not significant at 0.05. Complete intervals are in Appendix D.

### 5.5 Component ablation

The distributional head reaches 0.7214 SRCC and 0.6842 MAE. Adding the element branch raises SRCC to 0.7415. Cross-level coupling reaches 0.7526 SRCC and 0.6512 MAE, and the final guardrail gives the largest drop in MAE, to 0.5864. The element branch and cross-level coupling improve SRCC, while the final guardrail produces the largest reduction in MAE.

### 5.6 Cluster stability and calibration behavior

A fixed 500-image diagnostic subset is sampled at the image level and unfolded to 13,027 caption-level records. Image-level sampling preserves each local caption neighborhood. Table 7 reports paraphrase stability and within-cluster ordering.

Positive variants fall within 0.2 score units of the anchor in 89.42% of evaluated pairs. The anchor outranks available counterfactuals in 98.89% of cases, and weak positives outrank counterfactuals in 93.54%. Paraphrases receive similar scores, whereas counterfactuals are ranked below their corresponding anchors in most cases. On this subset, 14.13% of unfolded records exceed absolute error 1.0, 3.42% exceed 1.5, and 0.34% exceed 2.0.

### 5.7 Domain-wise behavior and residual errors

Table 8 reports five semantic regimes. General images have the highest SRCC. Count/Spatial and Text/Symbol are harder to rank.

Table 7: Cluster-level stability and ordering on the fixed 500-image diagnostic subset.

| Metric | Ratio / accuracy | Pairs |
|---|---|---|
| $\lvert T_0 - T_p \rvert < 0.2$ | 0.8942 | 2,526 |
| $T_0 > T_n$ | 0.9889 | 2,526 |
| $T_p > T_n$ | 0.9032 | 7,651 |
| $T_0 > T_w$ | 0.9488 | 8,182 |
| $T_w > T_n$ | 0.9354 | 42,869 |

Table 8: Domain-wise performance on the fixed diagnostic subset after cluster unfolding.

| Domain | $n$ | SRCC | PLCC | MAE | RMSE |
|---|---|---|---|---|---|
| Abstract / Symbolic | 4,570 | 0.7479 | 0.7254 | 0.5161 | 0.6402 |
| Count / Spatial | 4,998 | 0.6948 | 0.6632 | 0.5564 | 0.6961 |
| Diagram / Texture | 1,557 | 0.7587 | 0.7144 | 0.5814 | 0.7123 |
| General | 4,476 | 0.7810 | 0.7404 | 0.5793 | 0.7196 |
| Text / Symbol | 1,930 | 0.6951 | 0.7088 | 0.4691 | 0.5787 |

Text/Symbol has relatively low SRCC but the lowest MAE, indicating reasonable numerical placement despite difficult fine-grained ordering. Count/Spatial remains difficult in both correlation and absolute error.

Manual inspection of high-error samples identifies four recurring regimes. Table 9 summarizes these cases; it is a descriptive error taxonomy, not a trained classifier.

Table 9: Recurring error regimes from manual inspection of high-error cases.

| Category | Typical cases | Observed tendency |
|---|---|---|
| Text / symbol | Rendered text, logos, symbols | Correct content can be under-scored. |
| Diagram / texture | Blueprints, guides, texture or design references | Non-natural layouts are often scored conservatively. |
| Count / spatial | Counts, relative position, multi-object layouts | Objects are found, but relation or quantity constraints remain difficult. |
| Abstract / style-heavy | Slogans, surreal scenes, strongly stylized prompts | Global plausibility can mask weak decisive semantics. |

Text/symbol and diagram/texture cases often show underestimation even when local evidence is present. Style-heavy cases are more often over-scored despite weak decisive semantics. Figure 3 compares absolute-error distributions for the dual-level and full Qwen models. The full model shifts more predictions into the low-error range and reduces the largest-error tail.

### 5.8 Ranking versus score calibration

The final Qwen guardrail changes SRCC by 0.0057 but reduces MAE by 0.0648. In frozen CLIP, cluster-soft gives the larger correlation gain; restoring correct caption–target assignments improves both correlation and error, and gap-aware ranking mainly reduces MAE/RMSE. The matched anchor comparison also leaves a higher RMSE, so clustered labels alone do not remove the large-error tail. We keep the ordering losses and score constraints as separate training terms because they affect different metrics.

## 6 Limitations

Only a category-stratified subset of the reference scores is directly human-scored. GLM-4.5V assigns the remaining scores, followed by human review for clear inconsistencies, so judge-specific bias may remain. GLM-4.5V and the Qwen2.5-VL-7B-Instruct backbone are separate models. Counterfactual coverage is incomplete, and element votes capture local support but not full compositional correctness. Cluster-soft improves correlation and MAE over the matched anchor baseline but gives a higher

RMSE, leaving some large residuals. Count/spatial relations and non-natural symbolic layouts are still difficult. The Qwen and frozen-backbone studies also use different split protocols. The current evaluation is limited to continuous alignment scoring.

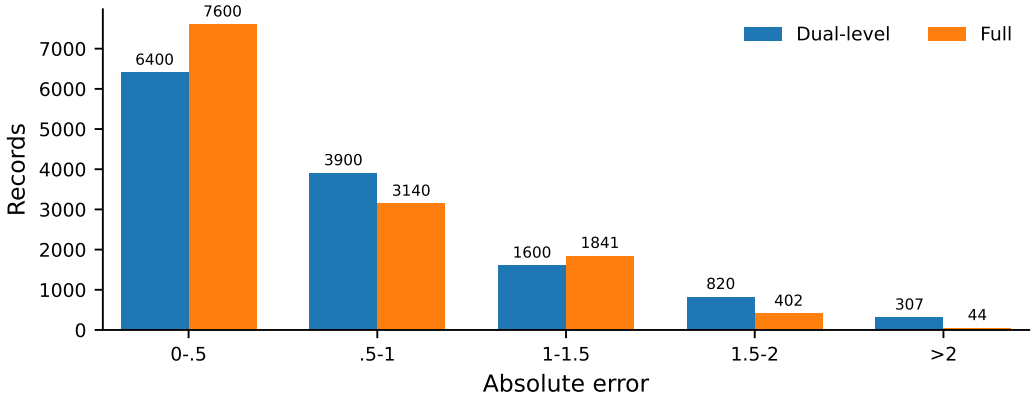


Figure 3: **Absolute-error distribution for the Qwen dual-level and full models.** The full model places more predictions below 0.5 absolute error and reduces the largest-error tail.

## 7 Conclusion

We introduced DECAR for continuous image–text alignment scoring with graded captions from the same image and element-level support during training. On Qwen2.5-VL, the element coupling and bound terms mainly reduce absolute error. In the three-seed frozen-CLIP study, clustered labels improve correlation, and the target-shuffle control shows an additional benefit from correct caption–score assignments. Gap-aware ranking lowers MAE/RMSE with little change in correlation.

Across these experiments, the results indicate that local caption ordering and numerical score accuracy benefit from different training signals. Clustered labels mainly improve ranking, whereas the guardrail and gap-aware loss have a larger effect on numerical error. This separation provides a practical way to train continuous image–text scorers that preserve fine-grained semantic ordering while producing more reliable score values.

## Reproducibility Statement

We report split rules, data counts, caption construction, score provenance, human review, model heads, sampling rules, loss margins, optimization settings, baseline checkpoints, and seeds. Matched controls use seeds 42/43/44; Appendix D reports seed-42 prompt-group bootstrap intervals from 5,000 resamples.

## AI Use Statement

Generative AI tools were used for auxiliary-caption generation, LLM-assisted counterfactual construction, GLM-4.5V score propagation described in Section 4 and language editing. GLM-4.5V is separate from the Qwen2.5-VL-7B-Instruct backbone used by DECAR. All reported experimental measurements come from the recorded training and evaluation runs.

## References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng,

Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12655–12663, 2020.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1218–1226, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330. PMLR, 2017.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. EvalMuse-40K: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. *arXiv preprint arXiv:2412.18150*, 2024.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

Tiep Le, Vasudev Lal, and Phillip Howard. COCO-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pp. 201–216, 2018.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705, 2021.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 12888–12900. PMLR, 2022.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742. PMLR, 2023.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4654–4662, 2019.

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, 2014.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10921–10930, 2020.

Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14354–14363, 2024.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3441–3450, 2015.

Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15661–15670, 2022.