CUEMEM: DISCOVERING SHARED VISUAL CUES IN LONG IMAGE COLLECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large visual collections contain recurring evidence that links images beyond their dominant content, enabling collection-level organization and exploration. In this paper, we define *query-free shared visual cue discovery*: given a long image collection, a model must identify image subsets grounded in recurring, interpretable visual cues and describe what each subset shares, without cue-specific queries or a predefined cue inventory. Conventional VLMs must recover sparse recurring evidence amid increasingly many distractors, yet standard token-level supervision does not directly separate supporting images from irrelevant ones. To study this problem, we introduce CUESEL, a diagnostic benchmark with 4k packed collections and 73k human-filtered cue groups spanning object, appearance, instance, and scene cues. To overcome the lack of explicit support modeling, we propose CUEMEM, which transforms internal VLM query/key interactions into cue-induced image affinities and uses inferred supports to ground cue generation. It achieves 0.747 Pairwise Grouping F1 (PGF1), compared with 0.405 for the strongest supervised flat VLM. Its advantage over flat VLMs grows as cue support becomes sparse relative to the collection, indicating that explicit relation learning and support-conditioned generation help VLMs recover recurring evidence under increasing visual distraction. Code and model data will be released.

1 INTRODUCTION

Large visual collections contain recurring evidence that links images beyond their dominant content, enabling collection-level organization and exploration. Yet such evidence is often not the dominant semantics of any single image: a few photographs may share the same object instance across viewpoints, a distinctive material pattern, or a subtle scene state while differing substantially in their remaining content. As vision-language models (VLMs) extend to longer image sequences (Qwen Team, 2024; Ye et al., 2024; Zhang et al., 2024; Chen et al., 2025), this motivates a question beyond recognizing or summarizing images: *can a model discover what visually connects subsets of a collection without being told what to search for?*

In this paper, we introduce *query-free shared visual cue discovery* to study this capability. Given a shuffled image collection and a general discovery instruction, the task is to infer an unordered set of support–description pairs $\{(\mathcal{S}_m, c_m)\}_{m=1}^{M}$, where $\mathcal{S}_m$ identifies the images supporting a recurring visual cue and c_m describes the evidence they share. Neither the cue vocabulary, its support, nor the number of cues is given. The task is therefore not exhausted by image grouping: grouping asks which images belong together under a similarity criterion, whereas cue discovery must also identify the open-vocabulary visual evidence that explains each support. We evaluate image-level supports and textual explanations rather than spatial cue localization.

Conventional VLMs face a fundamental evidence-dilution challenge in this setting. A recurring cue may be supported by only k images in a collection of size N, while the remaining images contribute salient but irrelevant objects, scenes, and events. Under a stylized image-level softmax, maintaining fixed interaction mass on the supporting images requires their score margin over distractors to grow with $\log((N-k)/(k-1))$ as N increases. Longer collections therefore do not merely add context; they intensify competition between weak recurring cues and dominant image-specific content. Token-level supervision provides no direct signal for strengthening within-cue image relations or suppressing distractor relations, making sparse supports difficult to recover.

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.

To evaluate this problem, we introduce CUESEL, a diagnostic benchmark comprising 4,008 long image collections and 73,852 human-filtered cue groups. Each cue is paired with the images in which it recurs, enabling joint evaluation of support discovery and cue description. By embedding these groups within long, shuffled collections, CUESEL tests whether models can recover sparse shared evidence without knowing in advance what to search for. More broadly, this capability would extend VLMs from answering predefined questions to autonomously structuring long visual contexts, supporting open-ended organization, retrieval, and analysis of large image collections.

Motivated by the evidence-dilution challenge, we propose CUEMEM, which formulates shared cue discovery as joint relational and semantic inference over a visual collection. Instead of leaving cross-image structure implicit in token prediction, CUEMEM transforms internal query/key (Q/K) interactions into an explicit image-level affinity geometry and aligns it with cue-induced relational structure. This learned geometry induces the image supports that condition semantic generation, grounding each description in the same relational evidence that determines its support. By coupling relational organization with semantic interpretation, CUEMEM preserves weak recurring cues against dominant image-specific content.

We compare CUEMEM with supervised flat VLMs and structured embedding controls. CUEMEM raises Pairwise Grouping F1 (PGF1) from 0.405 for the strongest flat VLM to 0.747, and also outperforms supervised cluster-based captioning (0.717), despite the latter using an explicit support-clustering step before captioning. Its advantage over flat generation increases as cue support becomes sparse relative to the collection, consistent with the evidence-competition analysis.

Our contributions are threefold:

- **Task.** Query-free shared visual cue discovery is formulated as jointly predicting image supports and open-vocabulary explanations, without cue-specific queries or a predefined cue inventory. This formulation extends VLMs toward autonomously organizing and analyzing long visual contexts.
- **Benchmark.** CUESEL provides 4,008 long image collections and 73,852 human-filtered cue groups for evaluating support discovery and cue description under sparse recurring evidence.
- **Method.** CUEMEM transforms internal VLM interactions into cue-induced image affinities and uses the inferred supports to ground cue generation, achieving gains over flat VLMs that increase as cue support becomes sparse.

2 RELATED WORK

Multi-image Understanding and Retrieval. Image–text alignment and instruction-following systems established the foundations of modern VLMs (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). Mantis and mPLUG-Owl3 extend learning and inference to multi-image observations (Jiang et al., 2024; Ye et al., 2024), while LongVA and LongVILA study longer visual contexts (Zhang et al., 2024; Chen et al., 2025). Benchmarks and surveys examine the resulting reasoning capabilities (Cheng et al., 2025; Yuan et al., 2026). Visual Haystacks and MMNeedle evaluate evidence retrieval among distractors given a question or target (Wu et al., 2025; Wang et al., 2025). Our setting instead leaves both the relevant evidence and its support unspecified. Long-context capacity makes this open-ended discovery possible, but does not by itself provide an objective for inducing the underlying image relations.

Shared Visual Content and Correspondence. Co-segmentation, co-saliency, and common-object discovery seek visual content shared across image sets (Rother et al., 2006; Rubinstein et al., 2013; Fu et al., 2013). This is the closest conceptual connection to our task: relevance depends on evidence shared across images rather than on single-image salience alone. Recent work also finds that explicit correspondence remains difficult for multimodal LLMs (Zhou et al., 2025). CUESEL studies multiple cue-induced relations in a mixed collection, ranging from appearance and scene context to object and instance correspondence. It requires image-level support recovery together with an open-vocabulary explanation, rather than a region mask or pairwise match alone. Conversely, its annotations do not provide the spatial localization evaluated by segmentation-based tasks.

Visual Representations and Grouping. DeepCluster, SCAN, and SwAV learn features through clustering or consistency objectives (Caron et al., 2018; Van Gansbeke et al., 2020; Caron et al., 2020). DINOv2/DINOv3 and SigLIP provide strong generic representations for visual similarity (Oquab et al., 2023; Siméoni et al., 2025; Zhai et al., 2023). We refer to clustering these embeddings and describing each resulting group as *Cluster-based Captioning*, a structured embedding control that is especially relevant to appearance and scene cues. CUEMEM instead adapts internal VLM interactions to the benchmark relation target and conditions cue descriptions on the inferred supports. Both approaches use affinity propagation (Frey & Dueck, 2007) to decode supports, providing a common structural decoder for the representation comparison.

Photo Selection and Visual Storytelling. Photo selection and album organization emphasize quality, coverage, diversity, or representative images (Obrador & Moroney, 2009; Potapova et al., 2009; Obrador et al., 2010). Visual storytelling produces coherent narratives across ordered image sequences (Huang et al., 2016; Yu et al., 2017; Wang et al., 2019). Although CUESEL draws images from VIST, it evaluates recurring visible evidence within shuffled pools rather than narrative continuity. A shared source story does not itself specify the cue: the output must identify its image support and explain the visual relation. This motivates assessing structural recovery separately from joint support-and-description agreement, rather than relying on the fluency or semantic similarity of a collection summary.

3 CUESEL: A SHARED VISUAL CUE DISCOVERY BENCHMARK

3.1 TASK DEFINITION

A *shared visual cue* is an interpretable unit of visual evidence that recurs across images, such as an appearance pattern, object, instance, material texture, visual state, or scene configuration. Its *support* is the set of images in which that evidence is visible. Given a shuffled image sequence $\mathcal{I} = (I_1, \ldots, I_N)$ and a general discovery instruction, the model predicts an unordered set $\mathcal{G} = \{(\mathcal{S}_m, c_m)\}_{m=1}^{M}$, where $\mathcal{S}_m \subseteq \{1, \ldots, N\}$ is a predicted support and c_m is its open-vocabulary description. Neither the cue vocabulary, support assignments, nor M is supplied at inference; training uses annotated examples. The task combines structural relation induction—recovering which images support a cue—with semantic abstraction of the evidence inducing that relation.

Compared with conventional image clustering, the distinction lies not only in granularity but also in what defines group membership. Clustering groups images according to similarity in a chosen representation, whereas shared visual cue discovery must infer the recurring, interpretable evidence that induces each relation and express it in open vocabulary. Such evidence can be finer-grained than dominant image semantics—for example, an instance detail, material pattern, visual state, or scene configuration—so globally dissimilar images may share a cue, while globally similar images need not share the annotated one. Thus, even under our controlled primary-partition protocol, the task requires both cue-conditioned support recovery and an explanation of why each support belongs together.

Unlike query-conditioned retrieval, shared visual cue discovery does not assume that the target relation is known in advance. Its support–description output provides an interpretable index over the collection: each discovered relation is expressed in open vocabulary and linked to the images that support it. This representation can serve as a primitive for browsing, organizing, and subsequently querying large visual collections, while making the evidence behind each discovered cue directly inspectable.

For reproducible evaluation, reference supports form a primary partition: $\bigcup_m \mathcal{S}_m = \{1, \ldots, N\}$ and $\mathcal{S}_m \cap \mathcal{S}_k = \varnothing$ for $m \neq k$. Group labels and their serialization order are arbitrary. This convention supplies a measurable structural target for PGF1 under the controlled packing protocol. Outputs specify image-level support rather than cue-region masks.

3.2 CUE ANNOTATION AND COLLECTION CONSTRUCTION

CUESEL uses Story-in-Sequence from the Visual Storytelling (VIST) dataset (Huang et al., 2016) as a source of candidate image sets. Annotation identifies a shared cue and its support subset within each short set; retained groups are subsequently packed into longer collections. Thus, the construction

unit is a cue-supported group, not an independently captioned image, and the resulting long samples are controlled compositions for evaluation. The supplement includes a qualitative reference-group visualization that illustrates how the annotation target is grounded in recurring visible evidence rather than story-level relatedness alone.

Each source set is independently reviewed by at least three trained annotators, who identify whether the set contains a global or partial visual relation and, when applicable, describe the visible shared content such as objects, textures, lighting, layouts, or instance details. We remove inconsistent judgments or supports, empty selections, underspecified cues, and malformed records, followed by a second quality-control pass that prioritizes precision over coverage. The raw candidate pool is approximately 1.3 times the retained pool, and a separate manual audit finds visible shared evidence in 191 of 200 sampled reference groups. Further annotation and audit details are provided in the supplement.

Before packing, retained groups are connected by a shared source story/candidate key or exact image URL, and each connected component is assigned wholly to one split. Packing preserves human-validated support–cue pairs while independently increasing the surrounding candidate pool. It therefore turns short, interpretable annotations into a controlled long-collection diagnostic: the visual evidence defining each cue remains localized to a small support, but must be recovered among increasingly many unrelated images and competing cue relations. Shuffling and fresh identifier assignment further prevent models from exploiting source order or persistent identifiers.

3.3 DATASET STATISTICS

Table 1: CUESEL split statistics. Smp. and Grp. denote the total numbers of samples and cue groups; Med. and Max denote median and maximum. Statistics under Images/sample and Groups/sample are computed per benchmark sample.

Split	#Smp.	#Grp.	Images / sample			Groups / sample		
			Mean	Med.	Max	Mean	Med.	Max
Train	3,498	34,974	31.36	31	45	10.00	10	10
Test	510	38,878	238.34	230	1,644	76.23	74	520
Overall	4,008	73,852	57.69	32	1,644	18.43	10	520

After filtering and packing, CUESEL contains 4,008 benchmark samples and 73,852 cue-centered reference groups. Table 1 summarizes the resulting splits.

CUESEL stresses three coupled dimensions of collection-level discovery. First, it introduces a substantial length shift: training collections contain 31.36 images on average and at most 45, whereas test collections average 238.34 images and reach 1,644; 363 of 510 test collections exceed 128 images and 223 exceed 256. Second, individual supports remain small, averaging 3.13 images, while singleton groups account for only 0.068% of all groups. Recurring evidence therefore becomes increasingly sparse relative to the candidate pool. Third, relational multiplicity grows with collection length: a test collection contains 76.23 cue groups on average and as many as 520. The resulting challenge is not merely processing more images, but recovering many locally supported visual relations without obscuring or conflating them amid competing content.

3.4 EVALUATION

We evaluate cue discovery from two primary perspectives: whether a model recovers the image supports induced by recurring cues, and whether its descriptions semantically agree with the evidence in those supports. Because flat VLMs must additionally serialize a complete partition, we separately diagnose output validity:

- **Pairwise Grouping F1 (PGF1)** converts predicted and reference supports into same-cue image pairs and computes $\mathrm{PGF1} = 2TP/(2TP + FP + FN)$. Here, true positives are pairs grouped together in both partitions, while false positives and false negatives capture incorrect merges and missed within-cue relations, respectively. This pairwise formulation avoids dependence on arbitrary group labels.

- **Cue Semantic Alignment (CSA)** matches predicted and reference supports by intersection-over-union, computes embedding similarity between their cue descriptions, and weights this similarity by the matched image overlap. It therefore measures end-to-end support–description agreement rather than standalone caption quality.
- **Partition Validity Score (PVS)** multiplies identifier coverage, exclusivity, and validity before repair. These factors check whether all input images are included, whether an identifier appears in at most one support, and whether generated identifiers belong to the input collection. PVS isolates serialization failures from cue-discovery quality.

In addition, the supplement adds a manual cue-compatibility audit and a fixed-support caption preference comparison.

Before scoring PGF1 and CSA, we remove invalid identifiers and duplicates within each predicted support. If an image appears in multiple supports, only its first assignment is retained. For PGF1, omitted images are added as singleton groups so that each prediction forms a complete partition. Metrics are computed per collection and macro-averaged over the 510 test collections. We use PGF1 as the primary measure of support recovery; detailed repair rules and CSA matching cases are provided in the supplement. Human audits separately assess whether the predicted descriptions correspond to visually valid cues.

4 Learning Relations for Shared Visual Cue Discovery

4.1 Overview: Relation Induction and Semantic Abstraction

Shared visual cue discovery resembles a natural human inspection process: one first scans the collection to determine which images are connected by recurring visual evidence, and only then abstracts a verbal description of what the shared evidence is. This suggests that the task involves two complementary forms of reasoning: recovering the support structure induced by shared cues, and explaining the semantics of each recovered support. Treating both requirements as a single autoregressive serialization problem entangles grouping with description, making structural errors difficult to separate from semantic errors.

CueMem follows this observation by factorizing the set-valued prediction into a structural stage that induces same-cue image relations and a semantic stage that explains the evidence underlying each inferred support (Figure 1). Both objectives adapt the same VLM through LoRA. Let $S_\theta(\mathcal{I})$ denote the image-level affinity matrix induced from a collection $\mathcal{I}$, and let $\widehat{\mathcal{S}} = \{\widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_{\hat{M}}\}$ denote its decoded support structure. At inference,

$$\widehat{\mathcal{S}} = \mathrm{Bind}(S_\theta(\mathcal{I})), \qquad \hat{c}_m = \mathrm{Decode}_\theta(\mathcal{I}, \widehat{\mathcal{S}}_m, \hat{c}_{<m}). \tag{1}$$

Here, Bind maps pairwise affinities to an unordered family of predicted supports. The identifiers of each predicted support are then deterministically serialized as a fixed prefix, and Decode_θ generates its cue description. Thus, support membership is inferred from the learned affinities, whereas only its textual identifier representation is programmatic.

Why sparse supports motivate relation learning. The structural stage is especially important when a cue is supported by only a few images within a larger collection. In this setting, relevant images must compete with many visually plausible distractors, so the model needs a sufficiently separated image-level relation signal before cue description can be reliable. A simple equal-score abstraction makes this dependence explicit. Consider a self-excluded image-level softmax with temperature $\tau_{\mathrm{attn}} > 0$, scores μ_+ for the other $n_m - 1$ supporting images and μ_- for $N - n_m$ distractors, where $1 < n_m < N$. Under this abstraction, the total softmax mass assigned to supporting images is

$$A_m = \frac{(n_m - 1)\exp(\mu_+/\tau_{\mathrm{attn}})}{(n_m - 1)\exp(\mu_+/\tau_{\mathrm{attn}}) + (N - n_m)\exp(\mu_-/\tau_{\mathrm{attn}})}. \tag{2}$$

Solving $A_m \geq \rho$ gives

$$\mu_+ - \mu_- \geq \tau_{\mathrm{attn}} \log \frac{\rho}{1-\rho} + \tau_{\mathrm{attn}} \log \frac{N - n_m}{n_m - 1}. \tag{3}$$

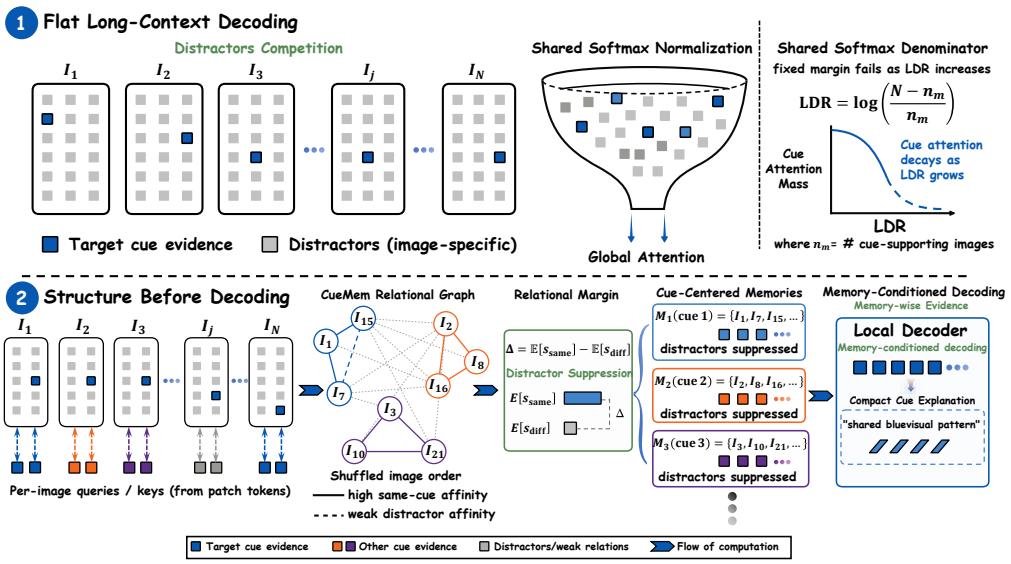


Figure 1: CUEMEM separates shared cue discovery into relation-based support recovery and support-conditioned cue description. Unlike flat long-context decoding, it first induces cue-centered relations from supervised Q/K interactions, then uses the inferred support memories to ground each description.

Under this equal-score approximation, the required gap grows with the distractor-to-support ratio. Since token-level generation supervision does not directly constrain this image-level separation, this trend motivates an auxiliary objective on cue-induced relations and an evaluation stratified by support sparsity. We use the calculation as diagnostic intuition for distractor competition; Appendix A.1 gives the derivation and assumptions.

4.2 RELATIONAL SUPERVISION FOR CUE-INDUCED IMAGE STRUCTURE

The relation stage aims to make the implicit cross-image comparisons inside the VLM explicit. If two images support the same cue, some of their visual patches should become mutually compatible under the model's internal representation, even when the cue is sparse or surrounded by distractors. Rather than learning a separate relation encoder, CUEMEM extracts this signal from the VLM's full-attention interactions, thereby coupling structural learning to the visual space used for generation.

Concretely, at selected full-attention layers ℓ and heads h, we use the normalized query and key vectors before rotary position embedding (RoPE). For patches u and v from images i and j, with head dimension d_h, their interaction score is

$$r^{\ell,h}_{i,u;j,v} = \frac{(q^{\ell,h}_{i,u})^\top k^{\ell,h}_{j,v}}{\sqrt{d_h}}. \tag{4}$$

These patch-level interactions are then aggregated into an image-level affinity. We normalize over destination visual patches, sum the resulting mass within each destination image, average over query patches and heads, and symmetrize the two directions. Uniform averaging over the selected layers yields $p_{ij} \in [0, 1]$, an estimate of whether images i and j support the same cue. This score should be interpreted as structural evidence derived from internal VLM interactions, not as a localized cue map or the model's native causal attention distribution.

Reference supports provide direct supervision for this relational signal. We set $y_{ij} = 1$ when images i and j belong to the same annotated cue group and $y_{ij} = 0$ otherwise, and optimize the mean binary cross-entropy

$$\mathcal{L}_{\text{rel}} = -\frac{2}{N(N-1)} \sum_{i<j} \left[y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij}) \right]. \tag{5}$$

This objective encourages the VLM to separate cue-supporting pairs from distractor pairs at the image-relation level, complementing token-level generation supervision. The labels correspond to

the primary reference partition and do not specify cue identity or spatial extent; negative pairs may still share unannotated visual properties. At inference, off-diagonal affinities are rescaled per sample to form S_θ, used only as the structural input to support decoding rather than calibrated probabilities.

4.3 Grounding Cue Descriptions in Inferred Support Structure

The learned affinity matrix provides a continuous view of cue-induced image structure, but the task requires discrete supports paired with natural-language explanations. We therefore first convert pairwise affinities into cue-level support sets, and then use each support as the visual grounding context for cue description. This design separates the decision of which images belong together from the description of why they belong together.

We decode supports with affinity propagation (AP) (Frey & Dueck, 2007), which infers the number of groups from the affinity structure rather than requiring it to be specified in advance. AP uses the median off-diagonal affinity as its preference. The resulting supports serve as temporary cue memories that anchor subsequent description generation; alternative structural decoders are evaluated in the ablation study.

During training, a reference assignment vector $\mathbf{Z}$ links each support–description block to its visual support. This assignment restricts direct full-attention access from the block to the corresponding support's visual patches, while preserving image–image interactions and causal text attention. The generation loss covers all assistant-target tokens $\mathcal{A}$, including identifiers and delimiters:

$$\mathcal{L}_{\text{gen}} = -\frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \log p_\theta(y_t \mid y_{<t}, \mathcal{I}, \mathbf{Z}). \tag{6}$$

The shared LoRA adapters are optimized with a weighted sum of the generation and relation losses, $\mathcal{L}_{\text{gen}} + \lambda_{\text{rel}} \mathcal{L}_{\text{rel}}$, with the weight specified in the supplement.

At inference, a global forward pass first estimates the affinity matrix and decodes the support structure. Each inferred support is then serialized as a fixed prefix and used to condition cue description. Because generation operates after support decoding, it explains the inferred support without changing its membership. Additional training and inference details are provided in the supplement.

5 Experiments

We evaluate CueMem on support recovery, generic-similarity comparisons, sparse-support robustness, and design ablations. PGF1 measures reference same-cue recovery, while CSA measures joint support–description agreement.

5.1 Experimental Setup

Models and implementation. We compare supervised flat baselines from GLM-Edge-V-2B (GLM Team, 2024), MiniCPM-V (Yao et al., 2024), Qwen2-VL-2B (Qwen Team, 2024), Qwen2.5-VL-3B (Qwen Team, 2025a), Qwen3-VL-4B (Qwen Team, 2025c), and Qwen3.5-4B (Qwen Team, 2026). Flat models are fine-tuned on CueSel to serialize assignments and cues. CueMem uses Qwen3.5-4B with LoRA adaptation (Hu et al., 2022), rank 16, alpha 32, dropout 0.05, and affinity propagation as its default structural decoder. We evaluate a Qwen3-VL-4B instantiation. Detailed preprocessing, training, resolution, and generation-budget settings are provided in the supplement.

Comparison scope. All systems predict supports and cue descriptions. Flat VLMs and CueMem use the same training annotations, from which CueMem derives pair labels and support-conditioned visual masks. Cluster-based Captioning measures generic similarity under explicit AP support clustering; its supervised variant adds task-specific SigLIP embedding adaptation before the same captioning procedure. Relation-loss, decoder, and output-interface controls examine the main design choices. Full protocols and additional comparisons are provided in the supplement.

5.2 Structural Recovery and Description Agreement

Table 2 shows that explicit structural prediction substantially improves cue-support recovery. Relative to flat Qwen3.5-4B, CueMem raises PGF1 from 0.405 to 0.747 and CSA from 0.371 to 0.748.

Table 2: Supervised flat VLMs and CUEMEM on the CUESEL test split. Superscripts denote model scale in billions of parameters; GLM denotes GLM-Edge-V. Higher is better for all metrics.

Metric	GLM^{2B}	Qwen2-VL^{2B}	Qwen2.5-VL^{3B}	MiniCPM-V^{8B}	Qwen3-VL^{4B}	Qwen3.5^{4B}	CUEMEM^{4B}
PVS ↑	0.508	0.256	0.429	0.151	0.334	0.434	**0.987**
PGF1 ↑	0.033	0.051	0.092	0.135	0.247	0.405	**0.747**
CSA ↑	0.216	0.140	0.118	0.204	0.246	0.371	**0.748**

The gains indicate better recovery of the annotated support–description pairs under the benchmark protocol, which is the intended controlled comparison target of CUESEL. PVS rises from 0.434 to 0.987, partly reflecting deterministic identifier serialization. A flat support-only control followed by programmatic identifier serialization and recaptioning reaches 0.553 PGF1 and 0.602 CSA. Separating support prediction from cue generation improves the flat baseline, while CUEMEM retains higher scores under this control. The supplement further verifies the effectiveness of the same formulation on Qwen3-VL-4B and reports additional description-quality checks.

5.3 CUE RECOVERY BEYOND GENERIC VISUAL SIMILARITY

Many shared cues correlate with generic visual similarity, raising the question of whether explicit cue-relation learning is necessary. We compare CUEMEM with structured pipelines that apply the same AP decoder to frozen DINOv2, DINOv3, and SigLIP representations. Their supports are described by the same flat Qwen3.5-4B LoRA, isolating the relation features used for support inference. We also train a two-layer SigLIP projection with pair labels from the CUESEL partition, then use the same decoder and captioner. The supplement describes the projection, objective, and decoding protocol.

Table 3: Structured support-inference pipelines on the test split. All use AP as the structural decoder. Cluster-based Captioning uses the CUESEL-trained flat Qwen3.5-4B LoRA for cue generation, whereas CUEMEM uses its jointly adapted VLM.

Pipeline	Relation features	PGF1 ↑	CSA ↑
Cluster-based Captioning	DINOv2-large	0.560	0.614
Cluster-based Captioning	DINOv3-ViT-L/16	0.650	0.669
Cluster-based Captioning	SigLIP-SO400M	0.685	0.686
Supervised Cluster-based Captioning	Adapted SigLIP projection	0.717	0.708
CUEMEM	Adapted Qwen3.5-4B Q/K	**0.747**	**0.748**

Table 3 shows that generic representations recover substantial support structure when paired with explicit AP support clustering: PGF1 ranges from 0.560 for DINOv2 to 0.685 for SigLIP, compared with 0.405 for flat Qwen3.5-4B. Supervised Cluster-based Captioning raises PGF1 to 0.717 and CSA to 0.708 through task-specific SigLIP adaptation. These controls are not flat VLM generators; they separate support inference from captioning. CUEMEM reaches 0.747 PGF1 and 0.748 CSA, gains of 0.030 and 0.040 over Supervised Cluster-based Captioning, indicating higher support-and-description performance among the evaluated structured systems.

5.4 SPARSE-SUPPORT RECOVERY IN LONG COLLECTIONS

We evaluate cue-support recovery on collections with more than 100 images (Table 4) and stratify performance by support sparsity.

Table 4: Results on test cases with more than 100 images. Gray superscripts indicate model scale.

Metric	Qwen2-VL^{2B}	Qwen2.5-VL^{3B}	Qwen3-VL^{4B}	Qwen3.5^{4B}	CUEMEM^{4B}
PVS ↑	0.166	0.350	0.178	0.309	**0.979**
PGF1 ↑	0.012	0.021	0.101	0.285	**0.709**
CSA ↑	0.082	0.035	0.108	0.260	**0.723**

On the >100-image subset, CUEMEM reaches 0.709 PGF1 compared with 0.285 for flat Qwen3.5-4B, a gain of 0.424. CSA increases from 0.260 to 0.723. The larger gap than on the full test split indicates that the structured model better retains the reference support relations in long inputs.

Support sparsity. We compute $\mathrm{LDR}_m = \log((N - |\mathcal{S}_m|)/|\mathcal{S}_m|)$ and average over reference cue groups per sample. Larger LDR indicates sparser support relative to the collection. This proxy includes the query image in its support count, unlike Equation 3. In Figure 2, flat Qwen3.5-4B is competitive in the Low bucket (0.845 versus 0.886 PGF1), but falls to 0.070 in the Extreme bucket while CUEMEM retains 0.673. The gap grows from 0.041 to 0.604 using unrounded means. Both methods decline, but CUEMEM preserves more support structure as evidence becomes sparse. Because LDR also reflects collection length and the number of competing cue groups, we interpret this trend as evidence of robustness under increasingly cluttered inputs rather than as an isolated effect of sparsity alone.

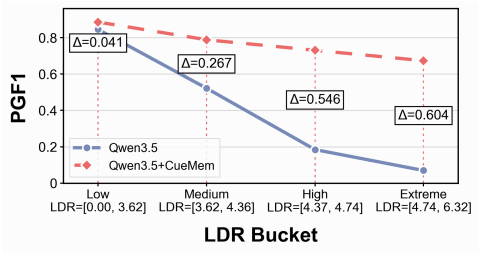


Figure 2: PGF1 across LDR buckets, where higher LDR indicates sparser cue support. CUEMEM degrades more slowly than the flat Qwen3.5 baseline as evidence becomes increasingly sparse.

5.5 Ablations and Design Analysis

Removing direct pair supervision while retaining support-conditioned generation and the same structural decoder lowers PGF1 from 0.747 to 0.706 and CSA from 0.748 to 0.727. Pair supervision adds 0.041 PGF1 within the structured pipeline, whose remaining components already provide a strong baseline.

Table 5: Relation-supervision ablation. Rel. sup. denotes relational supervision.

Setting	PGF1 ↑	CSA ↑
w/o Rel. sup.	0.706	0.727
w/ Rel. sup.	**0.747**	**0.748**

5.6 Cost of Discovering and Describing Cue Supports

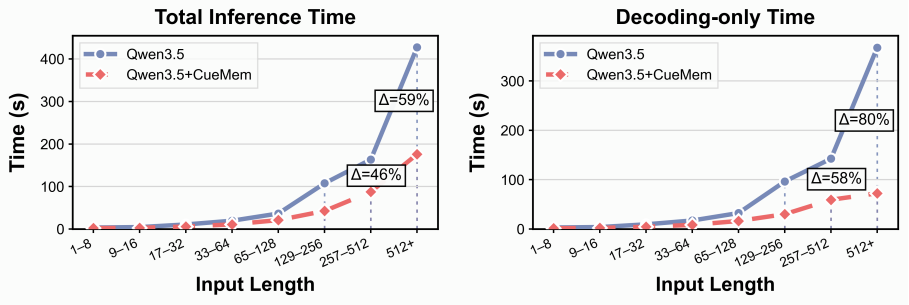


Figure 3: Sampled pipeline runtime: end-to-end (left) and decoding (right). Each bucket contains at most five length-spaced cases; the supplement details the timing protocol and measurements.

Figure 3 compares sampled end-to-end and decoding times under matched preprocessing, prompt format, and nominal output budgets. CUEMEM's total time includes relation estimation, structural decoding, and cue generation. In the sampled 129–256-image bucket, total time decreases from 107.6 to 42.5 seconds (60.5%); in the 513+ bucket, it decreases from 427.3 to 175.9 seconds (58.8%). Decoding time in the latter bucket falls from 367.6 to 72.4 seconds (80.3%), outweighing the additional structural computation on these cases.

6 Conclusion

We introduced query-free shared visual cue discovery, where a model must identify recurring visual evidence and describe the image subsets supported by that evidence without cue-specific queries or a predefined cue inventory. To study this problem, we built CUESEL, a benchmark of long image collections packed from human-filtered cue groups. CUEMEM addresses the evidence-dilution challenge by transforming internal VLM interactions into cue-induced image affinities and grounding cue descriptions in the inferred supports. It substantially improves reference-support recovery over flat VLMs and shows larger gains as cue support becomes sparse. These results suggest that explicit relation learning can help VLMs move beyond answering or summarizing images toward autonomously organizing long visual contexts.

AI USE STATEMENT

The authors employed generative AI tools exclusively for light language polishing of text originally written by the authors. This includes grammar checks, sentence rephrasing, improving clarity and consistency of academic prose. No generative AI was used to conceive research ideas, design experiments, derive technical claims, analyze experimental results, or draft any core technical content of this work. All modifications proposed by the AI tool were carefully examined and verified by the authors before adoption. The authors hold full responsibility for all content, claims, and conclusions presented in this manuscript.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision*, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. LongVILA: Scaling long-context visual language models for long videos. In *Proceedings of the International Conference on Learning Representations*, 2025.

Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, Yanlin Li, Lei Ren, Wei Chen, Zhiyuan Huang, Mingjie Zhan, Xiaojie Wang, and Fangxiang Feng. Evaluating MLLMs with multimodal multi-image reasoning benchmark, 2025. URL https://arxiv.org/abs/2506.04280.

Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing*, 2013.

GLM Team. GLM-Edge: Glm series edge models, 2024. URL https://github.com/zai-org/GLM-Edge. Vision-language checkpoint: glm-edge-v-2b.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 1233–1239, 2016.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Pere Obrador and Nathan Moroney. Automatic image selection by means of a hierarchical scalable collection representation. In *Proceedings of SPIE Visual Communications and Image Processing*, volume 7257, pp. 72570W, 2009. doi: 10.1117/12.806088.

Pere Obrador, Rodrigo de Oliveira, and Nuria Oliver. Supporting personal photo storytelling for social albums. In *Proceedings of the ACM International Conference on Multimedia*, 2010. doi: 10.1145/1873951.1874025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Ekaterina Potapova, Marta Egorova, and Ilia Safonov. Automatic photo selection for media and entertainment applications. In *Proceedings of the 19th International Conference on Computer Graphics and Vision*, 2009.

Qwen Team. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Qwen Team. Qwen2.5-VL technical report, 2025a. URL https://arxiv.org/abs/2502.13923.

Qwen Team. Qwen3 Embedding: Advancing text embedding and reranking through foundation models, 2025b. URL https://arxiv.org/abs/2506.05176.

Qwen Team. Qwen3-VL technical report, 2025c. URL https://arxiv.org/abs/2511.21631.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Public checkpoint: https://huggingface.co/Qwen/Qwen3.5-4B.

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 International Conference on Machine Learning*, 2021.

Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching incorporating a global constraint into MRFs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision*, 2020.

Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, and Feng Zhang. Hierarchical photo-scene encoder for album storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8909–8916, 2019. doi: 10.1609/aaai.v33i01.33018909.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, 2025. URL https://arxiv.org/abs/2406.11230.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.13766.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone, 2024. URL https://arxiv.org/abs/2408.01800.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. URL https://arxiv.org/abs/2408.04840.

Licheng Yu, Mohit Bansal, and Tamara Berg. Hierarchically-attentive RNN for album summarization and storytelling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 966–971, 2017. doi: 10.18653/v1/D17-1101.

Zhiqiang Yuan, Zexi Jia, Ying Deng, Jiapei Zhang, Kailin Lyu, Qiwei Yan, Yang Liu, Liwei Yang, Xiaoyue Duan, Peixiang Luo, Jinchao Zhang, and Jie Zhou. Visual macroscopic intelligence: A survey from multi-image understanding to agentic visual reasoning. *ChinaXiv*, 2026. doi: 10.12074/202606.00228.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision, 2024. URL https://arxiv.org/abs/2406.16852.

Yikang Zhou, Tao Zhang, Shilin Xu, Shihao Chen, Qianyu Zhou, Yunhai Tong, Shunping Ji, Jiangning Zhang, Lu Qi, and Xiangtai Li. Are they the same? exploring visual correspondence shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

SUPPLEMENTARY MATERIAL CONTENTS

A A Stylized Analysis of Sparse Cue Support **14**

A.1 Equal-Score Calculation 14

A.2 Conditional Implication Within the Abstraction 14

A.3 Using Predicted Support to Describe a Cue 14

B Dataset Supplementary Materials **14**

B.1 Benchmark Challenges 14

B.2 Additional Dataset Statistics 15

B.3 Shared Cues and Their Supporting Images 15

B.4 Full Cue Category Distribution 15

B.5 Annotation Quality and Split Hygiene 16

B.6 Reference-Partition Audit 17

C Method Implementation Supplementary Materials **17**

C.1 Fine-Tuned Flat Baseline Protocol 17

C.2 Backbone and Fine-Tuning 17

C.3 Cue-Group Affinities from Internal Q/K Interactions 18

C.4 Cue Description Conditioned on Support Groups 18

C.5 Inference Procedure 18

C.6 Input and Generation Settings 19

D Supplementary Experiments **19**

D.1 Cross-Backbone Validation Details 19

D.2 Output-Interface Control 19

D.3 External Stress Test on Visual Haystacks Multi-Needle 20

D.4 Ablation of Relational Supervision 20

D.5 Sensitivity of Cue-Support Inference to Clustering 20

D.6 Detailed Evaluation Metrics 21

D.7 Manual Consistency Audit for CSA 23

D.8 Cluster-based Captioning Baselines 23

D.9 LDR Bucket Details 24

D.10 Runtime Bucket Details 25

D.11 Qualitative Result Comparison 25

D.12 Failure Case Analysis 25

E Limitations **27**

F Ethics Statement **27**

A A Stylized Analysis of Sparse Cue Support

We analyze competition between supporting images and distractors using a self-excluded image-level softmax with equal scores within each set. This abstraction connects support mass to the score gap and distractor count, providing diagnostic intuition for the evidence-dilution effect studied in the main text.

A.1 Equal-Score Calculation

Consider a reference group of size $1 < n_m < N$ in a sequence of N images and a softmax temperature $\tau_{\text{attn}} > 0$. For a query image from this group, the competing candidates consist of $n_m - 1$ other supporting images and $N - n_m$ distractors. Let μ_+ be the common support-image score and μ_- the common distractor-image score. The total mass assigned to supporting images is

$$A_m = \frac{(n_m - 1)\exp(\mu_+/\tau_{\text{attn}})}{(n_m - 1)\exp(\mu_+/\tau_{\text{attn}}) + (N - n_m)\exp(\mu_-/\tau_{\text{attn}})}. \tag{7}$$

To maintain support mass at least $\rho \in (0, 1)$, we require $A_m \geq \rho$. Multiplying both sides by the denominator and collecting the support terms gives

$$(1 - \rho)(n_m - 1)\exp(\mu_+/\tau_{\text{attn}}) \geq \rho(N - n_m)\exp(\mu_-/\tau_{\text{attn}}). \tag{8}$$

Equivalently,

$$\exp\left(\frac{\mu_+ - \mu_-}{\tau_{\text{attn}}}\right) \geq \frac{\rho}{1 - \rho} \cdot \frac{N - n_m}{n_m - 1}. \tag{9}$$

Taking logs yields

$$\mu_+ - \mu_- \geq \tau_{\text{attn}} \log \frac{\rho}{1 - \rho} + \tau_{\text{attn}} \log \frac{N - n_m}{n_m - 1}. \tag{10}$$

Thus, for a fixed target mass ρ and temperature, the required score gap grows logarithmically with the distractor-to-support ratio. The empirical proxy $\text{LDR}_m = \log((N - n_m)/n_m)$ uses all n_m group members in the denominator rather than excluding the query image. At fixed group size, both the proxy and Equation 10 grow with the number of distractors, making LDR a practical statistic for stratifying support sparsity in the benchmark.

A.2 Conditional Implication Within the Abstraction

Let $\Delta = \mu_+ - \mu_-$ denote the score gap. At fixed ρ and τ_{attn}, increasing the gap from Δ_0 to Δ_1 multiplies the maximum admissible distractor-to-support ratio by $\exp((\Delta_1 - \Delta_0)/\tau_{\text{attn}})$. This implication motivates evaluating support recovery under increasing distractor pressure, while the experiments directly measure recovered supports and cue agreement.

A.3 Using Predicted Support to Describe a Cue

Conditioning cue generation on a predicted support restricts direct full-attention access to the selected images. Global representations, linear-attention state, and causal text context remain available. The mask therefore focuses the model's direct visual access on the inferred support while preserving the backbone's remaining context pathways.

B Dataset Supplementary Materials

This section reports cue composition, annotation quality control, and split construction for CueSel.

B.1 Benchmark Challenges

Cue discovery requires jointly identifying recurring visual evidence and its support from a general instruction. At inference, neither a cue inventory nor a reference group count is supplied. Training uses annotated examples.

The test collections contain 238.34 images on average and up to 1,644, while each reference group averages 3.13 images. Models must recover these small supports from a much larger candidate pool. They must also distinguish shared evidence from salient image-specific content and describe it at an appropriate level of detail. The cues span color, lighting, texture, objects, instances, and scenes, requiring comparisons at several levels of visual similarity.

B.2 Additional Dataset Statistics

Table 6 reports the bucketed image count distribution in the test split. The test set is deliberately length shifted: 363 out of 510 samples contain more than 128 images, 223 contain more than 256 images, and 4 contain more than 1,024 images.

Table 6: Image count distribution in the test split.

Images	**1–8**	**9–16**	**17–32**	**33–64**	**65–128**
Samples	3	8	25	34	77
Images	**129–256**	**257–512**	**513–1,024**	**>1,024**	**Total**
Samples	140	217	2	4	510

Table 7 reports group- and cue-level statistics. The average group size is 3.13 images and the median group size is 3, confirming that target evidence is usually sparse within a long candidate sequence. Singleton groups account for only 0.068% of all groups, suggesting that most retained annotations contain non-trivial cross-image correspondences.

Table 7: Group- and cue-level statistics. Singleton denotes single-image groups.

Split	**Grp. Size Mean/Med.**	**Sing. Ratio**	**Cue Tok. Mean/Med.**	**Uniq. Tok. Mean/Med.**	**Vocab. Size**
Train	3.14/3	0.086%	10.17/9	8.77/8	5,205
Test	3.13/3	0.051%	10.23/9	8.80/8	1,739
Overall	3.13/3	0.068%	10.20/9	8.79/8	5,450

B.3 Shared Cues and Their Supporting Images

Figure 4 illustrates one cue-centered reference group in CueSel. Multiple views contain recurring but partially occluded details that jointly motivate the reference description. The visualization explains the annotation semantics and the type of recurring evidence targeted by the benchmark.

B.4 Full Cue Category Distribution

We visualize frequent visual cue categories in Figure 5 and report an extended category table below. The English strings used in these textual analyses are translated from the cleaned Chinese annotations, and underspecified expressions such as “the same” are removed.

Similar color and similar lighting occur in 17.9% and 16.2% of cue annotations; shared-object and same-instance tags occur in 8.3% and 7.5%. The tags are non-exclusive. Appearance and scene cues make generic visual embeddings relevant baselines, while object and instance cues broaden the annotation inventory beyond global appearance similarity.

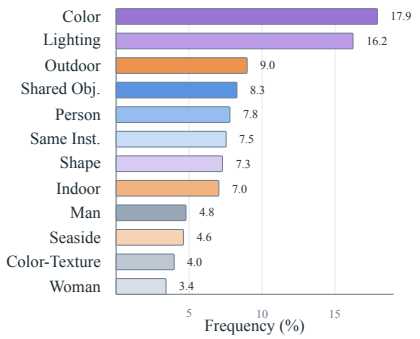


Figure 5: Top visual cue categories after cleaning and translation.

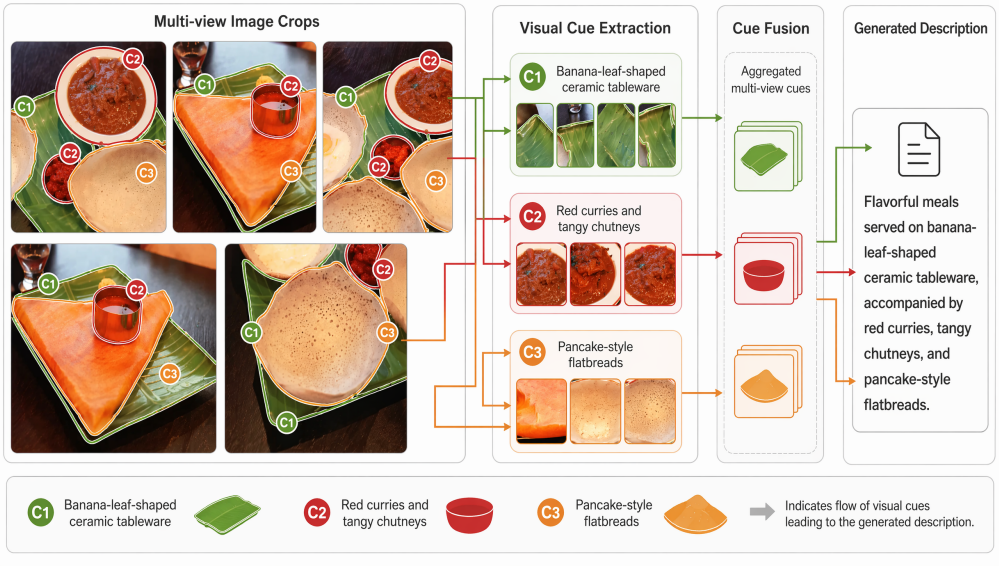


Figure 4: Qualitative visualization of a multi-view reference group. Repeated visible details include banana-leaf-shaped tableware, red curries and chutneys, and pancake-style flatbreads; the figure relates these observations to the group's cue description.

B.5 Annotation Quality and Split Hygiene

CueSel uses several safeguards to make construction auditable and reduce avoidable leakage. Before annotation, annotators received task-specific training and calibration examples emphasizing concrete visual evidence rather than generic story relatedness. Each five-image source candidate set was independently reviewed by at least three annotators as unrelated, globally related, or partially related. For a partial relation, the form required both a support subset and a cue tied to visible evidence. Cue descriptions were collected in Chinese and translated into English for model training and text-based evaluation. Support annotations are defined directly over image identifiers and are therefore unchanged by translation; all compared systems use the same English references.

Candidate sets with inconsistent relatedness judgments or selected subsets, empty selections, underspecified cue strings, or malformed records are removed rather than resolved by simple majority vote. A second quality-control pass checks whether each retained cue names visible evidence and whether its selected images support that cue. The raw candidate pool is approximately 1.3 times the retained pool, so roughly one quarter is removed. This redundancy and filtering improve annotation precision, and we treat the retained outputs as reference groups for controlled evaluation.

Table 8: Top-29 non-exclusive visual cue categories after cleaning, translation, and removal of underspecified labels. Counts are category assignments and frequencies use all cue annotations as denominator.

Rank	Category	Count	Freq.
1	Similar color	13,230	0.179
2	Similar lighting	11,991	0.162
3	Outdoor	6,630	0.090
4	Shared object	6,117	0.083
5	Person	5,760	0.078
6	Same instance	5,572	0.075
7	Similar shape	5,387	0.073
8	Indoor	5,201	0.070
9	Man (formal)	3,534	0.048
10	Seaside	3,403	0.046
11	Similar color texture	2,939	0.040
12	Woman (formal)	2,523	0.034
13	Woman	2,498	0.034
14	Street	2,189	0.030
15	Snow field	2,109	0.029
16	Grassland	2,053	0.028
17	Man	1,663	0.023
18	Similar color tone	1,598	0.022
19	Building	1,544	0.021
20	Wall	1,461	0.020
21	Sea	1,326	0.018
22	Snow	1,246	0.017
23	Ground	1,134	0.015
24	Table	1,104	0.015
25	Fireworks	1,080	0.015
26	Night	1,011	0.014
27	Clothing	1,003	0.014
28	Tree	978	0.013
29	Hat	922	0.013

For split construction, connected components are the atomic units before packing. Two retained groups are connected if they share a source story/candidate key or an exact image URL, and each

Table 9: Manual audit of reference-partition validity, alternative cue granularities, and fixed-support caption preference. CBC denotes Supervised Cluster-based Captioning.

Audit item	Samples	Outcome
Reference groups with visible shared evidence	200 groups	191 / 200 (95.5%)
Random cross-group pairs with alternative cue	200 pairs	14 / 200 (7.0%)
Major disagreements with alternative granularity	100	Flat: 24; CBC: 21; CUEMEM: 18
Fixed-support caption preference	200 groups	88 / 52 / 60 (CUEMEM / Tie / Flat)

component is assigned wholly to one split. Packing, shuffling, and fresh identifier assignment occur only afterward. This reduces direct leakage through exact image reuse, source grouping, persistent identifiers, and original order.

B.6 REFERENCE-PARTITION AUDIT

We conduct a manual audit to characterize the reliability and boundary of the constructed primary partition and to add a fixed-support check of cue descriptions. The audit checks whether sampled reference groups contain visible shared evidence, how often random cross-group pairs admit an alternative valid shared cue, how often major prediction disagreements across representative systems correspond to an alternative valid cue granularity, and which anonymized caption is preferred when both captioners describe the same reference support.

Visible shared evidence is present in 191 of 200 reference groups (95.5%). Among random cross-group pairs, 186 of 200 (93.0%) do not admit an additional valid cue, while 14 illustrate the expected boundary cases of open-vocabulary visual relations (Table 9). Together, these checks support the precision of the controlled primary partition while quantifying plausible alternatives rather than assuming uniqueness.

Alternative granularities account for 24 of 100 audited major disagreements for the flat model, 21 of 100 for Supervised Cluster-based Captioning, and 18 of 100 for CUEMEM, indicating that CUEMEM has the lowest rate among the audited systems.

With reference supports held fixed, the auditor prefers CUEMEM's description in 88 cases, records 52 ties, and prefers the flat captioner in 60. This comparison favors CUEMEM on the audited sample independently of differences in predicted support membership.

C METHOD IMPLEMENTATION SUPPLEMENTARY MATERIALS

This section details the training objectives, affinity computation, and support-conditioned generation used by CUEMEM.

C.1 FINE-TUNED FLAT BASELINE PROTOCOL

Flat VLMs are fine-tuned on CUESEL to jointly generate support identifiers and cue descriptions. CUEMEM uses the same annotations and output schema, deriving pair labels and visual masks from the existing reference partition without additional human annotation. It predicts supports before serializing their identifiers and generating descriptions. The output-interface control in Section D.2 separates support prediction and captioning for the flat model as well.

C.2 BACKBONE AND FINE-TUNING

Unless otherwise stated, CUEMEM and its matched flat baseline use `Qwen/Qwen3.5-4B` (Qwen Team, 2026) with the same preprocessing and chat format. LoRA adapts the q, k, v, attention-output, and feed-forward gate/up/down projections with rank 16, alpha 32, and dropout 0.05; backbone weights remain frozen. Training runs for three epochs with learning rate 10^{-4}, weight decay 0.01, warmup ratio 0.03, and gradient clipping at 1.0. The implementation uses bf16, SDPA attention, non-reentrant gradient checkpointing, and distributed data parallelism. This recipe applies to the matched Qwen3.5 comparison.

C.3 CUE-GROUP AFFINITIES FROM INTERNAL Q/K INTERACTIONS

We extract normalized pre-RoPE Q/K activations from the eight full-attention layers of Qwen3.5-4B and align the grouped-query heads. Scaled patch-to-patch scores are normalized over destination visual patches, summed within each destination image, and averaged over query patches and heads to obtain $a^{\ell}_{i\to j}$. Each layer's image affinities are symmetrized:

$$s^{\ell}_{ij} = \tfrac{1}{2}(a^{\ell}_{i\to j} + a^{\ell}_{j\to i}). \tag{11}$$

Uniform layer averaging yields bounded relation affinities $p_{ij} \in [0, 1]$ used in the relation BCE. Because the directed quantities $a^{\ell}_{i\to j}$ allocate normalized interaction mass across destination images, the symmetrized p_{ij} is interpreted as a pairwise compatibility score rather than an independently calibrated Bernoulli probability. BCE serves as a discriminative surrogate: it increases compatibility for same-cue pairs relative to different-cue pairs, while the structural decoder uses this induced separation rather than an absolute probability threshold. This interpretation matches the inference pipeline, which applies a monotone per-sample rescaling before binding.

At inference, off-diagonal entries are min–max normalized per sample, placed in a symmetric matrix with unit diagonal, clipped, and converted to logits before clustering. This deterministic transform preserves ordering up to clipping; it is separate from the training objective.

C.4 CUE DESCRIPTION CONDITIONED ON SUPPORT GROUPS

Reference outputs use blocks of the form `<id>[...]</id> <cue>...</cue>`. An additive mask restricts the full-attention visual access of every token in a block to its reference support, including identifier and delimiter tokens. Image–image interactions and causal text attention are preserved; separators outside these blocks use the standard causal mask. Qwen3.5's linear-attention layers retain their global recurrent state. Cross-entropy is averaged over all valid shifted assistant-target tokens.

The same forward pass yields the relation loss, averaged over unordered non-self image pairs labeled by reference-group membership. The shared LoRA adapters minimize $\mathcal{L}_{\text{gen}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}}$, with $\lambda_{\text{rel}} = 0.1$. The coefficient balances the numerical scales of the two losses and is fixed across the main runs. The relation-loss ablation sets it to zero while retaining the LoRA configuration, support masks, AP decoder, and generation settings.

C.5 INFERENCE PROCEDURE

At inference, one forward pass over the candidate sequence computes the Q/K affinities and initializes the decoding state. The default AP decoder uses the median off-diagonal affinity as preference, 500 maximum iterations, 15 convergence iterations, and random seed 42; it does not use the reference group count. The average-linkage variant uses threshold $\gamma_{\text{link}} = 0.2$, spectral clustering estimates its group count from thresholded connected components, and K-medoids selects the count by silhouette score.

For each predicted support, the system deterministically serializes `<id>[...]</id> <cue>` as a prefix and restricts the full-attention visual context to that support while retaining nonvisual context and the global linear-attention state. Groups are ordered by their minimum identifier. Support membership is therefore determined by relation induction and binding, not by autoregressive identifier generation.

CUEMEM-Select for Visual Haystacks. CUEMEM-Select is a training-free inference heuristic using native Qwen3.5-4B. It constructs Q/K-based groups, selects one using the query anchor, and sends those images and the original question to the same answer decoder. No parameters are updated on Visual Haystacks or CUESEL. This query-conditioned experiment complements the main benchmark by testing whether Q/K-based grouping can improve evidence selection in a related long-context retrieval setting.

Table 10: Cross-backbone results on the full test split. Every shown system has zero missing or failed records. Gray superscripts indicate scale.

Metric	Qwen3-VL^{4B}		Qwen3.5^{4B}	
	Flat	CUEMEM	Flat	CUEMEM
PVS ↑	0.334	**0.995**	0.434	**0.987**
PGF1 ↑	0.247	**0.513**	0.405	**0.747**
CSA ↑	0.246	**0.652**	0.371	**0.748**

C.6 Input and Generation Settings

Training uses a nominal sequence-length setting of 8192 with input truncation disabled. Test-time grouping includes all input images. The default total generation budget is approximately 4096 tokens, and each predicted group has a decoding-step cap of $\max(32, \lfloor 4096/\hat{M} \rfloor)$. The total budget is checked before each group and includes identifier prefixes, cue text, and formatting tokens; the last group may therefore extend beyond it.

The 507 standard test samples use pixel bounds 256^2–320^2 and the default generation budget. To accommodate memory limits, the three samples containing 1,254, 1,402, and 1,644 images use pixel bounds 128^2–192^2 and a nominal generation budget of 8192 tokens.

D Supplementary Experiments

The experiments in this section address complementary questions left open by the main comparison: whether the formulation transfers across backbones, whether its gains arise from output factorization alone or from relational supervision, how much they depend on the structural decoder, whether Q/K-based evidence selection extends beyond CUESEL, and how the automatic metrics reflect qualitative behavior. Metric definitions and prediction-repair rules are given in Section D.6.

D.1 Cross-Backbone Validation Details

Because internal Q/K geometry and long-context behavior depend on the VLM architecture, gains on Qwen3.5 alone could be specific to that checkpoint. We therefore repeat the matched flat-versus-structured comparison on Qwen3-VL-4B, asking whether relation induction and support-conditioned description remain beneficial under a second Qwen-family visual representation.

CUEMEM improves PGF1 and CSA on both evaluated 4B Qwen-family backbones (Table 10), showing that the observed benefit is not confined to the Qwen3.5 instantiation within this evaluated family. Absolute performance varies by backbone, as expected for VLM-dependent visual representations.

D.2 Output-Interface Control

The main comparison changes both how supports are inferred and how outputs are produced: the flat baseline serializes supports and descriptions jointly, whereas CUEMEM predicts structure before description and serializes identifiers programmatically. We introduce this control to determine how much of the gain can be explained by reducing the output-format burden alone, rather than by Q/K-based relation learning. The free-form flat baseline generates support identifiers and cue descriptions in a single sequence. In the support-only control, flat Qwen3.5 first predicts supports; their identifiers are serialized programmatically, and the cue descriptions are generated in a separate stage. This preserves flat autoregressive support prediction while matching the factorized output interface more closely.

The support-only control raises PGF1 by 0.148 and CSA by 0.231 over free-form generation (Table 11). Decoupling support prediction and captioning therefore improves the flat baseline. CUEMEM retains higher scores, with a PGF1 gap of 0.194 and a CSA gap of 0.146 relative to this control.

Table 11: Output-interface control for the Qwen3.5 flat baseline. The support-only control reduces free-form serialization burden before cue recaptioning.

Method	PGF1 ↑	CSA ↑
Flat Qwen3.5 free-form	0.405	0.371
Flat Qwen3.5 support-only + programmatic recaption	0.553	0.602
CUEMEM	**0.747**	**0.748**

Table 12: Visual Haystacks Multi-Needle stress test. Entries are answer accuracy (%), and columns indicate input image counts. Both pipelines use the same frozen Qwen3.5-4B answer model.

Method	5	10	20	50	100	Avg.
Qwen3.5^{4B}	63.29	62.30	60.12	56.87	54.34	59.38
CUEMEM-Select (Qwen3.5^{4B})	**64.10**	**64.05**	**63.40**	**61.55**	**60.20**	**62.66**

D.3 EXTERNAL STRESS TEST ON VISUAL HAYSTACKS MULTI-NEEDLE

Our main experiments evaluate query-free cue discovery on CUESEL, leaving open whether the underlying structural principle—selecting a small evidence-bearing image subset before semantic decoding—is useful beyond the benchmark's constructed collections and supervised partitions. We therefore use Visual Haystacks Multi-Needle (Wu et al., 2025) as an external stress test. Unlike CUESEL, this benchmark provides a query and embeds relevant images among an increasing number of distractors, allowing us to test the same evidence-selection principle under a different dataset and task formulation. This experiment is not intended as direct transfer of query-free cue discovery; rather, it probes whether Q/K structure can support inference-time evidence filtering when the relevant visual evidence is sparse.

We compare two frozen Qwen3.5-4B pipelines. Flat inference answers from the full image collection. The training-free CUEMEM-Select heuristic forms Q/K-based groups, selects one using the query anchor, and sends the selected images and original question to the same answer model. Both pipelines use identical frozen parameters, and neither is trained on Visual Haystacks or adapted with CUESEL supervision. The comparison therefore isolates the effect of Q/K-based selection from task-specific parameter learning.

The selection heuristic improves average accuracy by 3.28 percentage points (Table 12). Its advantage grows from 0.81 points at five images to 5.86 at 100 images, consistent with greater benefit in larger candidate pools. Because selection both filters evidence and shortens the decoder context, this stress test supports their combined inference-time effect rather than attributing the gain to either mechanism alone.

D.4 ABLATION OF RELATIONAL SUPERVISION

The complete system combines support-conditioned generation with an explicit pairwise relation loss, so its main result alone cannot identify whether direct relation supervision contributes beyond the structured pipeline. We therefore compare the complete objective with an otherwise matched variant that removes direct pair supervision while retaining the same Q/K affinity construction, structural decoder, and support-conditioned generation procedure. This isolates the incremental contribution of aligning internal affinities with reference same-cue relations.

Direct pair supervision raises PGF1 from 0.706 to 0.747 and CSA from 0.727 to 0.748. Both variants use Q/K affinities, reference-support masks during training, and the same structural decoder. The difference measures the incremental effect of the relation loss within this model; the generation objective continues to train the variant without pair supervision.

D.5 SENSITIVITY OF CUE-SUPPORT INFERENCE TO CLUSTERING

Pairwise affinities do not uniquely determine discrete supports: performance may depend on the algorithm used to infer group count and boundaries. We therefore test whether the learned representation

Table 13: Ablation of direct relational supervision.

Setting	PGF1 ↑	CSA ↑
Without relation supervision	0.706	0.727
With relation supervision	**0.747**	**0.748**

Table 14: Binding-operator sensitivity on the controlled ablation checkpoint. Complete and Average denote complete and average linkage.

Metric	K-med.	DBSCAN	Spectral	Complete	AP	Average
PVS ↑	**0.998**	0.693	0.920	0.843	0.982	0.880
PGF1 ↑	0.212	0.373	0.659	0.680	0.744	**0.759**
CSA ↑	0.269	0.462	0.674	0.661	**0.744**	0.710

remains useful across alternative binding operators rather than optimizing the main model separately for each decoder. We keep the affinity model and cue generator fixed within this decoder sweep and vary only the clustering operator, isolating how predicted pair relations are converted into cue-support groups. This sweep uses a separate controlled checkpoint, so its absolute values are not intended to reproduce the main-result checkpoint; group-count estimation and threshold settings are given in the implementation supplement.

Average linkage achieves the highest PGF1 in Table 14, exceeding AP by 0.015, while AP gives higher CSA (0.744 versus 0.710) and PVS (0.982 versus 0.880). This trade-off motivates AP as the default for the joint task. K-medoids attains high PVS but much lower PGF1, highlighting that output validity and support recovery capture complementary properties. The sweep compares complete decoder settings, including their respective group-count and threshold rules.

D.6 Detailed Evaluation Metrics

The task couples output validity, support recovery, and cue semantics, while generated partitions may omit, repeat, or invent identifiers. A single score would conflate these failure modes. We therefore define the repair protocol and three complementary metrics explicitly: PVS diagnoses raw serialization validity, PGF1 measures support structure after standardized repair, and CSA measures joint support–description agreement.

For each test sample, let X denote its valid input image identifiers and $\mathcal{G} = \{(G_1, c_1), \ldots, (G_M, c_M)\}$ its reference partition and cues. The raw prediction is an ordered sequence $\mathcal{P} = ((P_1, \hat{c}_1), \ldots, (P_K, \hat{c}_K))$, where each P_j is a list of emitted identifiers and may contain invalid or repeated entries. The evaluator uses different representations for validity and semantic scoring.

Prediction repair. PVS uses raw parsed identifier lists. For PGF1 and CSA, the evaluator removes invalid identifiers, deduplicates each group, and assigns an identifier appearing in multiple groups to its first occurrence in output order. Empty groups are discarded. PGF1 appends each missing input image as a singleton to obtain a complete partition. CSA retains the cues of surviving groups without adding missing-image singletons. Write $\widetilde{\mathcal{P}}$ for the repaired PGF1 partition and $(\bar{P}_j, \bar{c}_j)_{j=1}^{\bar{K}}$ for the retained CSA groups and cues. PGF1 is invariant to relabeling a valid partition, but repairing overlapping predictions is order dependent. CSA matching can also require tie-breaking when several assignments have equal overlap.

Partition Validity Score. Let U_P be the set of all emitted identifiers and $m(x)$ their total occurrence count across the raw lists, including repetitions within a group. We define

$$\begin{aligned} \text{Coverage} &= \frac{|U_P \cap X|}{|X|}, \\ \text{Exclusivity} &= \frac{1}{|X|} \sum_{x \in X} \mathbf{1}(m(x) = 1), \\ \text{ValidID} &= \frac{|U_P \cap X|}{|U_P|}. \end{aligned} \tag{12}$$

The final score is

$$\text{PVS} = \text{Coverage} \times \text{Exclusivity} \times \text{ValidID}. \tag{13}$$

When U_P is empty, ValidID and PVS are defined as zero. PVS is an output-validity diagnostic rather than a semantic grouping metric; it is therefore interpreted together with PGF1 and CSA, which assess support recovery and support–description agreement.

Pairwise Grouping F1. For any unordered image pair (x_a, x_b) with $x_a, x_b \in X$ and $a < b$, define

$$\delta_G(a,b) = \begin{cases} 1, & \exists i \text{ such that } x_a, x_b \in G_i, \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

and define $\delta_P(a,b)$ analogously using the repaired partition $\widetilde{\mathcal{P}}$. We count

$$\begin{aligned} TP &= \sum_{a<b} \mathbf{1}(\delta_G(a,b) = 1 \wedge \delta_P(a,b) = 1), \\ FP &= \sum_{a<b} \mathbf{1}(\delta_G(a,b) = 0 \wedge \delta_P(a,b) = 1), \\ FN &= \sum_{a<b} \mathbf{1}(\delta_G(a,b) = 1 \wedge \delta_P(a,b) = 0). \end{aligned} \tag{15}$$

PGF1 is $2TP/(2TP + FP + FN)$. The implementation returns zero when this denominator is zero or fewer than two valid input images are available. Filling omissions as singletons does not add positive pairs; missing members of a non-singleton reference group therefore incur false negatives.

Cue Semantic Alignment. CSA is a joint partition-and-text metric, not a standalone visual-grounding or description-quality score. It is useful for comparing complete cue-discovery pipelines because a cue description receives weight only through its matched support overlap. For the retained CSA groups, define

$$w_{ji} = \frac{|\bar{P}_j \cap G_i|}{|\bar{P}_j \cup G_i|}. \tag{16}$$

Hungarian matching maximizes total IoU and yields matched pairs $\mathcal{M}$; text similarity is not used to select the matching. For each matched pair, a frozen Qwen3-Embedding-4B encoder (Qwen Team, 2025b), denoted by ϕ, gives

$$\hat{s}_{ji} = \frac{1 + \cos(\phi(\bar{c}_j), \phi(c_i))}{2}. \tag{17}$$

The final score is

$$\text{CSA} = \sum_{(j,i) \in \mathcal{M}} \frac{|\bar{P}_j \cap G_i|}{|X|} \hat{s}_{ji}. \tag{18}$$

Unmatched groups and zero-overlap matches contribute zero, as do empty predicted group sets. We report CSA as a joint summary of matched support recovery and cue agreement, and assess description quality further through fixed-support preference and cue-compatibility audits.

PVS, PGF1, and CSA are computed per sample and macro-averaged over the 510 test cases in the full-test protocol. The same repair and aggregation rules apply to every method.

Table 15: Manual visual-compatibility check on 200 randomly sampled matched cue pairs.

Samples	Compatible	Flagged	Rate	95% CI
200	188	12	94.0%	[89.8, 96.5]

Table 16: Manual analysis of the twelve cue pairs flagged in the compatibility audit.

Flagged Issue	Count
Textually similar cues but different visual grounding	5
Visually plausible cue but overly generic description	4
Ambiguous or underspecified reference cue	3

Inference Time. For a model evaluated on T test cases, inference time is

$$\text{Time} = \frac{1}{T}\sum_{t=1}^{T}\tau_t, \tag{19}$$

where τ_t is the end-to-end wall-clock time for the t-th test case.

D.7 Manual Consistency Audit for CSA

CSA relies on embedding similarity, which alone cannot establish that two descriptions refer to the same visible evidence. We therefore complement it with a manual audit of visual compatibility rather than treating the automatic metric as sufficient evidence of grounding. We randomly sample 200 matched predicted–reference cue pairs after the same group matching used by CSA. An auditor sees the matched images, reference cue, and generated cue, and records whether the matched cue pair describes compatible visible evidence. The unit of assessment is the cue pair and its visual evidence. The resulting compatibility proportion complements the fixed-support caption preference comparison: the audit examines matched cue pairs, while the preference comparison evaluates two captioners on common reference supports.

The auditor judges 188 of 200 pairs compatible (94.0%; Wilson interval [89.8, 96.5]). The interval summarizes the sampled binary audit proportion under a binomial model. The flagged pairs involve different visual grounding despite similar wording, overly generic descriptions, and ambiguous reference cues. These cases illustrate how text similarity can differ from the quality of the shared-evidence explanation.

D.8 Cluster-based Captioning Baselines

Shared cues may correlate with ordinary visual similarity, so a strong clustering-and-captioning pipeline could recover much of the benchmark structure without using internal VLM interactions. This comparison asks two progressively stricter questions: whether frozen generic embeddings suffice when paired with the same structural decoder, and whether task-specific adaptation of a global visual embedding closes the remaining gap. We compare CueMem with conventional clustering references and Cluster-based Captioning pipelines. The latter apply affinity propagation to frozen DINOv2 (Oquab et al., 2023), DINOv3 (Siméoni et al., 2025), or SigLIP (Zhai et al., 2023) features before cue captioning. Supervised Cluster-based Captioning further adds task-specific relation learning to the SigLIP representation, providing a comparison with supervised global-embedding adaptation under the same primary cue-group annotations and AP decoder. The conventional references include DeepCluster-v2 (Caron et al., 2018; 2020), SCAN (Van Gansbeke et al., 2020), and SwAV (Caron et al., 2020).

Supervised SigLIP representation. For image I_i, the frozen SigLIP-SO400M image encoder produces a global feature $f_i \in \mathbb{R}^{d_f}$. A two-layer projection head with a GELU activation maps this feature to a relation embedding, followed by ℓ_2 normalization:

$$u_i = W_2\,\mathrm{GELU}(W_1 f_i + b_1) + b_2, \qquad z_i = \frac{u_i}{\|u_i\|_2}. \tag{20}$$

Table 17: Comparison with clustering-based grouping and Cluster-based Captioning pipelines on the full test split. Every method has zero failures. AP denotes affinity propagation.

Method	Relation Features	Binding	Cue Generator	PGF1 ↑	CSA ↑
SCAN	ResNet50-SwAV	KNN refine	—	0.065	—
SwAV	ResNet50-SwAV	K-means	—	0.161	—
DeepCluster-v2	ResNet50-DCv2	K-means	—	0.162	—
DINOv2 + AP	DINOv2-large	AP	—	0.560	—
DINOv3 + AP	DINOv3-ViT-L/16	AP	—	0.650	—
SigLIP + AP	SigLIP-SO400M	AP	—	0.685	—
Cluster-based Captioning	DINOv2-large	AP	Flat Qwen3.5^{4B} LoRA	0.560	0.614
Cluster-based Captioning	DINOv3-ViT-L/16	AP	Flat Qwen3.5^{4B} LoRA	0.650	0.669
Cluster-based Captioning	SigLIP-SO400M	AP	Flat Qwen3.5^{4B} LoRA	0.685	0.686
Supervised Cluster-based Captioning	Adapted SigLIP projection	AP	Flat Qwen3.5^{4B} LoRA	0.717	0.708
CUEMEM	Qwen3.5^{4B}	AP	CUEMEM Qwen3.5^{4B} LoRA	**0.747**	**0.748**

Here, $W_1 \in \mathbb{R}^{d_h \times d_f}$ and $W_2 \in \mathbb{R}^{d_z \times d_h}$ denote the hidden and output projections. Relation supervision updates the projection head while keeping the image encoder frozen. The adapted representation thus retains the pretrained visual features while learning a task-specific similarity space.

Pairwise relation objective. The normalized embeddings define a symmetric same-primary-group probability through a scaled and shifted cosine score,

$$p_{ij}^{\mathrm{Sig}} = \sigma\big(a\, z_i^\top z_j + b\big), \qquad a > 0, \tag{21}$$

where σ is the sigmoid function and a and b are the score scale and bias. For each training collection, $y_{ij} = 1$ when images i and j belong to the same primary reference group, and $y_{ij} = 0$ otherwise. The objective is the mean binary cross-entropy over all unordered, non-self image pairs within that collection:

$$\mathcal{L}_{\mathrm{Sig}} = -\frac{2}{N(N-1)} \sum_{i<j} \Big[y_{ij} \log p_{ij}^{\mathrm{Sig}} + (1 - y_{ij}) \log(1 - p_{ij}^{\mathrm{Sig}}) \Big]. \tag{22}$$

Each pair contributes once with equal weight; positive and negative pairs follow the reference partition without class reweighting or negative subsampling. This aligns the pair-label definition and within-collection BCE averaging with CUEMEM's relational objective. The SigLIP projection is trained on image-pair relations, independently of the captioning objective.

Support decoding and cue description. At inference, adapted pairwise similarities provide the affinity input to AP, using the same binding rule as CUEMEM with the median off-diagonal affinity as preference and no reference group count. The predicted supports are then described by the flat Qwen3.5-4B LoRA trained on CUESEL support–description targets, also used in the frozen-feature pipelines. Captioning leaves the support assignments unchanged. PGF1 therefore evaluates the support structure recovered by the adapted representation and AP, while CSA evaluates the resulting support–description pairs. In CUEMEM, relational and generative supervision instead update a shared VLM. This supervised pipeline is referred to as Supervised Cluster-based Captioning.

SigLIP + AP reaches 0.685 PGF1, compared with 0.162 for DeepCluster-v2. Captioning leaves these support assignments and PGF1 scores unchanged. Supervised Cluster-based Captioning improves to 0.717 PGF1 and 0.708 CSA, while CUEMEM reaches 0.747 and 0.748. These results show that explicit AP support clustering over strong embeddings is a competitive structured control, and that the jointly adapted VLM achieves higher support recovery and joint cue agreement among the evaluated systems.

D.9 LDR BUCKET DETAILS

The evidence-dilution analysis predicts that support recovery should become harder as relevant images grow sparse relative to distractors. Aggregate test scores cannot show whether the structured model specifically degrades more gracefully in this regime, so we stratify PGF1 by log distractor ratio (LDR). Because LDR also covaries with collection length and the number of competing groups,

Table 18: PGF1 by log distractor ratio bucket. Gap denotes CUEMEM minus the flat Qwen3.5-4B baseline.

Bucket	LDR Range	Mean LDR	Qwen3.5-4B	CUEMEM	Gap
Low	[0.00, 3.62]	2.67	0.845	**0.886**	0.041
Medium	[3.62, 4.36]	4.04	0.522	**0.788**	0.267
High	[4.37, 4.74]	4.56	0.184	**0.730**	0.546
Extreme	[4.74, 6.32]	4.96	0.070	**0.673**	0.604

Table 19: Sampled runtime by input-length bucket. Times are measured in seconds and averaged over the selected timing runs. Red. denotes relative reduction against flat Qwen3.5-4B.

Length	Flat Total	CUEMEM Total	Total Red.	Flat Decode	CUEMEM Decode	Decode Red.
1–8	3.5	2.1	39.4%	3.1	1.6	47.2%
9–16	4.5	2.8	37.6%	3.9	2.2	44.4%
17–32	10.7	5.8	45.2%	9.5	4.6	51.9%
33–64	19.5	10.9	44.2%	17.3	8.4	51.4%
65–128	36.5	21.0	42.5%	32.4	16.3	49.6%
129–256	107.6	42.5	60.5%	96.4	30.1	68.8%
257–512	163.5	87.5	46.5%	142.6	59.3	58.4%
513+	427.3	175.9	58.8%	367.6	72.4	80.3%

this analysis tests robustness under increasing distractor pressure rather than identifying a purely causal sparsity effect. Table 18 reports the values underlying the main-paper analysis. Gaps use the unrounded plotted means and can differ from subtraction of rounded table entries; bucket ranges and plotted error magnitudes are reported for transparency.

D.10 RUNTIME BUCKET DETAILS

Explicit relation estimation introduces structural computation, but support-conditioned generation can shorten decoding by avoiding a single long free-form partition serialization. We measure complete-pipeline runtime to determine whether the added structural stage is offset by reduced decoding and how this trade-off changes with collection length. Table 19 reports a stratified wall-clock comparison under one shared implementation environment. For each input-length bucket, at most five cases are selected at approximately evenly spaced positions after sorting by image count. Each selected case is timed after one warm-up pass under matched preprocessing, prompt format, and nominal output budgets. The measurements characterize the relative end-to-end behavior of these implementations rather than hardware-independent complexity.

D.11 QUALITATIVE RESULT COMPARISON

Aggregate metrics indicate whether predictions agree with the reference partition, but not what visual relations the systems recover or how grouping errors affect their descriptions. We therefore inspect representative outputs from both methods to make the support–semantics interaction concrete and to characterize differences hidden by scalar scores. These examples are illustrative rather than an additional estimate of overall performance.

Figure 6 provides illustrative outputs complementing the quantitative evaluation. In the protest example, the flat baseline identifies some images under a broad landscape or illumination cue but also includes distractors relative to the reference group. CUEMEM more closely recovers that reference support set and describes shared illumination and sign-holding evidence. In the horseback-riding example, the flat output splits evidence into partially overlapping cues, while CUEMEM produces a group closer to the reference partition. These examples illustrate how support recovery and cue description interact in representative long-collection cases.

D.12 FAILURE CASE ANALYSIS

The average gains do not establish that explicit relation learning resolves ambiguous cue granularity or all support errors. We analyze failures to identify where the structured factorization still breaks



Figure 6: Qualitative comparison between the reference annotations, the flat Qwen3.5-4B baseline, and CUEMEM. The examples show that the flat baseline can recover locally plausible cues, but may mix unrelated images, prefer coarse scene-level regularities, or split evidence into inconsistent groups. CUEMEM more closely matches the reference cue-centered groups in these selected cases.



Figure 7: Representative failure cases of CUEMEM and the flat Qwen3.5-4B baseline. Red boxes denote images assigned to the target cue, and blue boxes denote other predicted groups. The examples illustrate cue overgeneralization and cue fragmentation under ambiguous visual evidence.

down and whether errors arise from overly broad relations or incomplete evidence aggregation. This complements the successful examples and bounds the interpretation of the quantitative improvements.

Figure 7 illustrates two boundary cases of shared cue discovery. In the first, the reference describes similar illumination, white tables with matching color and texture, and people making clothes; CUEMEM selects related indoor images but gives a broader description. This is *cue overgeneralization*: a coarser relation replaces the annotated conjunction, changing both support and semantics. In the second, a cue involving hair, a bow, and clothing is split into separate descriptions of carousel context and pink clothing. This is *cue fragmentation*: locally plausible descriptions recover only part of the complete reference group. These examples highlight the need to jointly evaluate support membership and cue semantics, especially when multiple plausible visual relations coexist.

E LIMITATIONS

Data and reference structure. CUESEL packs short VIST cue groups into long collections, and its disjoint primary partition provides a consistent target for controlled evaluation. The audits quantify both within-group validity and plausible alternatives, so reported metrics should be interpreted as agreement with this controlled reference rather than an exhaustive enumeration of every possible visual relation. Source-story and exact-URL isolation reduce direct reuse across splits; overlapping cues and naturally occurring collections remain complementary settings for future evaluation.

Method and attribution. CUEMEM predicts image-level supports and descriptions without requiring spatial cue localization. Global representations remain available during support-conditioned generation, while support membership is determined before cue generation. The flat comparison jointly changes relational supervision, output serialization, and visual access; the relation-loss and decoder ablations isolate key components, and the supervised embedding comparison evaluates an alternative complete structured pipeline.

Description quality and efficiency. CSA measures joint support–text agreement rather than standalone caption faithfulness; the compatibility audit and fixed-support preference study provide complementary human evidence. Runtime results are sampled complete-system wall-clock comparisons under matched preprocessing and output-budget settings, not hardware-independent complexity estimates.

F ETHICS STATEMENT

CUESEL is derived from the publicly available VIST dataset and is used to study shared visual cue discovery in image collections. It is intended for diagnostic evaluation rather than identity inference or surveillance. Annotators were instructed to describe visible shared content and avoid unsupported personal attributes. Descriptions of recurring appearance should not be treated as verified identity or inferred personal characteristics. Future releases should preserve source-dataset license requirements and include documentation for responsible use.