SRET: SPATIAL RELATION-AWARE EVIDENCE TRANSPORT FOR VISUAL TOKEN REDUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual token reduction is key to lowering vision language inference costs, but redundancy in object semantics does not imply redundancy in relational evidence. Spatial reasoning relies on information distributed across objects, reference landmarks, and observations. Importance-based deletion or semantic merging may preserve object appearance while disrupting the structure needed to infer direction, position, and viewpoint changes. We propose SRET, a training-free method for spatial relation-aware evidence transport. Inspired by geometric constraints on spatial observations, SRET formulates compression as relational representative selection and evidence transport. Visual semantics, shallow context, and local layout define token relations, guiding coverage repair and cross-group budget adjustment under a fixed budget. Based on source–anchor relations, source evidence mass is split into a node component and a relational residual, both transported to structure-aware receivers. Selection and aggregation share evidence weights, while tangential updates preserve anchor norms. Experiments on two vision language models and multiple spatial reasoning benchmarks demonstrate that SRET preserves spatial reasoning ability while compressing visual sequences.

1 INTRODUCTION

Vision language models (Liu et al., 2023; Alayrac et al., 2022; Bai et al., 2023; Chen et al., 2024c) are expanding beyond object recognition and captioning to multi-view understanding, video analysis, and spatial interaction, demanding better observation integration and reference-frame transformations (Yang et al., 2026; Yin et al., 2025; Li et al., 2026). Questions such as *Which object is behind the chair?* or *How do left–right relations change from the person's viewpoint?* require recognizing objects, establishing reference frames, and relating regions or observations (Johnson et al., 2017; Hudson & Manning, 2019; Santoro et al., 2017). Higher resolutions, more views, and longer videos enrich evidence but substantially lengthen visual sequences. Spatial reasoning thus poses a structurally constrained efficiency problem: how can visual sequences be compressed while preserving evidence for relational judgments?

Existing methods primarily exploit importance and redundancy via **importance-driven selection**, which retains high-scoring tokens and removes the rest (Rao et al., 2021; Chen et al., 2024b), and **redundancy-driven aggregation**, which pools similar or complementary content into a few representatives (Bolya et al., 2022; Shang et al., 2025; Yang et al., 2025). Yet tokens with similar semantic content may provide different spatial references. A low-attention region may delineate a target boundary, visually similar regions may lie on opposite sides of a target, and a less salient view may resolve an otherwise ambiguous front–back relation. Token removal should therefore depend on whether the retained set jointly explains the evidence, not merely its individual score.

Spatial substitutability depends not only on semantic similarity, but also on observation-local geometry: similar-looking regions may serve different reference roles, and 2D coordinates across views do not directly denote the same spatial location. We instantiate this principle through visual, contextual, and layout compatibility within each observation, using observation-local coordinates rather than shared 3D geometry. Under a fixed token budget, these constraints guide representative selection and feature aggregation: compression should preserve not only appearance but also references supporting direction, position, and viewpoint judgments.

We propose **SRET**, a training-free method that couples representative selection with evidence transport through within-group token relations. Starting from query-selected representatives, SRET re-

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.

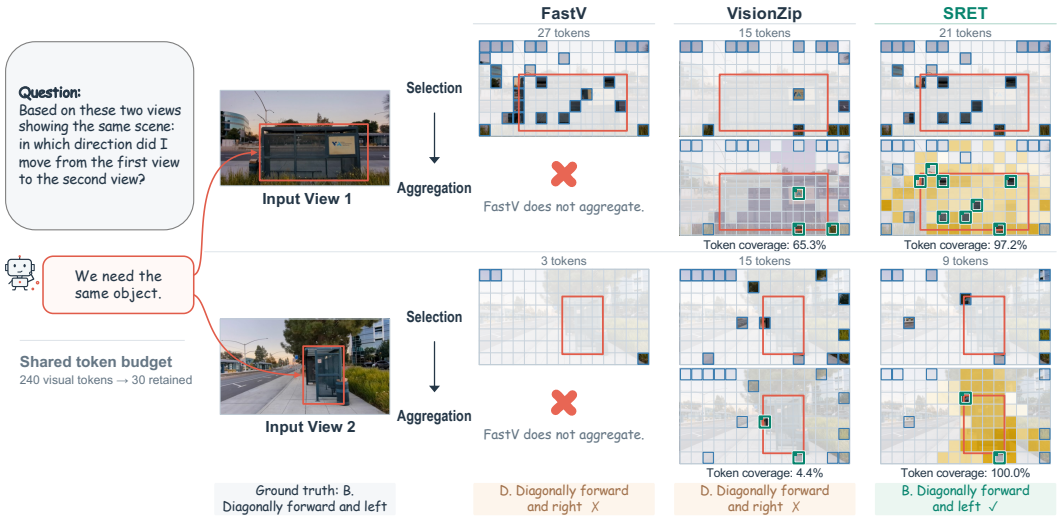

Figure 1. Comparison of FastV, VisionZip, and SRET on a cross-view motion question under a shared 240-to-30 visual token budget. Aggregation panels color source tokens covered by anchors in or near the red boxes. Token coverage is the colored-area fraction within each box. SRET more completely covers the shared landmarks across both views and answers correctly.

fines their positions and reallocates group budgets when needed. The same relational coverage scores and source weights guide the distribution of removed evidence to the retained set, connecting where tokens are kept with what information they receive. SRET uses existing model features and image coordinates without requiring shared 3D coordinates across views. Figure 1 illustrates the resulting evidence preservation on a cross-view motion question.

We evaluate two models on spatial reasoning benchmarks spanning single-image, multi-view, and video inputs. SRET achieves strong overall accuracy across compression budgets, especially under tight budgets. Request FLOPs and prefill KV cache characterize the trade-off between relational evidence preservation and resource budgets. Our main contributions can be summarized as:

1. We recast spatial visual-token compression as a relational evidence preservation problem rather than independent importance or semantic redundancy estimation. We model whether one token can substitute for another using visual, contextual, and local-layout compatibility, yielding a unified coverage objective for representative selection and budget allocation.
2. We develop a relation-aware evidence transport rule that couples where tokens are retained with where removed evidence is written. Source evidence is split according to reconstructability and transported through structure-aware receivers, while norm-preserving updates and frozen transport maps maintain stable propagation through later layers.
3. Extensive evaluation shows that relational evidence preservation is especially beneficial in the low-token regime. Across two VLMs, eight spatial benchmarks, and five retention ratios, SRET consistently improves compressed-model accuracy while enabling favorable FLOPs–accuracy and KV-cache–accuracy trade-offs.

2 Related Work

2.1 Visual Token Reduction in Vision Language Models

Visual token reduction shortens the sequences processed by vision language models. Learned interfaces, including the resampler in Flamingo (Alayrac et al., 2022) and the Q-Former in BLIP-2 (Li et al., 2023), produce compact visual representations through trained modules. TokenPacker (Li et al., 2025) enriches coarse visual queries with local visual details, while LLaVA-Mini (Zhang et al., 2025) fuses visual information into text representations before compressing the visual sequence. Training-free methods reduce tokens using representations available in pretrained models. FastV (Chen et al., 2024b) prunes visual tokens according to attention from early language-model layers. VisionZip (Yang et al., 2025) retains dominant tokens and aggregates the remaining tokens

into contextual representatives through key-feature matching. LLaVA-PruMerge (Shang et al., 2025) combines token selection with feature aggregation, while SparseVLM (Zhang et al., 2024) combines text-guided pruning with token recycling. DivPrune (Alvar et al., 2025) selects a representative subset by maximizing feature diversity. Beyond token importance and feature redundancy, SRET uses spatially constrained relations and shared source weights to jointly guide representative selection and the transport of removed evidence.

2.2 Spatial Reasoning and Spatially Structured Compression

Spatial reasoning requires representations that support object localization, relative-position judgments, and changes of viewpoint. SpatialVLM (Chen et al., 2024a) develops these capabilities through spatial supervision, SpatialRGPT (Cheng et al., 2024) combines region representations with depth information, and SpatialBot (Cai et al., 2025) learns from RGB and depth inputs. Our focus is preserving these capabilities when compressing the visual sequence of a fixed pretrained model. Recent compression methods have begun to consider spatial organization. GridPrune (Duan et al., 2026) allocates query-dependent budgets to image regions before selecting tokens within each region. S^2Prune (Jia et al., 2026) adapts regional token density to local image structure. Nüwa (Huang et al., 2026) uses regional representatives and spatially constrained aggregation, followed by text-guided pruning in the language model. CoverPrune (Ling et al., 2026) formulates representative selection through a coverage objective with transport costs based on features, global 3D coordinates, and time. CoVeR (Bui et al., 2026) selects tokens through coverage in 3D coordinate space. SRET models within-group token relations to refine anchors and guide source-feature aggregation, while allowing retention slots to move across groups. It uses existing model features and image coordinates without requiring a shared 3D coordinate system across views.

3 Method

As shown in Figure 2, SRET performs Relation Modeling, Anchor Adjustment, and Evidence Transport: establishing spatial substitutability, refining anchor positions and group budgets under a fixed total token budget, and writing removed evidence into retained representations.

3.1 Query-Driven Relation Modeling

3.1.1 Query-guided Anchor Initialization

Let $H = \{h_i\}_{i=1}^{N}$ denote the visual representations at the input to the third decoder block. Given a retention ratio ρ, SRET selects $K \approx \rho N$ visual positions as anchors. Tokens are divided into layout groups according to input preprocessing, such as images, sampled frames, or image tiles. The views in Figure 2 denote these groups. All text tokens and the original positional encodings of retained visual tokens are preserved.

Head-averaged attention from the last text position to visual positions provides query scores d_i. We initialize the anchors as $A_0 = \mathrm{TopK}(d, K)$. The final vision-encoder layer also provides visual saliency scores v_i. After separately normalizing d_i and v_i over all visual positions, we define shared evidence weights:

$$m_i = \max\{\bar{d}_i, \bar{v}_i\}, \qquad M = \sum_i m_i. \tag{1}$$

Here, $\bar{d}_i$ and $\bar{v}_i$ each sum to one. Taking their maximum allows both query-relevant and visually salient sources to influence compression. The same weights are used for relational coverage, the reference allocation across groups, and feature aggregation.

3.1.2 Within-group Token Relation Construction

We construct relations between all tokens within each group to evaluate which positions can represent a source token. Let z_i^v and z_i^c denote its normalized visual-encoder key features and shallow contextual features, respectively. Its normalized layout coordinate is $\tilde{p}_i \in [0,1]^D$, where $D = 3$ consists of one group-constant component and two image-coordinate components. We define visual,

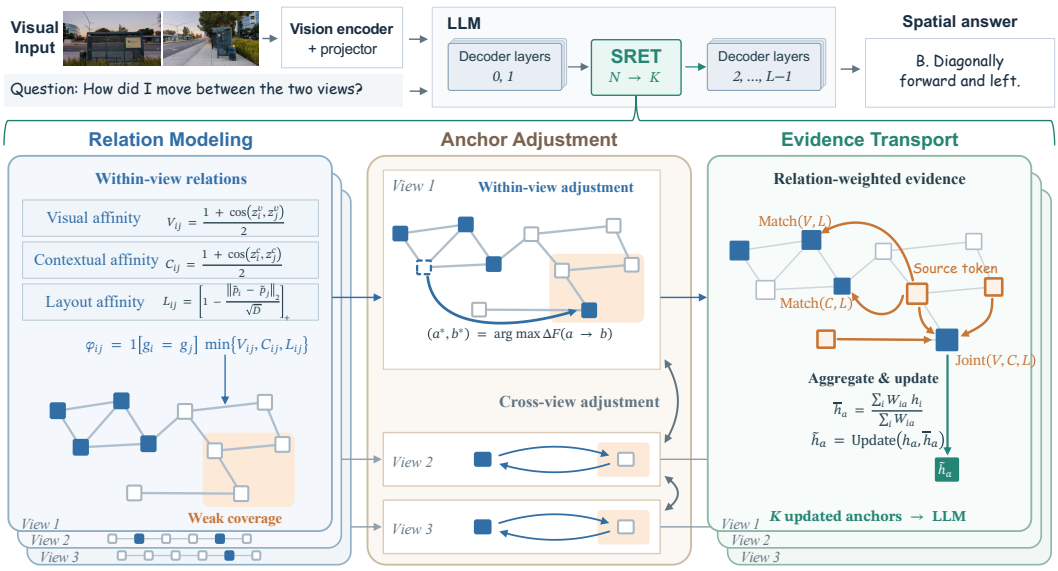

Figure 2. SRET framework. At the third decoder block's input, N visual tokens are compressed to K through relation modeling, anchor adjustment, and evidence transport. Anchor adjustment refines retained positions and group budgets; transport aggregates evidence while preserving anchor scale. Blue, white, and green nodes denote retained, unselected, and updated tokens. Dashed outlines mark replaced anchors; orange regions and arrows indicate weak coverage and evidence transport.

contextual, and layout affinities as

$$V_{ij} = \frac{1 + \cos(z_i^v, z_j^v)}{2}, \quad C_{ij} = \frac{1 + \cos(z_i^c, z_j^c)}{2}, \quad L_{ij} = \left[1 - \frac{\|\tilde{p}_i - \tilde{p}_j\|_2}{\sqrt{D}}\right]_+, \tag{2}$$

where $[x]_+ = \max(x, 0)$. With g_i denoting the group of token i, the relation weight and coverage score are

$$\phi_{ij} = \mathbf{1}[g_i = g_j] \min\{V_{ij}, C_{ij}, L_{ij}\}, \qquad c_i(A) = \max_{a \in A} \phi_{ia}. \tag{3}$$

Taking the minimum yields a high relation weight only when visual, contextual, and layout compatibility are all high, while $c_i(A)$ measures how well the best anchor represents source i. This shared score guides anchor selection and the directly transported fraction. Relations remain within groups; subsequent language layers combine observations across groups. Appendix B.1 details feature and coordinate definitions.

3.2 Budget-Constrained Anchor Adjustment

SRET refines A_0 through within-group anchor replacement and cross-group budget adjustment. Both operations optimize the same objective under a fixed total budget K, with query scores, evidence weights, and token relations held constant.

3.2.1 Coverage-Guided Within-Group Anchor Adjustment

Query-based initialization may concentrate anchors in a few regions. We evaluate an anchor set by combining relational coverage with retained query attention:

$$F(A) = \frac{1}{M} \sum_i m_i c_i(A) + \frac{\sum_{a \in A} \bar{d}_a}{\sum_{a \in A_0} \bar{d}_a}, \qquad |A| = K. \tag{4}$$

The terms balance evidence-weighted coverage against query attention retained relative to the initial Top-K set. Replacing anchor $a \in A$ with unselected token $b \notin A$ gives gain $\Delta F(a \to b) = F((A \setminus \{a\}) \cup \{b\}) - F(A)$. Within-group exchanges preserve group counts and are accepted when gains exceed numerical tolerance.

3.2.2 CROSS-GROUP BUDGET ADJUSTMENT

Within-group replacement cannot compensate for too few anchors allocated to a group. Let q_g^0 denote the initial anchor count in group g. We derive a reference allocation from its share of evidence weight and measure the mismatch with the initial allocation:

$$q_g^* = K\frac{\sum_{i:g_i=g} m_i}{M}, \qquad \Delta_{\text{cap}} = \frac{1}{K}\sum_g |q_g^* - q_g^0|. \tag{5}$$

Cross-group exchanges are enabled when the initial mismatch $\Delta_{\text{cap}} > 1/3$; otherwise, group counts remain fixed. The threshold is shared across models, benchmarks, and budgets. The reference allocation determines whether to enable these exchanges, rather than imposing quotas on the final counts.

Each cross-group exchange removes an anchor from one group and adds an unselected token from another, transferring one retention slot while preserving K. As with within-group replacement, gains in F determine whether an exchange is accepted and how the final budget is distributed.

Joint optimization. Each round jointly considers within-group and enabled cross-group exchanges over all unselected tokens, applying positive-gain exchanges with disjoint group endpoints in parallel. Optimization stops when no allowed gain exceeds numerical tolerance; Appendix B.1 and Algorithm 1 detail proposal ranking and execution.

3.3 STRUCTURE-PRESERVING EVIDENCE TRANSPORT

3.3.1 SOURCE-TO-ANCHOR ASSIGNMENT AND AGGREGATION

Given the adjusted anchor set A, SRET distributes each source token's evidence weight m_i among the retained anchors.

For a non-anchor source token i, let $a_i = \arg\max_{a\in A}\phi_{ia}$ be its most compatible retained anchor and $c_i = c_i(A)$. We assign a fraction c_i of its routing weight to a_i and split the remaining fraction $1-c_i$ between two receivers defined below. These portions are the node component and the relational residual, respectively. The split applies to routing weights: all receivers obtain weighted contributions from the same source representation h_i.

The visual and contextual endpoints b_i^v and b_i^c correspond to $\mathrm{Match}(V, L)$ and $\mathrm{Match}(C, L)$ in Figure 2. Let C_{ia}^{tr} denote contextual compatibility in the transport receiver space, x_i the 2D component of $\widetilde{p}_i$, and $\psi_{ia} = [1 - \|x_i - x_a\|_2/\sqrt{2}]_+$. Both endpoints balance feature and layout compatibility:

$$\begin{aligned} b_i^v &= \arg\max_{a\in A:\, g_a=g_i} \frac{V_{ia}+\psi_{ia}}{2}, \\ b_i^c &= \arg\max_{a\in A:\, g_a=g_i} \frac{C_{ia}^{\text{tr}}+\psi_{ia}}{2}. \end{aligned} \tag{6}$$

If no same-group anchor exists, selection falls back to global feature matching.

For distinct-position endpoints in the source's group, the source's projection onto their segment splits the residual:

$$t_i = \mathrm{clip}_{[0,1]}\left(\frac{\langle \widetilde{p}_i - \widetilde{p}_{b_i^v}, \widetilde{p}_{b_i^c} - \widetilde{p}_{b_i^v}\rangle}{\|\widetilde{p}_{b_i^c} - \widetilde{p}_{b_i^v}\|_2^2}\right). \tag{7}$$

Transport proportions are

$$P_{ia} = c_i\mathbf{1}[a=a_i] + (1-c_i)\big((1-t_i)\mathbf{1}[a=b_i^v] + t_i\mathbf{1}[a=b_i^c]\big). \tag{8}$$

Each row has at most three receivers and sums to one; anchor sources instead use $P_{ia} = \mathbf{1}[a=i]$ to preserve self-loops. The Joint and two Match branches in Figure 2 form this sparse directed transport graph. Aggregation uses shared mass:

$$W_{ia} = m_i P_{ia}, \qquad \mu_a = \sum_i W_{ia}, \qquad \bar{h}_a = \frac{\sum_i W_{ia}h_i}{\mu_a}. \tag{9}$$

Received mass μ_a and mean $\bar{h}_a$ satisfy $\sum_a \mu_a = M$. Local projection additionally ties residual allocation to source–endpoint geometry.

3.3.2 NORM-PRESERVING ANCHOR UPDATES

To incorporate $\bar{h}_a$ without changing the anchor's norm, we update h_a using the component of $\bar{h}_a - h_a$ orthogonal to h_a. Let $u_a = h_a/\|h_a\|_2$ and $T_a(x) = x - \langle x, u_a\rangle u_a$. The update is

$$r_a = T_a(\bar{h}_a - h_a), \qquad s_a = \frac{\|h_a\|_2}{\|h_a\|_2 + \|r_a\|_2}, \quad \widetilde{h}_a = \|h_a\|_2 \frac{h_a + s_a r_a}{\|h_a + s_a r_a\|_2}. \tag{10}$$

The factor s_a attenuates large tangential updates, and normalization restores anchor norm. Without learning a mixing coefficient, this rule preserves nonzero anchor norms and bounds directional rotation below 45° in real arithmetic.

3.4 FROZEN TRANSPORT MAPS AND EFFICIENT INFERENCE

For Qwen3-VL's subsequent DeepStack injections (Bai et al., 2025), SRET reuses a request-local frozen transport map without reselecting endpoints or re-estimating parameters. This map is linear in the increment and consistent with initial transport; InternVL3 has no such injections. Full visual encoding and the first two decoder blocks remain, while later blocks process retained visual representations and all text tokens. Appendices B.1–B.3 provide execution details and map properties.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We use SenseNova-SI-1.3-Qwen3-VL-8B and SenseNova-SI-1.3-InternVL3-8B (Cai et al., 2026), based on Qwen3-VL and InternVL3 (Bai et al., 2025; Zhu et al., 2025), and use the backbone names below. SRET shares compression rules and hyperparameters across models, benchmarks, and budgets without additional training.

Benchmarks and budgets. We evaluate on MMSI-Bench, MindCube Tiny, BLINK, EmbSpatial, ViewSpatial, 3DSRBench, SITE, and ReVSI, covering single-image, multi-view, and video inputs. SITE's image and video subsets are scored separately and equally averaged as one benchmark; ReVSI uses 16 frames. Full accuracy evaluation uses all valid examples at 50%, 25%, 15%, 10%, and 5% visual token retention. Full denotes the uncompressed reference.

Baselines. Besides Full, we compare FastV, VisionZip, DivPrune, GridPrune, and six diagnostic baselines: Random, Uniform Spatial, Vision Attention, Query Relevance, Feature Diversity, and Query Spatial. Random averages three seeds; other methods use a fixed seed. Appendix A gives implementation and evaluation details.

Inference and resource measurement. All methods use FlashAttention-2 (Dao et al., 2022; Dao, 2024). Accuracy evaluation generates at most 16 tokens; videos sample at most 32 frames except ReVSI's fixed 16. Preprocessing and invalid-video exclusions are identical across methods. Resource measurement uses an NVIDIA A100-SXM4-80GB and 50 fixed inputs per benchmark, shared across methods, with generation fixed to 16 tokens. Estimated request FLOPs and logical prefill KV cache sizes are equally averaged over benchmarks and paired with full-set accuracy.

4.2 FULL-SET ACCURACY

Table 1 reports macro-average official scores over eight benchmarks, while Avg further averages over two models and five retention budgets (Appendix A.2). Across the ten model–budget settings, SRET achieves the strongest overall compressed performance, with an Avg of 55.426%, exceeding the strongest baseline average by 2.060 percentage points (pp). It ranks first among compressed methods in nine settings and second in the remaining one.

The advantage of SRET becomes more pronounced as the token budget decreases. From 25% to 5% retention, SRET outperforms all compression baselines on both backbones at every budget, whereas the differences are smaller at 50% retention, where more original visual evidence remains available. Among the diagnostic baselines, feature diversity and uniform spatial sampling generally outperform visual-saliency- or input-level query-similarity-based selection, indicating that broader content and spatial coverage can be beneficial under compression. SRET improves further by combining relation-aware representative selection with structured transport of removed-token evidence.

Table 1. Full-set eight-benchmark macro scores (%). Avg equally averages two models and five budgets. Full is uncompressed; Random averages three seeds.

	Qwen3-VL					InternVL3					
Method	50%	25%	15%	10%	5%	50%	25%	15%	10%	5%	**Avg**
Full	60.07	60.07	60.07	60.07	60.07	60.74	60.74	60.74	60.74	60.74	60.41
Random	57.37	53.05	49.02	46.15	42.27	59.70	56.05	53.20	51.15	47.14	51.51
Uniform Spatial	58.14	55.10	52.17	49.39	43.64	59.33	57.21	54.73	51.55	46.89	52.82
Vision Attention	58.74	55.84	52.28	49.00	42.80	59.42	56.01	52.18	48.40	42.22	51.69
Query Relevance	53.13	46.85	43.83	41.67	37.99	54.93	47.29	42.86	40.36	38.94	44.79
Feature Diversity	58.73	55.41	52.40	49.43	44.47	59.62	56.90	54.39	51.44	46.99	52.98
Query Spatial	55.98	50.03	46.57	43.87	39.65	57.99	52.59	48.23	45.08	40.93	48.09
FastV	58.27	54.22	50.48	47.64	42.56	60.16	58.98	56.95	54.39	50.03	53.37
DivPrune	57.82	54.75	50.88	48.29	43.57	59.73	56.80	54.55	52.32	48.81	52.75
GridPrune	58.83	55.26	48.51	45.16	41.29	60.07	57.03	52.18	48.89	44.27	51.15
VisionZip	**59.18**	56.06	53.28	49.93	43.35	59.62	55.92	52.42	48.70	42.38	52.08
SRET	59.02	**56.98**	**54.18**	**52.03**	**47.05**	**60.45**	**59.40**	**57.91**	**55.85**	**51.39**	**55.43**

Table 2. Module ablations. Sel.: relational coverage selection; Trans.: structure-preserving transport. Setting (5) removes layout constraints in both modules.

			Qwen3-VL						InternVL3					
Setting	Sel.	Trans.	50%	25%	15%	10%	5%	Avg	50%	25%	15%	10%	5%	Avg
(1) Query pruning	×	×	57.5	53.1	49.4	46.7	41.3	49.6	57.8	56.8	53.7	50.7	48.3	53.5
(2) Selection only	✓	×	57.4	54.0	51.2	48.3	44.2	51.0	57.7	56.1	54.8	52.0	48.6	53.8
(3) Transport only	×	✓	**57.8**	55.0	50.9	49.0	44.0	51.3	57.8	**57.1**	53.6	51.9	48.9	53.8
(4) Full SRET	✓	✓	57.2	**56.2**	**52.9**	**50.5**	**45.5**	**52.5**	**57.9**	56.8	**55.5**	**53.1**	**49.2**	**54.5**
(5) No layout	✓	✓	57.5	54.6	51.8	48.2	44.2	51.3	57.4	56.9	54.5	52.5	48.5	54.0

SRET also exhibits the smallest average degradation relative to Full. Paired analysis in Appendix B.6 shows that its gains arise primarily from preserving answers that Full already solves correctly, rather than from recovering more previously unsuccessful examples. This pattern is consistent with the intended role of SRET as a compression method that preserves existing spatial-reasoning evidence.

4.3 Core Mechanism Ablations

4.3.1 Contributions of Relational Selection and Evidence Transport

We evaluate the two core components of SRET from complementary perspectives: module ablations measure the contributions of relational selection and evidence transport, while fixed-anchor interventions examine whether transport benefits depend on preserving content–path correspondence. Both analyses use the same fixed panel of 200 inputs per benchmark.

Table 2 combines relational coverage selection (Sections 3.1–3.2) and structure-preserving transport (Sections 3.3–3.4). Setting (1) applies query Top-K pruning. Settings (2) and (3) add relational selection or evidence transport individually, while setting (4) enables both components. Setting (5) removes explicit layout coordinates and grouping constraints from the complete method while retaining the backbone's native positional encoding. All settings share the same token budgets, inputs, and inference configurations, and results are reported separately for each backbone.

Selection and transport provide complementary gains. Averaged equally over the ten model–budget settings, relational selection and structure-preserving transport improve over query pruning by 0.902 and 1.067 pp, respectively, while combining them yields a 1.972 pp improvement. The near-additive gain indicates that the two components capture largely complementary aspects of compression: selection improves which visual positions are retained, whereas transport improves what evidence those retained positions receive. At wider budgets, where more original visual evidence remains directly available, the benefit of jointly modifying both retained positions and received content becomes smaller; transport alone performs better at Qwen3-VL 50% and InternVL3 25%.

Table 3. Fixed-anchor transport interventions at 10%, 15%, and 25% retention. Within-group shuffling reports mean ± sample standard deviation over three seeds.

	Qwen3-VL			InternVL3		
Transport	10%	15%	25%	10%	15%	25%
Relational (SRET)	**50.5**	**52.9**	**56.2**	**53.1**	**55.5**	**56.8**
Anchors only	48.3	51.2	54.0	52.0	54.8	56.1
Source shuffle	48.0 ± 0.6	50.8 ± 0.3	53.9 ± 0.6	50.9 ± 0.2	53.9 ± 0.4	56.1 ± 0.6

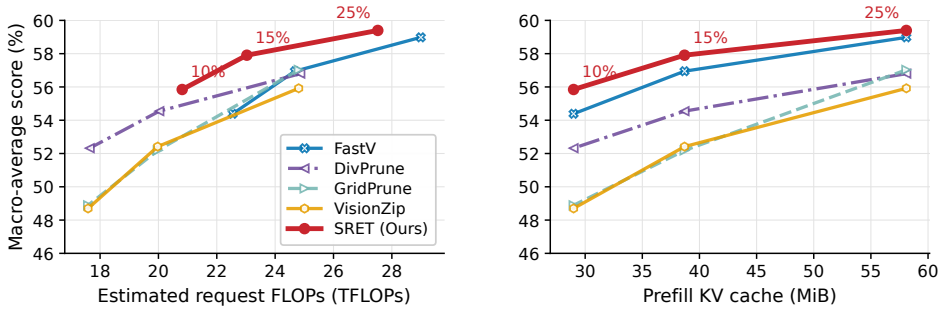

Figure 3. Accuracy–resource trade-offs on InternVL3. Horizontal axes show estimated request FLOPs (left) and logical prefill KV cache (right); the vertical axis is full-set eight-benchmark average accuracy. Curves connect 25%, 15%, and 10% budgets. Costs use 400 shared fixed inputs and 16 generated tokens.

Explicit layout complements feature matching. Full SRET outperforms the no-layout variant in eight of ten settings, with an average gain of 0.890 pp. Removing explicit grouping and local-coordinate constraints therefore degrades performance even though native positional encoding and feature similarity remain available. This indicates that implicit positional information in the backbone does not fully replace the observation-local structure used by SRET. Since setting (5) removes layout constraints from both selection and transport, this ablation should be interpreted as measuring their combined contribution rather than isolating a single geometric term.

4.3.2 Transport-Correspondence Interventions with Fixed Anchors

Table 3 isolates the effect of transport correspondence by fixing SRET's retained anchors and token budget. We compare normal transport, anchor retention without transport, and within-group source shuffling. The shuffling intervention preserves the source-feature multiset, transport weights, and received mass, while permuting the correspondence between non-anchor source features and transport slots within each group. Retained anchors remain unchanged, and subsequent visual increments follow the same shuffled correspondence. By holding selection and aggregate transport mass fixed, this intervention tests whether transport benefits depend on routing source content to appropriate receivers rather than merely injecting additional removed-token information.

Transport benefits depend on content–path correspondence. Normal transport outperforms both anchor retention without transport and mean shuffled transport in all six settings, by 1.461 and 1.916 pp on average, respectively. Moreover, shuffling performs worse than no transport in five of the six settings. Because the retained positions, source-feature multiset, transport weights, and received mass are preserved, the degradation under shuffling cannot be explained by reduced source content or altered anchor selection. Instead, the results show that aggregation is beneficial only when removed-token content is assigned through appropriate transport paths. Since shuffling also changes the tangential updates induced by the transported content, this intervention supports the importance of complete content–path correspondence rather than any isolated geometric component.

4.4 Accuracy versus Computation and Cache Budgets

Figure 3 relates full-set InternVL3 accuracy from Table 1 to estimated request FLOPs and logical prefill KV-cache size, characterizing the accuracy–resource trade-offs of different compression methods across token budgets.

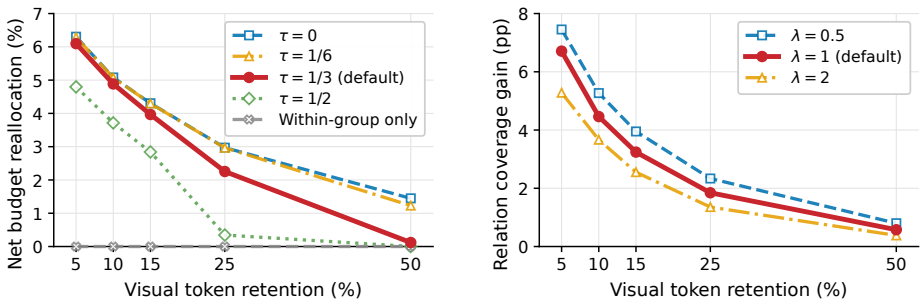

Figure 4. Hyperparameters and representative construction on InternVL3. Left: net cross-group reallocation (%) under different capacity thresholds, with $\lambda = 1$. Right: coverage gain (pp) under different decision weights, with $\tau = 1/3$. Both horizontal axes show token retention; curves denote hyperparameter settings, with defaults in red. Results use 1,600 shared inputs and equal benchmark weights.

Improved accuracy retention enables more aggressive compression at comparable quality. SRET at 15% retention exceeds GridPrune at 25% by 0.879 pp while using 7.16% fewer estimated request FLOPs. This comparison shows that stronger accuracy retention can support a tighter token budget and thereby offset part of the additional compression cost. It does not imply that SRET has lower FLOPs than other methods at every identical retention ratio.

SRET achieves higher accuracy for a given cache budget. At identical retention ratios, logical prefill KV-cache sizes are equal across methods, while SRET attains the highest accuracy at all three plotted budgets. Moreover, SRET at 10% retention is only 0.071 pp below VisionZip at 25%, while using 50.11% less logical KV cache. Thus, SRET can reach comparable accuracy with substantially fewer retained visual positions, and its cache advantage arises from preserving more useful evidence per retained position rather than from a different cache representation.

4.5 Hyperparameters and Representative Construction

On InternVL3's fixed 1,600-input panel, we separately vary the capacity threshold τ for cross-group exchange and the decision weight λ in the coverage objective. Figure 4 reports net cross-group slot transfer and coverage gain excluding the decision term, characterizing budget reallocation and representative repair. Defaults are $\tau = 1/3$ and $\lambda = 1$.

Higher capacity thresholds generally suppress cross-group reallocation, while the default induces larger changes under tight budgets. Smaller decision weights yield greater coverage gains at all five budgets but retain less initial query attention. Thus, τ controls the extent of budget redistribution, whereas λ governs the coverage–decision trade-off. Task accuracy is comparatively less sensitive, indicating that larger coverage gains do not necessarily translate into higher accuracy. Appendix B.4 provides the full settings, metric definitions, and accuracy curves.

4.6 Discussion and Limitations

SRET introduces an inference-time structural prior for preserving spatial evidence under visual token compression. Its relational graph combines visual semantics, shallow context, and local layout without requiring explicit 3D reconstruction or physical scene modeling. These lightweight cues provide effective proxies for relational substitutability across a broad range of spatial reasoning tasks, while performance may still vary with scene and observation characteristics.

5 Conclusion

SRET is a visual token reduction method inspired by geometric constraints on spatial observations. Its central idea is that compression should jointly select representatives and determine where removed evidence is reassigned under spatial structure constraints. Relations guide anchor repair, budget allocation, and evidence transport, with scale-preserving updates and frozen maps carrying information forward. Evaluations on two models, eight benchmarks, and five budgets show improved spatial reasoning accuracy at low token budgets, while FLOPs and KV-cache analyses characterize accuracy–resource trade-offs.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, including language editing, paragraph organization, and consistency checks. We did not use generative AI tools to generate experimental data, replace the experimental backbones, or provide independent scientific evidence for the claims in this paper. All experimental values, algorithm descriptions, mathematical derivations, and references were checked and verified by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text and claims produced with the aid of generative AI.

ETHICS STATEMENT

This work does not raise any specific ethical concerns.

REPRODUCIBILITY STATEMENT

The main code is provided in the supplementary material.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4895–4901, 2023.

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

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

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

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Nhat-Tan Bui, Varshini Elangovan, Arun Reddy Anugu, Sreyas Mohan, Wei Ye, Dilin Wang, JQ Huang, Rakesh Ranjan, Aviral Chharia, and Fernando De la Torre. Cover: Coverage-based token pruning for multi-view 3d reasoning in vlms. *arXiv preprint arXiv:2609.08345*, 2026.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498. IEEE, 2025.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7879–7890, 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024a.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024c.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024.

Yuxiang Duan, Ao Li, Yingqin Li, Luyu Li, and Pengwei Wang. Gridprune: From "where to look" to "what to select" in visual token pruning for mllms. *Neurocomputing*, pp. 133887, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong Yu, Laizhong Cui, and Qi Tian. Nüwa: Mending the spatial integrity torn by vlm token pruning. *arXiv preprint arXiv:2602.02951*, 2026.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Yuanyuan Jia, Shunpu Tang, and Qianqian Yang. S^2 prune: Spatially structured visual token pruning for multimodal large language models. *arXiv preprint arXiv:2609.01224*, 2026.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. In *European Conference on Computer Vision*, pp. 95–111. Springer, 2026.

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

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133(10):6794–6812, 2025.

Peng Ling, Yingda Yin, Lingting Zhu, Weikai Chen, Shengju Qian, Zeyu Hu, Xin Wang, and Wenming Yang. Coverprune: Coverage-driven token pruning for 3d vlms via optimal transport. In *European Conference on Computer Vision*, pp. 499–516. Springer, 2026.

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

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934. IEEE, 2025.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. Site: towards spatial intelligence thorough evaluation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9058–9069. IEEE, 2025.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, volume 2025, pp. 53285–53310, 2025.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X Chang. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Full Accuracy Results and Evaluation Details

A.1 All Methods at Very Low Retention

Table 1 gives full-set macro scores for two models and five budgets. Tables A1a and A1b give benchmark-level results at 5% retention, covering Full, ten baselines, and SRET in the same order. SITE averages image and video subsets before Avg equally averages the eight benchmarks. Random computes each seed's official score before averaging seeds.

Table A1a. Per-benchmark scores (%) for all methods at 5% retention on Qwen3-VL. Full is uncompressed; MindCube denotes MindCube Tiny; ReVSI uses 16 frames. SITE averages image and video scores; Avg equally averages eight benchmarks. Bold/underlined: best/second-best compressed scores per column.

Method	MMSI-Bench	MindCube	BLINK	Emb-Spatial	View-Spatial	3DSR-Bench	SITE	ReVSI	Avg
Full	38.700	68.286	64.755	83.187	56.250	57.282	53.905	58.210	60.072
Random	27.000	<u>51.206</u>	44.292	40.595	47.216	**51.666**	30.042	<u>46.170</u>	42.274
Uniform Spatial	<u>29.900</u>	49.333	42.031	<u>60.357</u>	48.757	<u>51.549</u>	25.009	42.208	43.643
Vision Attention	28.800	49.619	43.661	53.242	48.267	50.950	26.560	41.296	42.799
Query Relevance	26.800	45.238	39.821	44.176	45.413	48.939	16.602	36.902	37.987
Feature Diversity	27.800	50.286	43.346	58.077	<u>49.947</u>	51.532	<u>30.744</u>	44.000	<u>44.467</u>
Query Spatial	27.700	45.810	40.084	50.330	46.709	48.845	19.159	38.576	39.652
FastV	29.400	46.571	46.554	55.907	47.269	48.639	26.467	39.654	42.558
DivPrune	27.600	51.048	43.819	51.016	48.792	51.164	28.038	**47.112**	43.574
GridPrune	28.500	47.238	42.767	51.566	47.532	50.924	22.998	38.796	41.290
VisionZip	29.700	50.476	<u>47.238</u>	52.088	49.177	50.873	26.835	40.413	43.350
SRET	**30.100**	**56.667**	**51.289**	**64.066**	**50.158**	50.890	**32.169**	41.064	**47.050**

Table A1b. Per-benchmark scores (%) for all methods at 5% retention on InternVL3. Full is uncompressed; MindCube denotes MindCube Tiny; ReVSI uses 16 frames. SITE averages image and video scores; Avg equally averages eight benchmarks. Bold/underlined: best/second-best compressed scores per column.

Method	MMSI-Bench	MindCube	BLINK	Emb-Spatial	View-Spatial	3DSR-Bench	SITE	ReVSI	Avg
Full	42.000	90.381	48.343	80.962	60.172	62.399	50.268	51.438	60.745
Random	33.333	71.397	40.540	56.658	51.161	52.795	30.468	40.792	47.143
Uniform Spatial	33.100	70.190	41.136	56.484	51.208	51.857	30.701	40.464	46.892
Vision Attention	31.300	60.667	37.927	47.033	46.534	50.847	24.077	39.357	42.218
Query Relevance	29.200	54.571	38.769	44.725	43.890	47.005	17.502	35.841	38.938
Feature Diversity	33.800	70.190	38.401	56.566	52.293	52.379	31.972	40.339	46.993
Query Spatial	30.500	58.571	38.664	46.511	46.078	49.529	20.689	36.902	40.931
FastV	<u>37.400</u>	73.048	<u>42.241</u>	60.989	<u>55.147</u>	<u>56.127</u>	<u>33.642</u>	<u>41.670</u>	<u>50.033</u>
DivPrune	34.700	<u>74.571</u>	38.453	<u>61.209</u>	52.749	54.629	31.638	**42.525**	48.809
GridPrune	32.000	65.810	41.557	50.055	48.547	51.249	25.379	39.595	44.274
VisionZip	28.800	62.762	39.926	46.703	47.234	50.813	23.355	39.487	42.385
SRET	**38.800**	**75.714**	**42.504**	**63.929**	**56.057**	**56.931**	**35.790**	41.401	**51.391**

Advantages and differences across benchmarks. SRET achieves Avg scores of 47.050% and 51.391% on Qwen3-VL and InternVL3, above all compression baselines. It leads six of eight benchmarks on Qwen3-VL and seven on InternVL3, excluding Full. Relative to each benchmark's strongest baseline, Qwen3-VL gains 5.460, 4.050, and 3.709 pp on MindCube, BLINK, and Emb-Spatial; InternVL3 gains 2.720 and 2.149 pp on EmbSpatial and SITE. Receiving relational evidence thus helps preserve accuracy across tasks when representatives are scarce.

The advantage is not universal. On Qwen3-VL, SRET trails the best compression baseline by 0.776 and 6.049 pp on 3DSRBench and ReVSI; on InternVL3, it trails by 1.124 pp on ReVSI.

The strongest methods are Random and DivPrune for the former two, and DivPrune for the latter. Local layouts and shallow features therefore provide proxies of differing effectiveness across scenes; average gains do not imply improvement on every video spatial question.

A.2 Valid Samples and Scoring Consistency

We intersect valid samples across all five budgets, all methods, and Random's three seeds to avoid missing-input confounds. Shared invalid-video exclusions remove 23 question–answer pairs from SITE video and 16 from ReVSI; Table A2 lists retained counts. SITE subset and benchmark weights do not depend on sample counts. The module ablations and fixed-anchor interventions use seed 20260921 to fix 200 inputs per benchmark, with 100 per SITE subset, totaling 1,600 inputs. InternVL3 hyperparameter experiments reuse this panel.

We follow official scoring and internal aggregation. SITE computes chance-adjusted accuracy (CAA) from each subset's chance-adjusted numerator and denominator, then averages the two scores. ReVSI follows official question-type aggregation. Random computes metrics independently for each seed and averages them. Summaries use saved per-example scores rather than repeatedly parsing answers.

Let $\mathcal{B}$ contain the eight benchmarks, with SITE image and video scores S_{SI} and S_{SV}:

$$S_{\mathrm{SITE}} = \frac{S_{\mathrm{SI}} + S_{\mathrm{SV}}}{2}, \qquad S_{\mathrm{macro}} = \frac{1}{8}\left(\sum_{b \in \mathcal{B} \setminus \{\mathrm{SITE}\}} S_b + S_{\mathrm{SITE}} \right). \tag{11}$$

Average accuracy in the main text denotes this macro average of official metrics, not pooled question-level accuracy. Scores are percentages and differences are pp; Avg weights two models and five budgets equally. Except for the explicitly supplementary parsing in Appendix B.7, all primary results and mechanism analyses use original strict answer parsing. The BLINK format-sensitivity analysis does not replace them.

A.3 Implementation of Diagnostic Baselines

Random uses seeds 0, 1, and 2, scoring each before averaging; other methods fix seed 0. The six diagnostic baselines in Table 1 isolate selection criteria without additional training. They select after visual encoding and projection, before the language model; retained positions preserve sequence order, and unselected representations are deleted without aggregation. Query Relevance and Query Spatial use input embeddings rather than the shallow language attention used by FastV or SRET.

Method	Signals and selection rule	Budget
Random	Sample tokens uniformly without replacement within each layout group. Evaluate seeds 0, 1, and 2 separately, then average metrics.	Groupwise
Uniform Spatial	Greedy farthest-point sampling using layout coordinates only. Start at the point farthest from the group's coordinate mean, then maximize distance to the nearest selected point. Uniformity refers to coverage, not fixed-stride grid sampling.	Groupwise
Vision Attention	Take Top-K by aligned final-layer visual attention importance supplied by the model adapter. No question text is used.	Global
Query Relevance	Average input embeddings over valid non-visual, non-special text positions; select Top-K by cosine similarity to projected visual tokens. The mean covers the input prompt, not a separate text encoder's output.	Global
Feature Diversity	L2-normalize projected visual features and perform greedy farthest-point sampling with Euclidean distance, starting from the token farthest from the group feature mean.	Groupwise
Query Spatial	Min–max normalize Query Relevance scores within each group. Start with the highest-relevance token, then greedily maximize $0.7 \times$ relevance $+ 0.3 \times$ spatial novelty. Novelty is distance to the nearest selected coordinate, min–max normalized each round; selected entries are excluded.	Groupwise

Table A2. Common valid sample counts for full-set comparisons.

Benchmark	Qwen3-VL	InternVL3
MMSI-Bench	1000	1000
MindCube Tiny	1050	1050
BLINK	1901	1901
EmbSpatial	3640	3640
ViewSpatial	5712	5712
3DSRBench	11686	11686
SITE	8045 (4449 image / 3596 video)	8045 (4449 image / 3596 video)
ReVSI	4552	4552

Groupwise methods follow input image, tile, or frame layouts, allocating integer budgets in proportion to group token counts and retaining at least one token per group when the total budget permits. Rounding remainders are redistributed to satisfy the exact total. If the budget is smaller than the number of groups, the first K groups in layout order receive one token each and the rest receive zero. Global methods rank over the full visual sequence without minimum group quotas.

Vision Attention uses backbone-adapted visual scores: InternVL3 uses final-layer CLS-to-patch attention, whereas Qwen3-VL averages received final-layer visual attention and aligns it to merged visual tokens. Both are visual-side scores, not question-conditioned language attention. Feature Diversity is not identical to DivPrune: the former is groupwise and starts from the point farthest from the group mean; the latter is global by default and initializes max–min diversity search from the point with the largest nearest-neighbor distance.

A.4 Benchmarks and Spatial Abilities

BLINK covers fourteen fine-grained visual perception tasks (Fu et al., 2024); 3DSRBench focuses on 3D spatial relations (Ma et al., 2025); EmbSpatial-Bench examines egocentric understanding for embodied tasks (Du et al., 2024); and ViewSpatial-Bench distinguishes camera and person reference frames (Li et al., 2026). They assess local references, directional judgments, and related foundational perception.

MMSI-Bench evaluates multi-image spatial intelligence (Yang et al., 2026); MindCube studies limited-view reasoning through cognitive maps, perspective transformations, and motion imagination, and we use its Tiny evaluation set (Yin et al., 2025). SITE includes images and videos (Wang et al., 2025); ReVSI emphasizes visibility and answerability under sampled-frame video spatial QA, for which we fix 16 frames (Zhang et al., 2026). The algorithm does not use task-category labels, and all benchmarks share compression rules and hyperparameters.

A.5 Core Ablation and Transport Intervention Protocols

Tables 2 and 3 use the same 1,600-input panel: 200 per benchmark, with 100 per SITE subset. Compression layers, budgets, inputs, prompts, and decoding configurations are shared. Aggregation uses deterministic FP32 summation to avoid paired-comparison variability from atomic accumulation.

A.5.1 Module Ablations

Relational coverage selection comprises relation modeling, anchor exchange, and cross-group budget adjustment. Structure-preserving transport comprises source-mass weighting, endpoint allocation, scale-preserving writing, and frozen incremental mapping where applicable.

From query pruning to relational selection and transport. Setting (1) retains the query Top-K anchors A_0 without writing removed evidence. Setting (2) repairs anchors using the complete coverage objective but only slices retained positions. Setting (3) keeps A_0 unchanged and computes full transport on that set. Setting (4) enables both modules. Settings (1)/(3) and (2)/(4) have identical anchors. For later Qwen3-VL DeepStack injections, no-transport conditions also slice anchors, while transport conditions reuse their initial frozen maps.

Explicit layout constraints. Setting (5) removes local coordinates and group constraints from both selection and transport: all sources form one semantic relation group, joint affinity takes the minimum of visual and contextual compatibility, residual endpoints match features only, and distinct endpoints split residual mass equally. Source weights, the decision-retention term, writing, and frozen maps remain. Native positional encoding and implicit spatial information in features are unchanged, so this tests explicit layout constraints rather than all spatial information.

A.5.2 Transport Correspondence with Fixed Anchors

We fix full SRET's anchors A, source mass m_i, and budget, comparing normal transport, no transport, and within-group shuffling. Shuffling uses three seeds; Table 3 reports their mean and sample standard deviation. Normal transport follows Section 3.3; no transport keeps H_A without writing removed evidence.

Shuffling also fixes normal transport weights W_{ia} and permutes same-group non-anchor sources using π, leaving anchors unchanged:

$$\bar{h}_a^{\pi} = \frac{\sum_i W_{ia} h_{\pi(i)}}{\mu_a}, \qquad \pi(a) = a \ (a \in A), \qquad g_{\pi(i)} = g_i. \tag{12}$$

The source feature multiset, slots, weights, and received mass are preserved, but content–path correspondence is disrupted. Permutations are random cycles over non-anchors in each group; groups with fewer than two movable sources remain unchanged. Later Qwen3-VL increments use the frozen map normally, the same permutation when shuffled, and anchor slicing without transport. All aggregation conditions use deterministic FP32 summation.

B Mathematical Properties and Supplementary Experiments

B.1 Properties of Relational Coverage and Exchanges

Query and context. At the third decoder block's input, native query–key computation on normalized inputs provides head-averaged attention from the last text position to visual positions, yielding d_i. Normalized visual-position inputs give z_i^c. Query, contextual, and visual evidence are aligned to the language model's visual token granularity; $\bar{d}$ and $\bar{v}$ are separately normalized over all visual positions in the current input.

Visual evidence. Features z_i^v come from final-layer visual attention keys, not projector outputs or language hidden states. InternVL3 uses native patch keys, including enabled key normalization, concatenates heads, and averages over spatial downsampling units. Its importance v_i is head-averaged CLS-to-patch attention with the same pooling. Qwen3-VL uses keys after visual rotary position encoding, concatenates heads, and averages within spatial merge units. Importance is the attention received by each visual token within its visual attention sequence, averaged over query positions and heads and then spatially pooled. This single-token importance differs from pairwise affinity V_{ij}.

Layout groups and coordinates. Each dynamic InternVL3 tile forms a group, with a separate thumbnail group. Under the current input protocol, each Qwen3-VL image or sampled frame is a group. One image need not be one group, and groups are not semantic objects. Given preprocessing coordinates p_i and groups g_i, compute, per group and dimension,

$$\begin{aligned} l_{g,d} &= \min_{j:g_j=g} p_{j,d}, \qquad u_{g,d} = \max_{j:g_j=g} p_{j,d}, \\ \widetilde{p}_{i,d} &= \frac{p_{i,d} - l_{g_i,d}}{\max(1, u_{g_i,d} - l_{g_i,d})}. \end{aligned} \tag{13}$$

The transform removes offsets without stretching spans below one to the full $[0, 1]$ interval. InternVL3 uses tile patch centers mapped to original-image coordinates, preserving relative tile scale; thumbnails use full-image coordinates. Qwen3-VL uses unit-interval coordinates on its merged grid, with a constant temporal component for images and individually supplied frames. Both use three coordinates: the first becomes zero after within-group translation, and the final two support endpoint matching. No 3D scene is reconstructed.

Relational affinities. For $\tilde{p}_i \in [0,1]^D$ with $D = 3$, we define the visual, contextual, and layout affinities as

$$V_{ij} = \frac{1 + \cos(z_i^v, z_j^v)}{2}, \qquad C_{ij} = \frac{1 + \cos(z_i^c, z_j^c)}{2}, \\ L_{ij} = \left[1 - \frac{\|\widetilde{p}_i - \widetilde{p}_j\|_2}{\sqrt{D}}\right]_+ . \tag{14}$$

Cross-group affinity is zero; within-group affinity takes their minimum, with self-affinity set to one numerically. Visual and contextual features are L2-normalized for cosine similarity. Contextual features obey backbone causal attention and do not additionally access text after their visual positions. Coordinates support within-group layout comparison.

Coverage and exchanges. For a fixed nonnegative relation graph, $f_i(A) = \max_{a \in A} \phi_{ia}$ has diminishing marginal returns. If $A \subseteq B$,

$$f_i(A \cup \{j\}) - f_i(A) = [\phi_{ij} - f_i(A)]_+ \geq [\phi_{ij} - f_i(B)]_+. \tag{15}$$

Nonnegatively weighted coverage is monotone submodular, and the decision term is modular, so their sum remains monotone submodular (Nemhauser et al., 1978). SRET uses one-for-one local search under a fixed budget. Within-group exchanges preserve capacities; cross-group exchanges reallocate them without changing token count. Group-disjoint actions affect independent evidence blocks, so combining positive-gain actions preserves budget and improves the objective. Termination gives a local solution in the allowed exchange neighborhood.

Each round computes the best local exchange per group and, if enabled, cross-group proposals using donor removal costs and receiver addition gains. Proposals are jointly sorted by gain, choosing actions with disjoint group endpoints and gains above numerical tolerance. All unselected tokens are candidates; simultaneous disjoint gains add. Within-group repair and cross-group adjustment are not separate sequential optimizations.

Writing $\pi_g^* = q_g^*/K$ and $\pi_g^0 = q_g^0/K$, the gate satisfies

$$\frac{\Delta_{\text{cap}}}{2} = \text{TV}(\pi^*, \pi^0). \tag{16}$$

The threshold $1/3$ therefore corresponds to total variation distance exceeding $1/6$.

Blockwise execution and computational scale. Figure 2 shows only some edges; computation covers all within-group pairs. For group sizes N_g, visual/contextual dimensions d_v, d_c, and coordinate dimension D, affinity construction scales as $O(\sum_g N_g^2(d_v + d_c + D))$ without a full cross-group matrix. Optimization reuses relation blocks and indices, fusing candidate gains and local exchanges. With at least eight groups, affinity matrix multiplication uses FP16; evidence weights and objectives accumulate in FP32. Once endpoints are known, sparse aggregation costs $O(Nd)$ for language dimension d; endpoint search and iterative exchanges are additional. Position encodings, masks, and existing KV states are aligned during compression.

Resource interpretation. SRET's cache advantage at comparable accuracy comes from using fewer retained positions, not from reducing KV-head counts (Ainslie et al., 2023) or allocation fragmentation (Kwon et al., 2023). The reported logical prefill cache size excludes temporary compression tensors and allocator overhead; it is not a peak-memory measurement.

B.2 Evidence Conservation and Local Spatial Moments

Contextual transport features and endpoints. The transport compatibility C^{tr} uses native keys shared with query evidence:

$$z_i^k = \text{Normalize}\Big(\text{Concat}_{h=1}^{H_{\text{KV}}} \text{Normalize}(k_i^{(h)})\Big). \tag{17}$$

Here H_{KV} is the number of KV heads, and $k_i^{(h)}$ includes native input normalization, key projection, model-specific key normalization, and position encoding. We define $C_{ia}^{\text{tr}} = (1 + \cos(z_i^k, z_a^k))/2$, distinguishing receiver-space matching from the hidden-state compatibility C_{ij} used in relation modeling. Visual and contextual endpoints maximize an equally weighted mean of feature and local 2D

layout compatibility over same-group anchors, falling back to global feature matching when none exist. This differs from the joint representative's minimum intersection of visual, contextual, and layout compatibility.

Local projection. For endpoints at distinct positions in the source's group,

$$t_i = \mathrm{clip}_{[0,1]}\left(\frac{\langle \widetilde{p}_i - \widetilde{p}_{b_i^v}, \widetilde{p}_{b_i^c} - \widetilde{p}_{b_i^v}\rangle}{\|\widetilde{p}_{b_i^c} - \widetilde{p}_{b_i^v}\|_2^2}\right). \tag{18}$$

Distinct endpoints without a valid within-group segment use $t_i = 1/2$; identical endpoints coalesce. Anchors use self-loops with effective $c_i = 1$, preventing redistribution of their original content.

Evidence conservation. Since $\sum_a P_{ia} = 1$, total evidence mass is preserved:

$$\sum_a \mu_a = \sum_a \sum_i m_i P_{ia} = \sum_i m_i = M. \tag{19}$$

For a valid segment, $(1-t_i)\widetilde{p}_{b_i^v} + t_i \widetilde{p}_{b_i^c}$ is the closest segment point to the source. Barycentric allocation preserves this projected position of the relational component, not general 3D spatial moments; source-to-projection distance measures endpoint position error.

B.3 Tangential Updates and Incremental Maps

The feature-space tangential update r_a differs from the mass-based relational residual: the former controls the writing direction, while the latter allocates source content among receivers. Initial norm restoration uses

$$\eta_a = \frac{\|h_a\|_2}{\|h_a + s_a r_a\|_2}. \tag{20}$$

Proposition 1 (Scale and direction constraints of initial transport). For a nonzero anchor in real arithmetic, initial transport preserves $\|\widetilde{h}_a\|_2 = \|h_a\|_2$ and rotates the direction by less than $45°$.

Proof. Let $A_a = \|h_a\|_2 > 0$ and $R_a = \|r_a\|_2$. Since $r_a \perp h_a$,

$$\tan\theta_a = \frac{s_a R_a}{A_a} = \frac{R_a}{A_a + R_a} < 1. \tag{21}$$

Norm restoration η_a scales without changing direction and restores A_a, proving the claim.

Proposition 2 (Linear consistency in a frozen frame). Fix the current input's initial transport parameters W, μ, u, s, η. The subsequent map is linear in its increment and consistent with initial transport:

$$\overline{D}_a = \frac{\sum_i W_{ia} D_i}{\mu_a}, \qquad \mathcal{L}_H(D)_a = \eta_a\big[D_a + s_a T_a(\overline{D}_a - D_a)\big], \tag{22}$$

$$\begin{aligned} \mathcal{L}_H(\alpha D_1 + \beta D_2) &= \alpha \mathcal{L}_H(D_1) + \beta \mathcal{L}_H(D_2), \\ \mathcal{L}_H(H) &= \widetilde{H}_A. \end{aligned} \tag{23}$$

Proof. With parameters frozen, aggregation, anchor extraction, tangent projection, and scaling are linear, and so is their composition. Substituting $D = H$ recovers initial writing in Section 3.3.2. For zero received mass, aggregation uses the anchor input as fallback, remaining linear. This applies to a fixed map from one compression event, not the full compressor under reselection or parameter re-estimation. These properties concern ideal real arithmetic, while implementation uses finite precision.

In Qwen3-VL, the map receives DeepStack injections after compression, replacing only the increment D without re-estimating endpoints, weights, tangent projections, or scales. InternVL3 has no such injections and performs initial transport only. Frozen state belongs to the current input's prefill and is not reused across requests.

Numerical implementation. Weights, received masses, and weighted representation sums accumulate in FP32 and are cast back to model dtype for writing. Denominators in received mass, normalization, tangent scaling, and norm restoration are lower-bounded by 10^{-12}. Zero received mass uses the original anchor as mean; zero-norm anchors remain zero after initial writing. Frozen maps reuse initial parameters without re-estimating update directions or scales.

B.4 Hyperparameter Settings, Mechanism Metrics, and Task Accuracy

We examine how hyperparameters change compact representative construction through budget reallocation and coverage gains. InternVL3 uses the same panel as Table 3: 200 inputs per benchmark, with retention 5%, 10%, 15%, 25%, and 50%. Only one parameter changes at a time; others remain at shared defaults without benchmark- or budget-specific tuning. Mechanism metrics are computed per input, then averaged equally over benchmarks, with SITE subsets averaged first.

Capacity threshold and budget reallocation. Section 3.2 permits cross-group exchanges when $\Delta_{\mathrm{cap}} > \tau$. With the objective weight fixed, we compare $\tau \in \{0, 1/6, 1/3, 1/2\}$ and fully prohibited cross-group exchange; the default is $1/3$. If q_g^0 and q_g are initial and final group anchor counts, net reallocation is

$$R_{\mathrm{alloc}} = \frac{1}{2K} \sum_g |q_g - q_g^0|. \tag{24}$$

This measures the final fraction of retained slots transferred between groups, not cumulative exchanges. Figure 4 (left) reports percentages. Prohibiting cross-group exchange gives zero reallocation while retaining within-group repair and full transport.

Balancing coverage and decision support. We generalize the objective to

$$F_\lambda(A) = \frac{1}{M} \sum_i m_i c_i(A) + \lambda \frac{\sum_{a \in A} \bar{d}_a}{\sum_{a \in A_0} \bar{d}_a}, \qquad |A| = K. \tag{25}$$

Fixing $\tau = 1/3$, we compare $\lambda \in \{0.5, 1, 2\}$; $\lambda = 1$ recovers SRET. Only the repair objective changes, not initial anchors, source masses, relations, or transport. To avoid comparing differently weighted objectives, Figure 4 (right) reports coverage gain without the decision term:

$$C(A) = \frac{1}{M} \sum_i m_i c_i(A), \qquad \Delta C = C(A) - C(A_0). \tag{26}$$

The plot scales ΔC by 100 to report pp. This internal substitutability proxy is not the accuracy of real spatial relations.

Both studies use the primary evaluation's aggregation implementation. Existing full-set records for the default and no-cross-group settings are filtered by panel IDs; other settings are evaluated on the same panel. Mechanism metrics come from corresponding records. Figure B1 reports task accuracy for comparison with Figure 4.

Average accuracy is relatively stable across capacity thresholds. The five-budget mean scores of the four finite thresholds span 0.091 pp. The default achieves the highest average but is not optimal at every budget. Prohibiting cross-group exchanges exceeds the default at 25% but trails it at the other four budgets, indicating budget-dependent value. Together with Figure 4 (left), substantially different reallocation magnitudes can yield similar accuracy.

No decision weight is consistently optimal across budgets. The three weights' five-budget mean scores span 0.118 pp, although individual budgets are more sensitive. A smaller weight slightly improves wide-budget performance but worsens the tightest budget. Combined with Figure 4 (right), larger coverage gains need not improve accuracy; query support and coverage must be considered together.

B.5 Initial Relational Coverage Distortion and Task Gains

Improving how retained positions explain sources need not improve answer accuracy. We ask whether inputs with more relational evidence missed by initial query selection benefit more from

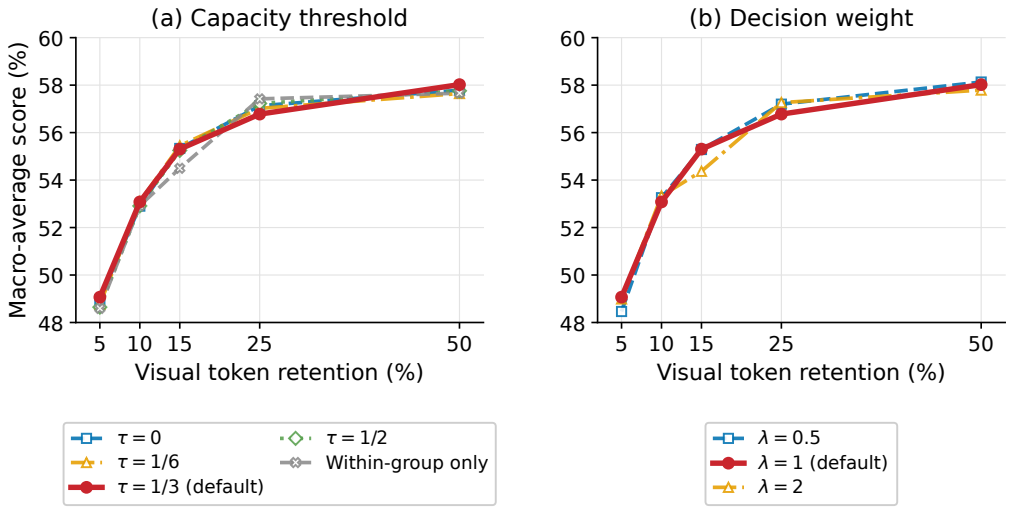

Figure B1. Task accuracy under hyperparameter changes. (a) Capacity threshold with $\lambda = 1$; (b) decision weight with $\tau = 1/3$. Horizontal axes show retention, vertical axes the eight-benchmark official macro score (%), and red the shared default.

SRET. We reuse Table 2's 1,600-input panel across both models and five budgets, comparing query-only selection, relational selection without transport, and full SRET without regenerating answers.

For initial anchors A_0, define

$$D_0 = D_{\mathrm{rel}}(A_0) = 1 - \frac{1}{M}\sum_i m_i c_i(A_0). \tag{27}$$

Within each model, budget, and benchmark, examples are split by D_0 into approximately equally sized low, medium, and high terciles. SITE subsets are split separately and equally averaged. Grouping uses pre-optimization coverage, not correctness or optimized gains. Each stratum computes per-question score gains over query-only selection, then averages equally over eight benchmarks and five budgets. These are question-level task scores: binary for classification, partial credit for numerical ReVSI questions, and raw accuracy for SITE. The diagnostic therefore differs from official macro scores; we additionally recompute official scores within strata as a sensitivity check. Confidence intervals use 2,000 within-benchmark question-level bootstrap resamples, sharing indices across methods and budgets with fixed strata. They neither cluster shared media nor form simultaneous confidence bands.

The distortion–gain association depends on the backbone. InternVL3 gains increase from 0.673 to 0.863 to 1.415 pp across low, medium, and high strata. Qwen3-VL gains are 3.049, 3.143, and 2.328 pp, so the highest-distortion group does not gain most. High-minus-low gains are $+0.743$ pp on InternVL3, with 95% interval $[-0.616, 2.089]$, and -0.721 pp on Qwen3-VL, with interval $[-2.791, 1.415]$. Both include zero and do not establish a model-independent monotonic relationship. Recomputing official metrics within strata preserves the overall trends, while individual budgets are not uniformly monotonic either.

Coverage is a compression proxy, not a per-input benefit predictor. Full SRET outperforms relational selection alone on average in all three strata of both models, but initial distortion alone does not predict transport benefits. D_0 measures a substitutability gap on a fixed graph; answers also depend on whether missing evidence matters to the question and how the backbone uses compact representations. Relational distortion for representative construction must therefore be distinguished from task scores. The analysis supports neither a universal higher-distortion/larger-gain rule nor a causal interpretation of stratification.

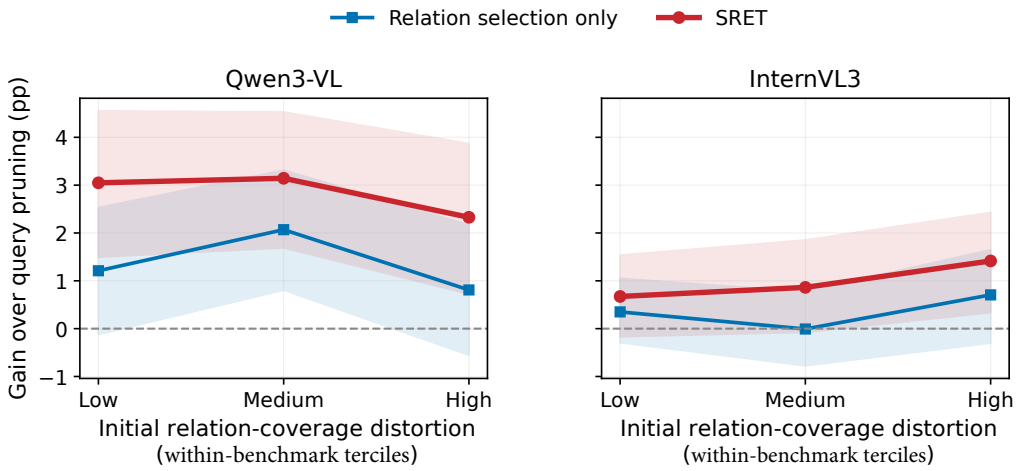

Figure B2. Initial coverage distortion and task gains. Horizontal axes show initial relation-coverage distortion; vertical axes show gains over query-only selection, averaged equally across five budgets and eight benchmarks. Shading denotes question-level 95% bootstrap intervals.

B.6 Preserving Uncompressed Correct Answers and Decomposing Gains

Average accuracy under compression combines damage to previously correct answers with improvements on previously unsuccessful ones. We pair predictions from all ten baselines and SRET on Table 1's common full-set inputs. Each model uses the same 37,586 inputs at every budget; Random's metrics are computed per seed before averaging.

Let s_j^{Full} and s_j^m be per-question scores for Full and method m. Split examples into $\mathcal{C} = \{j : s_j^{\text{Full}} = 1\}$ and $\mathcal{U} = \{j : s_j^{\text{Full}} < 1\}$. Full-credit retention and recovery are

$$R^m_{\text{keep}} = \frac{\sum_{j\in\mathcal{C}} \mathbf{1}[s_j^m = 1]}{|\mathcal{C}|}, \qquad R^m_{\text{recover}} = \frac{\sum_{j\in\mathcal{U}} \mathbf{1}[s_j^m = 1]}{|\mathcal{U}|}. \tag{28}$$

Full credit means correctness for classification. For numerical ReVSI questions with partial credit, $\mathcal{U}$ also contains partially correct answers, not only zero-score errors. Conditional rates are computed within each benchmark, averaging SITE subsets first and then all eight benchmarks. Full-credit counts for Qwen3-VL and InternVL3 are 22,374 and 22,687; plotted rates do not simply pool these inputs.

SRET better preserves the uncompressed model's correct answers. Five-budget mean retention is 79.548% on Qwen3-VL and 85.541% on InternVL3, exceeding the strongest baseline for each model by 2.882 and 1.779 pp. SRET leads nine of ten model–budget settings, except Qwen3-VL at 50%. Gains are clearer under tight budgets, matching the primary accuracy trends.

Accuracy gains primarily reduce lost correct answers rather than increase recovery. SRET's five-budget mean recovery rates are 16.024% and 12.910%, below every baseline. To connect this binary diagnostic to official scores, decompose its difference from each baseline:

$$\Delta = \Delta_{\mathcal{C}} + \Delta_{\mathcal{U}}, \qquad \Delta_{\mathcal{S}} = 100 \sum_{j\in\mathcal{S}} w_j (s_j^{\text{SRET}} - s_j^m), \quad \mathcal{S} \in \{\mathcal{C},\mathcal{U}\}. \tag{29}$$

Weights w_j preserve official question-type, SITE chance-adjustment, and benchmark weights without renormalizing within conditional subsets. The components therefore exactly sum to the official macro-score difference, including ReVSI partial-credit changes. Across the twenty model–baseline comparisons averaged over five budgets, $\Delta_{\mathcal{C}} > 0$, $\Delta_{\mathcal{U}} < 0$, and total gains are positive. For example, Qwen3-VL's $+1.492$ pp over VisionZip comprises $+1.960$ pp on Full-credit inputs and -0.468 pp elsewhere. Reduced damage to correct answers outweighs weaker improvement on the remaining inputs.

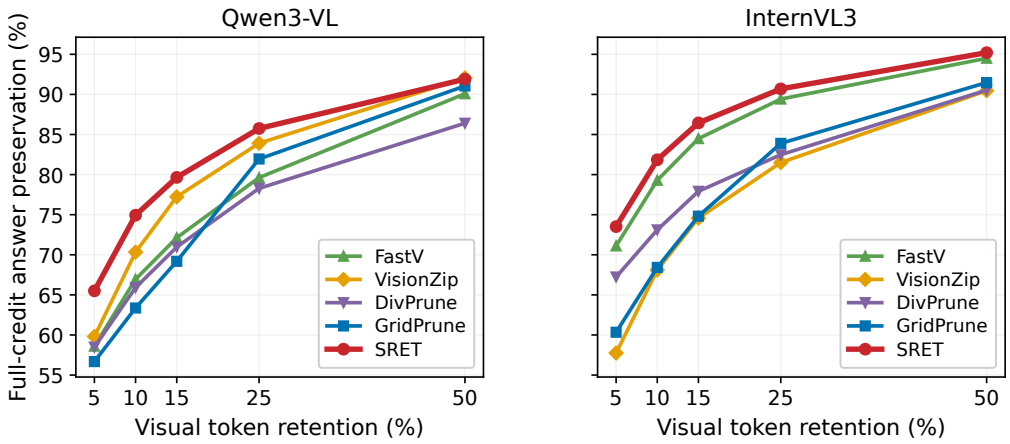

Figure B3. Retention of Full's full-credit answers. Horizontal axes show token retention; vertical axes show the eight-benchmark macro retention rate. Four literature baselines and SRET are shown. This diagnostic differs from official task scores.

These gains reflect preservation of existing abilities, consistent with transport reducing deletion losses. Full is not ground truth: improving on its unsuccessful answers can be genuine, not merely noise. This paired decomposition describes gain sources rather than establishing a module's causal role; transport correspondence is intervened on in Table 3.

B.7 BLINK Task-Wise Compression Behavior and Scoring Sensitivity

BLINK's fourteen native categories offer a fine-grained view of how visual abilities respond to compression. This section considers SRET only, comparing 50%, 25%, 15%, 10%, and 5% retention on 1,901 common inputs per model, with Full as reference. Categories contain 117–172 questions; no relabeling or additional inference is performed. Category accuracy is computed within each category; overall BLINK accuracy pools questions rather than weighting categories equally.

B.7.1 Answer Format and Scoring

An audit of saved answers found that some InternVL3 outputs express options in complete sentences, such as `The point C is correct.` The original parser requires an option letter at the beginning, marking such outputs unparsed and incorrect. This concentrates in three Correspondence categories and affects some Visual Similarity answers; Qwen3-VL does not exhibit it. Strict scores therefore reflect both content and format compliance; zero Correspondence scores do not directly imply absent correspondence recognition.

We retain strict scoring and apply a uniform format-sensitivity analysis. Existing parses are unchanged; only unparsed answers are processed with twelve whole-sentence templates for explicit affirmative choices. Templates are based on this batch's format audit, without ground truth or method identity, and do not force extraction from negation, multiple choices, or uncertain expressions. Across 132 result files and 250,932 prediction records, 29,289 were unparsed; supplementary extraction recovers 29,288, yielding 15,592 additional correct answers and leaving one unparsed. Records reuse questions across configurations; there are still 1,901 independent inputs per model.

This is a non-official sensitivity analysis of saved answers, not a replacement for the main table or earlier appendix scores. The audit covers every method to apply a consistent rule, but this section shows only SRET. Qwen3-VL scores are unchanged. InternVL3 SRET's five-budget average changes from 45.650% under strict parsing to 61.915% under supplementary parsing. Curves below use supplementary scores; Table B1 retains both where they differ, avoiding the interpretation of format-induced zero curves as compression insensitivity.

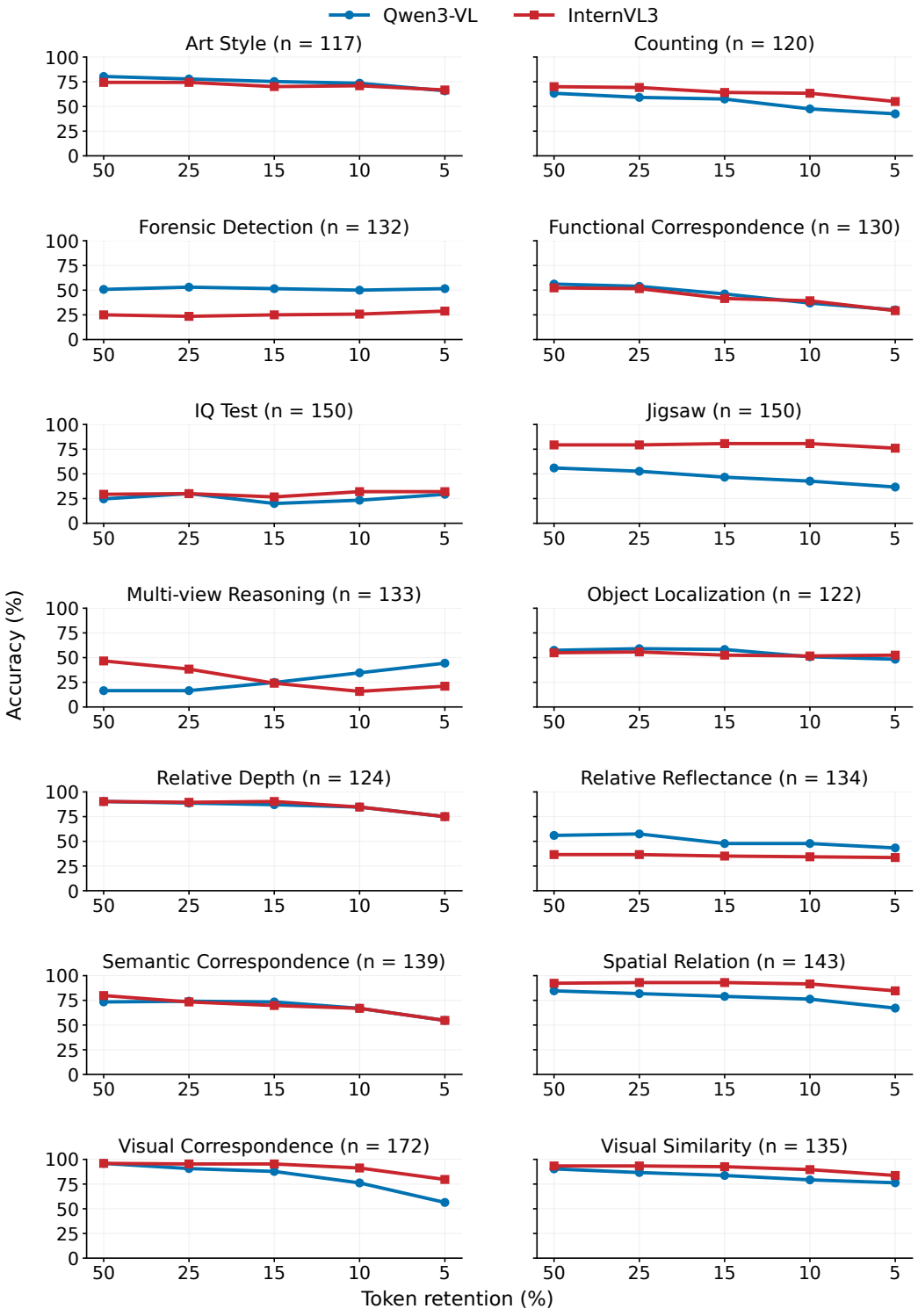

Figure B4. SRET accuracy–retention curves on fourteen BLINK tasks. Blue and red denote Qwen3-VL and InternVL3. Retention levels from 50% to 5% are equally spaced, with a shared 0–100% vertical range. Supplementary parsing is used; Qwen3-VL scores equal the original scores.

B.7.2 Task Responses to Retention Budgets

High-accuracy plateaus can coexist with low-budget degradation. InternVL3 Spatial Relation stays at 92.31%–93.01% from 50% to 15%, falling to 84.62% at 5%. Visual Correspondence stays at 95.35%–95.93% over the same range, falling to 79.65%. Some relational judgments thus withstand substantial compression before tighter budgets expose irreplaceable evidence losses. This is not universal: Semantic Correspondence falls from 79.86% to 54.68% between 50% and 5%, and Functional Correspondence from 52.31% to 29.23%. Even correspondence tasks have different budget requirements.

Compression sensitivity strongly depends on the backbone. Qwen3-VL Visual Correspondence falls from 95.93% to 56.40% between 50% and 5%, a 39.53 pp drop, versus 16.28 pp on InternVL3. InternVL3 Jigsaw changes from 79.33% to 76.00%, whereas Qwen3-VL falls from 56.00% to 36.67%. A shared compression rule does not yield uniform category-level degradation: compact representations also depend on backbone features and reasoning. Future allocation may account for these responses, but category labels do not guide this algorithm.

Non-monotonic changes do not imply stronger reasoning under compression. Qwen3-VL Multi-view Reasoning rises from 16.54% at 50% to 44.36% at 5%, but the latter remains below 50% uniform-choice chance accuracy for this binary task. It does not establish improved multi-view understanding. InternVL3 instead falls from 46.62% to 15.79% at 10% before recovering to 21.05% at 5%. Opposing responses and below-chance scores call for inspecting original questions and answers rather than attributing every trend to information loss.

Flat curves must be interpreted with absolute accuracy. InternVL3 Jigsaw remains near 80% at wider budgets, indicating relatively good preservation on this sample. IQ Test and Forensic Detection also lack clear monotonic decline, but at low absolute accuracy; low-score plateaus do not show preserved ability. With roughly a hundred questions per category, a few percentage points can reflect few examples. These are descriptive trends, not claims of significant category differences or causal mechanisms.

B.7.3 Complete Category and Budget Results

Table B1 reports all Full and SRET categories and budgets. Separate strict/supplementary tables and audit records accompanying the results retain parsing rules, source checks, and score changes.

These results reuse the same generated answers to separate format effects from budget-dependent behavior. Supplementary parsing is used only here, not for primary macro scores or Full-answer preservation in Appendix B.6.

Table B1. Full and SRET category accuracy on BLINK (%). Retention columns use SRET. Changed scores are strict → supplementary; single values agree. ALL is the sample-weighted BLINK score. Qwen3-VL is shown here; InternVL3 follows.

		Qwen3-VL					
Category	n	Full	50%	25%	15%	10%	5%
ALL	1901	64.76	64.18	63.12	59.97	56.44	51.29
Art Style	117	79.49	80.34	77.78	75.21	73.50	65.81
Counting	120	67.50	63.33	59.17	57.50	47.50	42.50
Forensic Detection	132	53.03	50.76	53.03	51.52	50.00	51.52
Functional Correspondence	130	58.46	56.15	53.85	46.15	36.92	30.00
IQ Test	150	27.33	24.67	30.00	20.00	23.33	29.33
Jigsaw	150	52.00	56.00	52.67	46.67	42.67	36.67
Multi-view Reasoning	133	9.77	16.54	16.54	24.81	34.59	44.36
Object Localization	122	59.84	57.38	59.02	58.20	50.82	48.36
Relative Depth	124	88.71	90.32	88.71	87.10	84.68	75.00
Relative Reflectance	134	58.21	55.97	57.46	47.76	47.76	43.28
Semantic Correspondence	139	75.54	73.38	74.10	73.38	66.91	54.68
Spatial Relation	143	83.92	84.62	81.82	79.02	76.22	67.13
Visual Correspondence	172	97.09	95.93	90.70	87.79	76.16	56.40
Visual Similarity	135	93.33	90.37	86.67	83.70	79.26	76.30

Table B1. Full and SRET category accuracy on BLINK (%), continued: InternVL3. Scoring conventions match the Qwen3-VL panel.

		InternVL3					
Category	n	Full	50%	25%	15%	10%	5%
ALL	1901	48.34 → 66.49	48.08 → 66.23	47.45 → 65.02	45.55 → 62.13	44.66 → 60.49	42.50 → 55.71
Art Style	117	72.65	74.36	74.36	70.09	70.94	66.67
Counting	120	71.67	70.00	69.17	64.17	63.33	55.00
Forensic Detection	132	23.48	25.00	23.48	25.00	25.76	28.79
Functional Correspondence	130	0.00 → 51.54	0.00 → 52.31	0.00 → 51.54	0.00 → 41.54	0.00 → 39.23	0.00 → 29.23
IQ Test	150	29.33	29.33	30.00	26.67	32.00	32.00
Jigsaw	150	78.00	79.33	79.33	80.67	80.67	76.00
Multi-view Reasoning	133	48.87	46.62	38.35	24.06	15.79	21.05
Object Localization	122	57.38	54.92	55.74	52.46	51.64	52.46
Relative Depth	124	91.94	90.32	89.52	90.32	84.68	75.00
Relative Reflectance	134	36.57	36.57	36.57	35.07	34.33	33.58
Semantic Correspondence	139	0.00 → 80.58	0.00 → 79.86	0.00 → 73.38	0.00 → 69.78	0.00 → 66.91	0.00 → 54.68
Spatial Relation	143	92.31	92.31	93.01	93.01	91.61	84.62
Visual Correspondence	172	0.00 → 96.51	0.00 → 95.93	0.00 → 95.35	0.00 → 95.35	0.00 → 91.28	0.00 → 79.65
Visual Similarity	135	93.33	92.59 → 93.33	92.59 → 93.33	92.59	89.63	83.70

C QUALITATIVE EXAMPLES

We present nine selected examples from EmbSpatial, MindCube Tiny, MMSI-Bench, and ViewSpatial-Bench, covering single-image spatial relations, camera motion, and reasoning across two to eight views. Within each case, FastV, VisionZip, and SRET retain the same number of visual tokens. The selection panels show retained positions and predicted answers, while the aggregation panels compare the removed-token contributions received by selected anchors in VisionZip and SRET. FastV does not aggregate removed features.

In the laptop-and-pillow example, SRET retains a token on the pillow and routes features from the laptop region to a nearby anchor. In the black-chair-and-iMac example, the displayed SRET anchors receive features concentrated around the monitor and chair, while the displayed VisionZip anchors receive features from more widely separated regions. SRET answers all nine selected questions correctly, while FastV and VisionZip give incorrect answers. These examples illustrate differences in retained positions and aggregation assignments; aggregate performance and controlled comparisons are reported in Section 4.

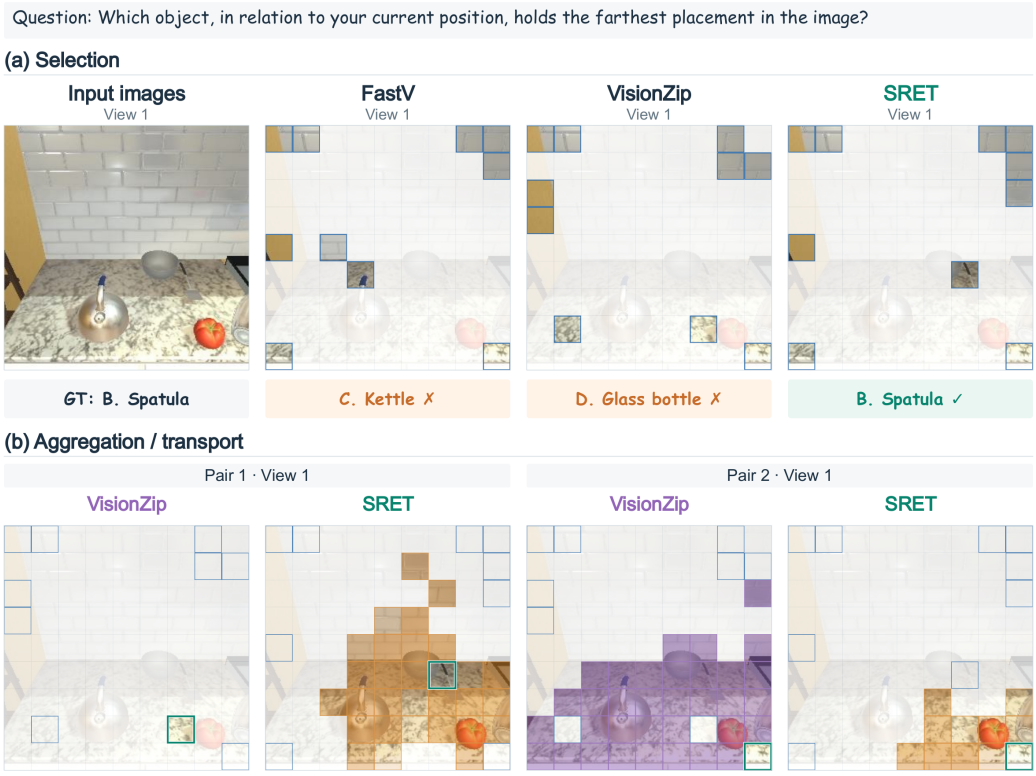

Figure C1. Single-image depth ordering. The question asks which object is farthest from the observer. Panel (a) shows the input, retained visual positions, and model predictions. Panel (b) compares contributions from removed tokens to two selected receiving anchors in VisionZip and SRET. Blue outlines indicate retained positions, and green outlines mark the displayed receiver. Purple and orange regions indicate source tokens assigned to that receiver by VisionZip and SRET, respectively; darker shading indicates a larger routing fraction before pooling. Receiver positions are selected separately for the two methods. Some displayed VisionZip receivers are dominant tokens and receive no aggregated contributions from removed tokens. The same visual conventions apply to the following examples.

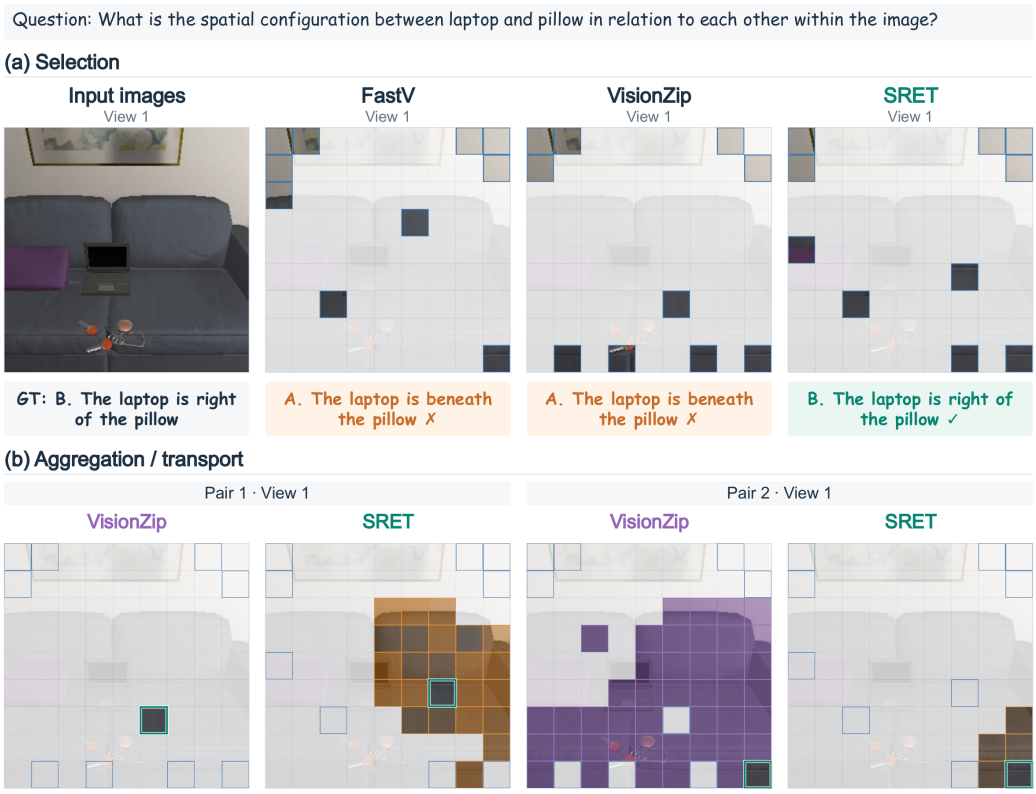

Figure C2. The spatial relation between a laptop and a pillow. The example compares retained positions and aggregation assignments for a single-image left–right judgment. Visual conventions follow Figure C1.

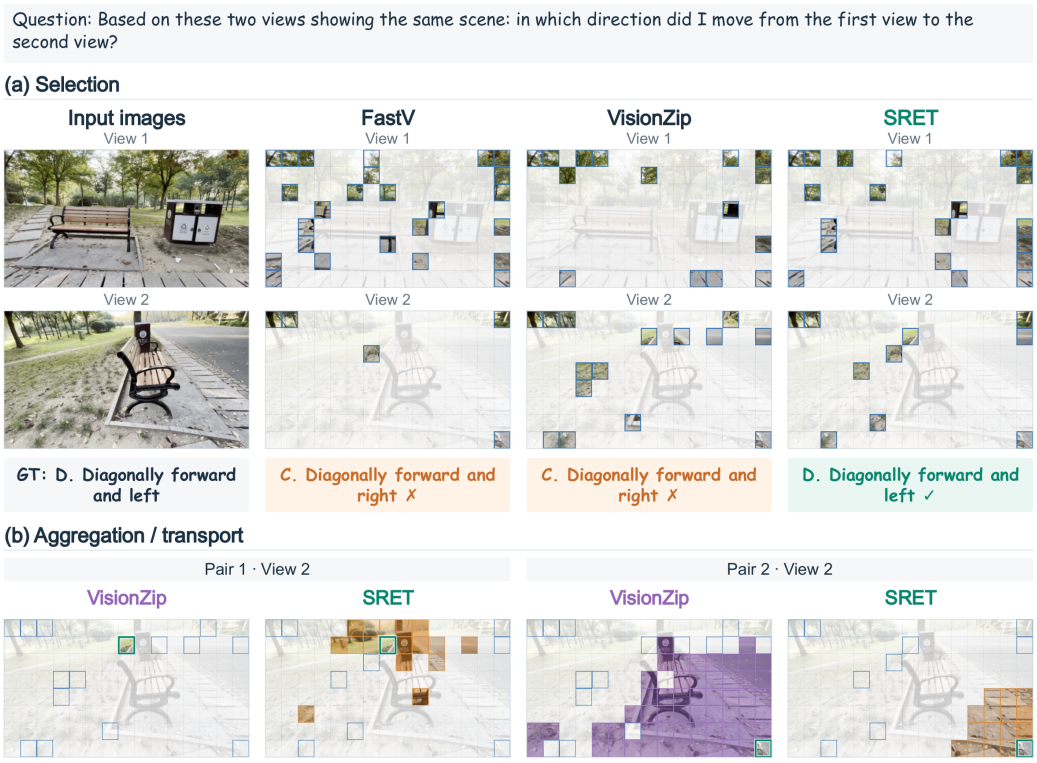

Figure C3. Camera motion across two views. The question asks how the observer moved between two views of a scene containing a bench. Visual conventions follow Figure C1.

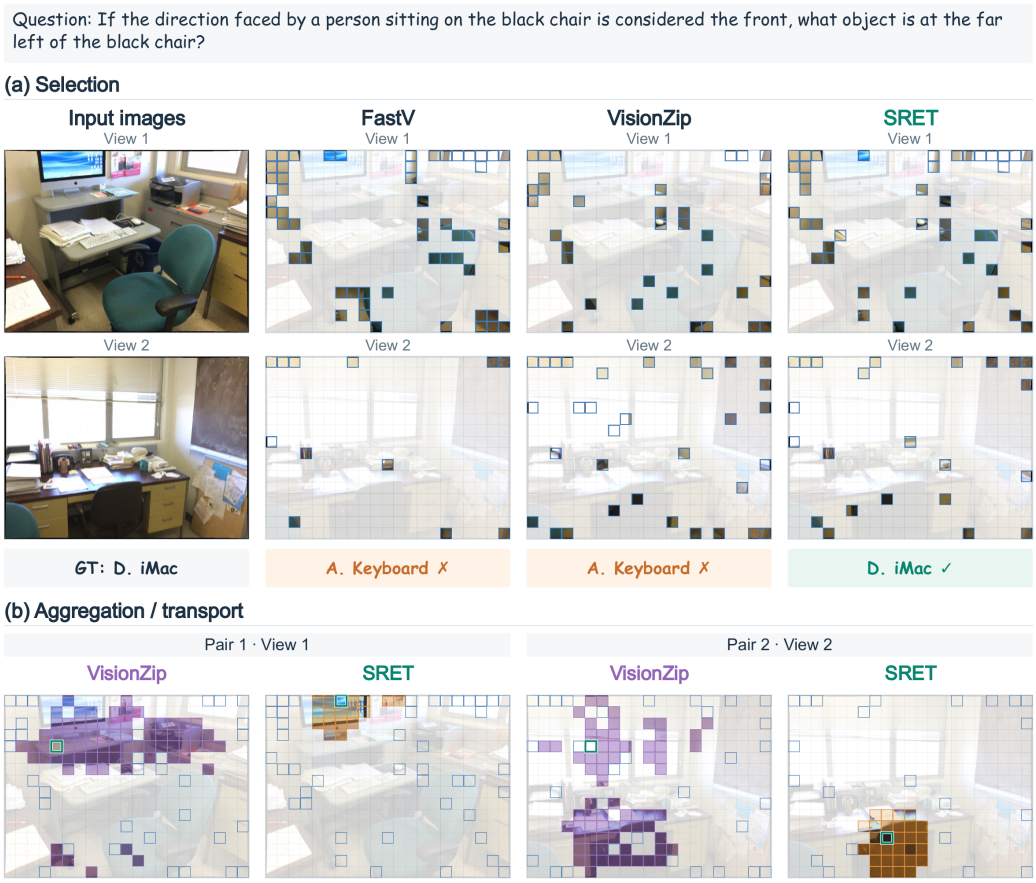

Figure C4. An object-relative reference frame across two views. The question asks which object lies at the far left of the black chair, using the facing direction of a seated person as the front. The aggregation panels show receivers associated with the monitor and chair regions. Visual conventions follow Figure C1.

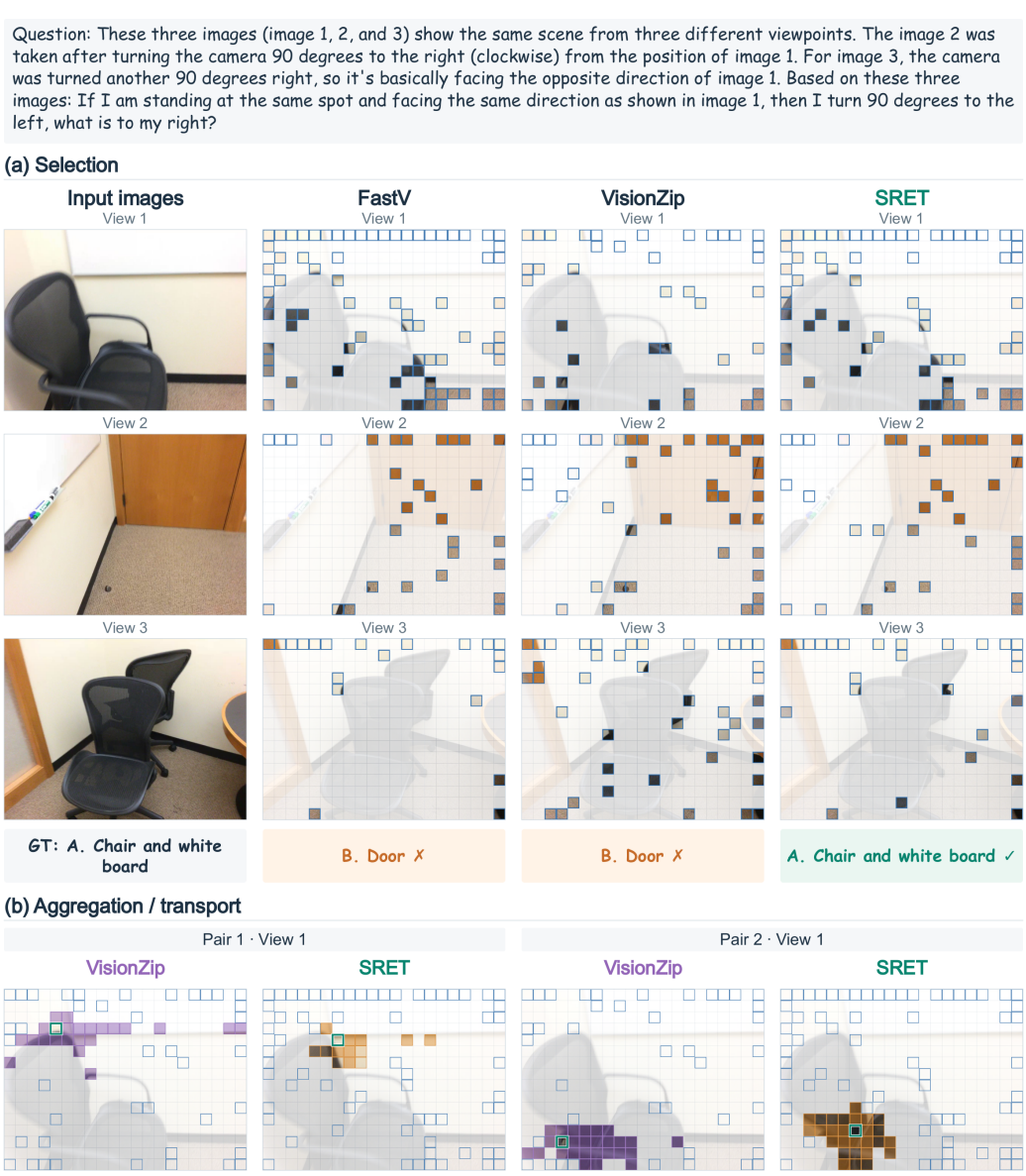

Figure C5. Reference-frame rotation across three views. The question asks what would lie to the observer's right after a specified turn, given three views of chairs and a whiteboard. Visual conventions follow Figure C1.

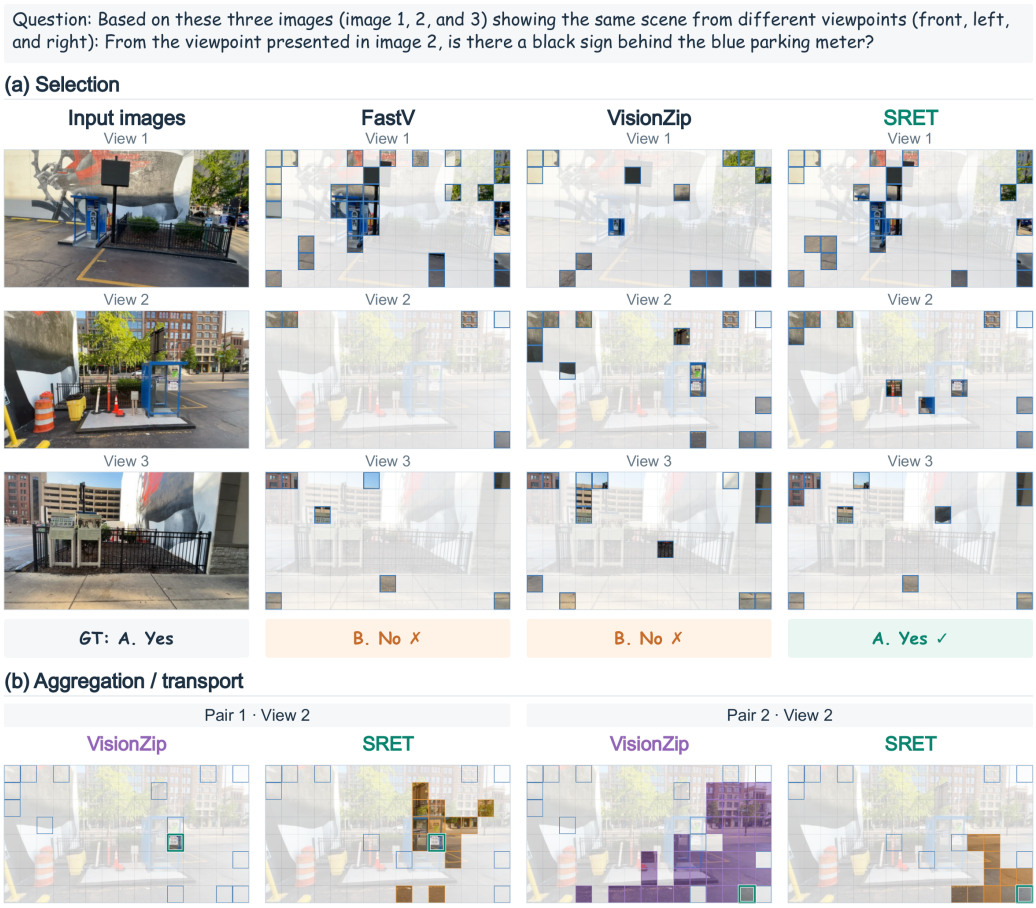

Figure C6. A viewpoint-dependent front–back relation. Three views provide evidence for deciding whether a black sign is behind the blue parking meter from the viewpoint of the second image. Visual conventions follow Figure C1.



Figure C7. A spatial relation across four views. The question asks which object is to the right of the pink ball from a specified viewpoint. Visual conventions follow Figure C1.

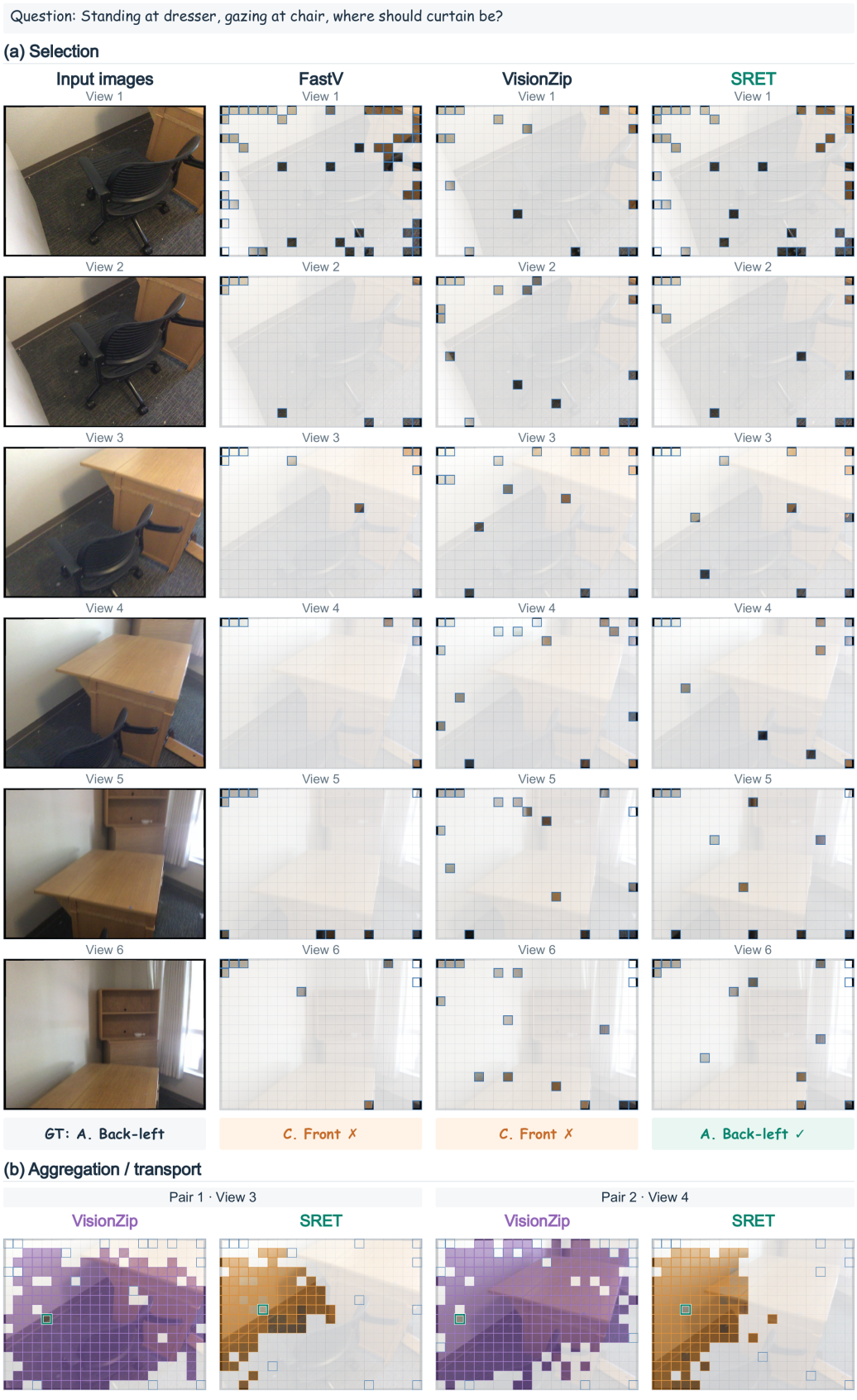

Figure C8. An imagined observer position across six views. The question asks where the curtain would be relative to an observer standing at the dresser and facing the chair. Visual conventions follow Figure C1.



Figure C9. Room-relative localization across eight views. The question asks where the desk lamp is located, using the mineral water bottle as a reference for the room's south wall. Visual conventions follow Figure C1.

D ALGORITHM PSEUDOCODE

Algorithm 1 **SRET representative selection and evidence transport**

Input: Visual representations H at the third decoder block; query scores d, visual importance v, visual/contextual/language-key features z^v, z^c, z^k, layout coordinates p, groups g, and budget K.

Output: Anchors A in sequence order, compact representations $\widetilde{H}_A$, and optional frozen-map parameters Θ.

1. $A_0 \leftarrow \mathrm{TopK}(d, K)$; $A \leftarrow A_0$.
2. Normalize d, v; compute shared mass m and within-group graph Φ.
3. Compute initial capacity mismatch Δ_{cap} from A_0, m.
4. `allow_cross` $\leftarrow (\Delta_{\mathrm{cap}} > 1/3)$.
5. **while** true:
6. Compute each group's best within-group exchange and gain in F.
7. **if** `allow_cross`:
8. Add best cross-group exchanges and gains for donor–receiver pairs.
9. In descending gain order, select positive-gain actions E with disjoint group endpoints.
10. **if** $E = \varnothing$: **break**.
11. Apply E simultaneously and update A; replaced anchors reenter the candidate set.
12. Sort A in sequence order; compute anchor assignments a_i and reconstructability scores c_i.
13. Select visual and contextual endpoints; compute projection coefficients t_i.
14. Construct P and W with self-loops for anchor sources; compute μ and $\bar{h}$.
15. Compute u, r, s, η and the updated representations $\widetilde{H}_A$.
16. $\Theta \leftarrow \varnothing$.
17. **if** the backbone has subsequent visual injections:
18. $\Theta \leftarrow (A, W, \mu, u, s, \eta)$.
19. **return** $A, \widetilde{H}_A, \Theta$.

Positive gains on line 9 must exceed numerical tolerance. Subsequent decoding retains all text tokens and aligns positions, masks, and KV states. At native visual injection sites, if $\Theta \neq \varnothing$, increments enter the compact sequence through $\mathcal{L}_H(D)$ in Appendix B.3. Frozen state is request-local.