BIG: BIDIRECTIONAL INFORMATION GAIN FOR VISUAL INSTRUCTION DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing visual instruction data selectors assess quality, coverage, or input contributions without jointly evaluating how the image and the particular instruction support the same reference answer. They can therefore retain examples whose answers require the image but are weakly constrained by the instruction, or follow from the instruction without requiring visual evidence. We propose **BIG (Bidirectional Information Gain)**, which measures both conditional contributions using one frozen vision-language scorer. Each gain is the relative reduction in reference-answer loss from adding one input given the other. To align gain scales across sources and directions, BIG converts them to within-source percentiles and ranks examples by their minimum under fixed source quotas. This makes the weaker contribution decisive, preventing a strong direction from compensating for a weak one. BIG consistently outperforms random and single-direction selection in aggregate performance across two model–data settings, and also leads the comparison with established selectors under a shared training recipe. On Vision-Flan with Qwen2-VL-2B, BIG exceeds full-data aggregate performance across seven benchmarks using 50% of the data. Ablation analysis shows that removing calibration or allowing compensation shifts selection toward examples with weaker support in one direction. Selected-data analysis shows that BIG systematically recovers examples with stronger contributions in the direction overlooked by single-direction criteria.

1 INTRODUCTION

Visual instruction tuning uses large mixtures of image-instruction-answer examples spanning different tasks and answer formats. Under a limited training budget, data selection must identify which supervision is most useful to retain. For each example, this requires asking how the image and the particular instruction support its reference answer. The image supplies instance-specific evidence, while the instruction specifies which content the answer should report.

However, selecting by only one contribution can overlook weakness in the other. First, strong visual dependence does not ensure that the particular instruction meaningfully constrains the answer: a broad description request leaves open which objects, attributes, or relations to report. Second, an informative instruction can constrain the answer while requiring little evidence from the particular image, because linguistic cues or general knowledge already supply the relevant information.

Figure 1 illustrates these two cases. A generic image-description request draws its answer from image content but leaves the object and attribute open, whereas the ostrich-position question specifies what the answer should report. Similarly, a light's general purpose can be inferred from the question and common knowledge, whereas the cat-color question requires an attribute of the depicted animal. These examples motivate two paired comparisons: what the image adds given the instruction, and what the particular instruction adds given the image, with the reference answer held fixed in both.

Existing selection approaches assess sample quality or learning utility (Li et al., 2024b;a), representation and concept–skill coverage (Lee et al., 2024), or the contribution of conditioning inputs. Contribution-based methods examine visual evidence through answer prediction with and without the image (Lee & Hwang, 2026; Dong et al., 2026), or question relevance through an evaluator's answer-validity judgment (Sun et al., 2026). Other methods combine multiple contribution signals, but use different prediction targets or conditioning structures (Zhou et al., 2025; Hong et al., 2026).

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.

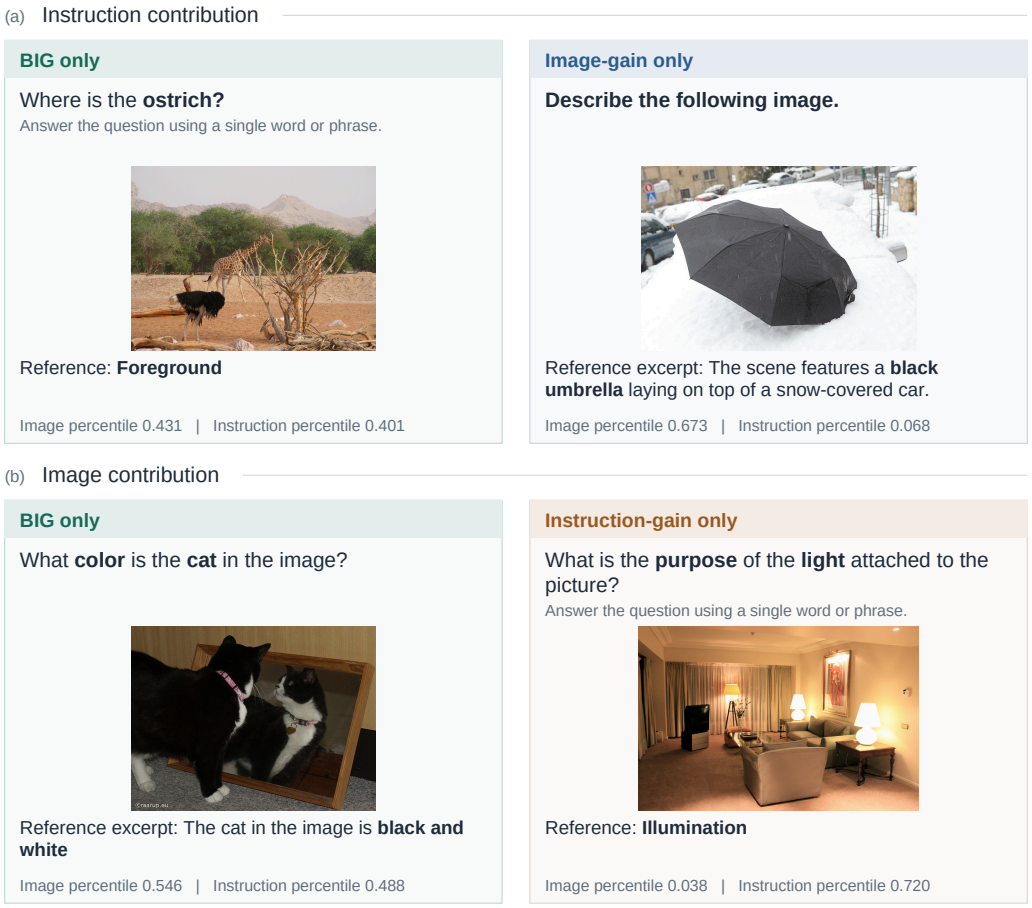

Figure 1: Complementary image and instruction contributions. **(a)** "Describe the following image" requires visual details but leaves the object and attribute to report unspecified; "Where is the ostrich?" constrains the answer to the depicted animal's position. **(b)** A light's general purpose can be inferred from its name and common knowledge, whereas identifying the cat's color requires the particular image. Each row compares BIG with one single-direction selector: the left and right examples are selected exclusively by BIG and that comparator, respectively. Examples come from the COCO source of LLaVA-665K. Percentiles give within-source standing in each gain, with larger values indicating stronger relative contributions. Long answers are excerpted.

To address this gap, we propose **BIG (Bidirectional Information Gain)**. One frozen vision–language scorer evaluates the same reference answer under full, image-absent, and instruction-absent conditions. Image gain measures how much adding the image improves answer prediction given the instruction, while instruction gain measures what the particular instruction adds given the image. Since raw gain scales vary across sources and directions, BIG converts each gain to a within-source percentile and uses the smaller percentile as the selection score. A strong direction therefore cannot compensate for a weak one. Selection retains high-scoring examples within fixed source quotas (§3, Figure 2).

Our contributions are as follows:

- **We measure both input contributions on a common answer-prediction objective.** Our bidirectional gains quantify what the image adds given the instruction and what the particular instruction adds given the image, using one frozen scorer and a fixed reference answer (§3.2).

- **We combine source calibration with a non-compensatory selection rule.** Within-source percentiles align the gain scales, and their minimum makes the weaker direction decisive. Analysis of the exchanged samples shows how these choices preserve stronger support in the weaker direction (§5.3). The selector operates under fixed source quotas without auxiliary scorer training or a tunable mixing weight (§3.3).
- **Joint selection yields consistent gains and recovers complementary supervision.** BIG leads Random and both single-direction criteria in aggregate performance across two model–data settings, and also leads the comparison with existing selectors. On Vision-Flan / Qwen2-VL-2B, BIG exceeds full-data aggregate performance using 50% of the data. Selected-data analysis shows that BIG introduces examples stronger in the direction omitted by each single-direction criterion (§5).

2 Related Work

We organize prior work by what its selection criterion evaluates: learning value, subset coverage, or input contribution. Each perspective connects to a different part of BIG's scoring and selection procedure.

2.1 Quality- and Utility-Based Selection

Quality- and utility-based methods evaluate how difficult an example is to learn or how much it can improve a model. IFD compares response difficulty with and without the instruction, and Superfiltering obtains this signal from a weaker model (Li et al., 2024b;a). DEITA combines complexity and quality with diversity, while Self-Filter combines instruction difficulty with redundancy control (Liu et al., 2024b; Chen et al., 2024b). Nuggets evaluates demonstration utility, whereas LESS and multimodal ICONS estimate an example's influence on target tasks (Li et al., 2024c; Xia et al., 2024; Wu et al., 2025).

BIG's gain definition follows the relative-loss perspective of IFD, with a baseline tailored to conditional input contribution. The baseline removes one input while retaining the other, and the gain measures the fraction of baseline answer loss removed by restoring the missing input. This makes each input's incremental predictive value the scoring target.

2.2 Diversity, Coverage, and Budget Allocation

Coverage-based methods seek subsets that represent the candidate pool. COINCIDE and CoIDO emphasize coverage, importance, and diversity (Lee et al., 2024; Yan et al., 2025), while Adapt-∞ adapts selection to evolving multimodal task mixtures (Maharana et al., 2025). ScalSelect selects examples by their contribution to instruction-conditioned visual representations (Wu et al., 2026). PreSel uses instruction relevance for task-budget allocation: it compares answer losses with and without the question, then aggregates this signal to allocate samples across tasks (Safaei et al., 2025).

BIG separates source allocation from candidate ranking. Fixed quotas preserve the source mixture, while within-source percentiles put the gains on a common relative scale before joint ranking. Where PreSel aggregates instruction relevance to set task budgets, BIG uses calibrated input contributions to decide which examples fill a fixed quota.

2.3 Contribution-Aware Multimodal Selection

The closest approaches examine how conditioning inputs affect prediction. Their key distinctions are the prediction target and which inputs remain available during the comparison. VIG and VisNec compare reference-answer prediction with and without the image (Lee & Hwang, 2026; Dong et al., 2026), leaving the instruction's contribution given the image unmeasured. CVS instead examines how adding the question changes a binary answer-validity judgment (Sun et al., 2026). BIG pairs the visual comparison with the instruction's effect on reference-answer prediction, keeping the prediction target fixed across both directions.

Joint selection also depends on what its component signals predict. mIFD combines a visual-following signal for an entire dialogue with an instruction-following signal for a text-only answer

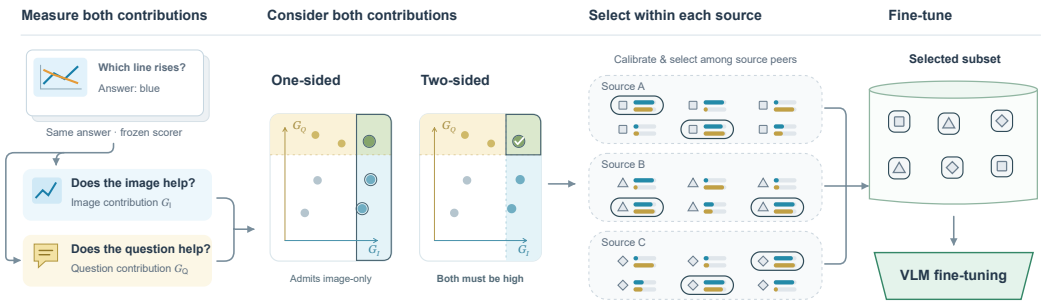

Figure 2: BIG end to end. One frozen scorer assesses how the image and instruction each support the same reference answer (left). The two contributions are considered jointly to favor examples supported by both inputs (centre). Selection operates within fixed source quotas, and the resulting subset is used for fine-tuning (right).

(Zhou et al., 2025). DICS combines answer-prediction and reverse instruction-reconstruction gains using the minimum (Hong et al., 2026). BIG aligns both components to the same reference answer, with the other input present in each comparison. Its minimum then acts on calibrated percentiles of these paired contributions, making the weaker contribution decisive.

3 METHOD

BIG measures how much the image and instruction each improve prediction of the same reference answer, then uses these two contributions to select training examples. Figure 2 summarizes the procedure.

3.1 PRELIMINARIES: DATA SELECTION AND ANSWER PREDICTION

Consider an eligible visual instruction pool $\mathcal{D} = \{(I_i, Q_i, A_i, h_i)\}_{i=1}^{N}$, where I_i, Q_i, and A_i are the image, instruction, and reference answer, and h_i identifies the source stratum. Let $\mathcal{D}_h$ contain the n_h examples from source h. Given a retention ratio r and the resulting sample budget K, selection constructs a subset $\mathcal{S}$ of K examples for instruction tuning.

We measure prediction difficulty using the mean negative log-likelihood (NLL) of the reference-answer tokens:

$$L_\theta(A \mid C) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(a_t \mid a_{<t}, C) = -\frac{1}{T}\log p_\theta(A \mid C), \tag{1}$$

where T is the answer length, C is the available conditioning input, and $a_{<t}$ is the reference-answer prefix supplied by teacher forcing. Prompt tokens are excluded. This empirical cross-entropy measures how well the model predicts the answer; perplexity expresses the same difficulty as $\mathrm{PPL}_\theta(A \mid C) = \exp(L_\theta(A \mid C))$. Lower loss or perplexity means the answer is easier to predict. BIG compares the relative reduction in NLL (log-perplexity) across input conditions. Appendix A.1 gives the likelihood, cross-entropy, and perplexity derivations.

3.2 BIDIRECTIONAL GAIN ESTIMATION

The image supplies instance-specific evidence, while the instruction constrains which aspects the answer should report. One frozen scorer evaluates the same reference answer under three conditions:

$$L_{\text{full}} = L(A \mid I, Q), \qquad L_{\text{txt}} = L(A \mid Q), \qquad L_{\text{img}} = L(A \mid I). \tag{2}$$

The image-absent pass preserves the prompt structure while suppressing visual embeddings and attention at visual positions. The instruction-absent pass retains the image and replaces the sample-specific instruction with a fixed scorer-specific turn (Appendix A).

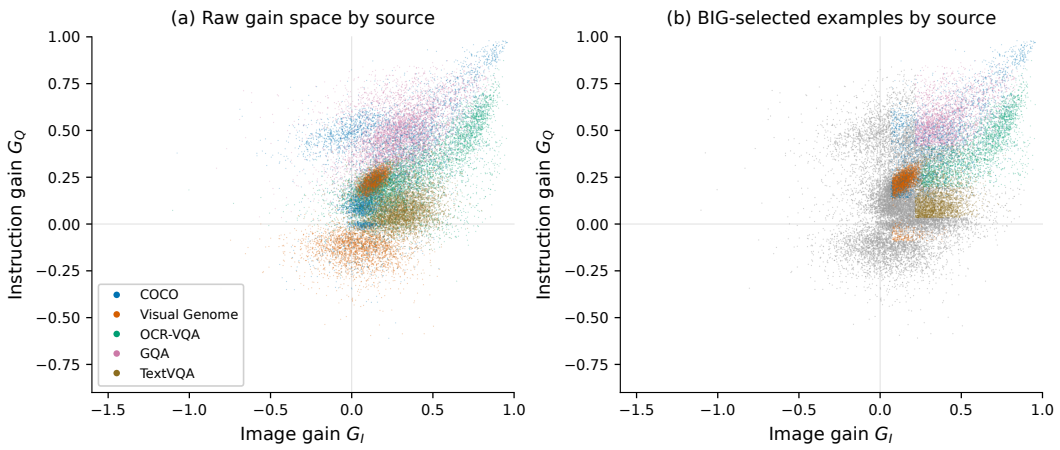

Figure 3: Raw gains and selected examples from LLaVA-665K, visualized using its five source strata for readability. Each source contributes 5,000 randomly sampled examples. **(a)** Image- and instruction-gain (G_I, G_Q) distributions. **(b)** The same examples with BIG-retained members highlighted.

We then define each gain as the fraction of its baseline loss removed by restoring the missing input:

$$G_I = 1 - \frac{L_{\text{full}}}{L_{\text{txt}}}, \qquad G_Q = 1 - \frac{L_{\text{full}}}{L_{\text{img}}}. \tag{3}$$

The **image gain** G_I measures what the image adds given the instruction; the **instruction gain** G_Q measures what the particular instruction adds given the image. Normalizing by each ablated loss expresses improvement relative to that baseline's prediction difficulty. Positive gains indicate reduced answer loss, while negative gains indicate increased loss. BIG retains these signed gains for ranking without clipping or sign filtering.

3.3 Source-Calibrated Joint Selection

Figure 3 shows that gain scales vary systematically across sources and directions, making raw minimum aggregation sensitive to scale differences between the two directions. Accordingly, BIG converts each gain to a within-source percentile before aggregation, while fixed source quotas preserve the source composition. Minimum aggregation then favors examples that rank highly in both directions.

Within-source calibration. Let rank_h denote ascending rank among eligible examples in $\mathcal{D}_h$. The percentiles

$$p_I = \frac{\text{rank}_h(G_I)}{n_h}, \qquad p_Q = \frac{\text{rank}_h(G_Q)}{n_h} \tag{4}$$

express each gain's relative standing within its source: a higher percentile indicates a stronger contribution in that direction. Calibration preserves each single-direction ordering and puts the directions on a common scale for aggregation.

Non-compensatory aggregation. Even after calibration, a mean allows a strong direction to offset a weak one. BIG instead uses the weaker percentile:

$$s_{\text{BIG}} = \min(p_I, p_Q). \tag{5}$$

A high score requires both percentiles to be high; increasing the stronger direction alone cannot improve it.

Source-stratified retention. For selection, each source receives a fixed quota k_h, proportional to n_h up to rounding, with $\sum_h k_h = K$. Selection takes the highest-scoring examples within each

source:

$$\mathcal{S}_h = \mathrm{Top}_{k_h}\{s(x) : x \in \mathcal{D}_h\}, \qquad \mathcal{S} = \bigcup_h \mathcal{S}_h. \tag{6}$$

The quota map controls source composition; calibration determines the relative scale used for joint ranking. Appendix C shows retention in percentile space. BIG requires three teacher-forced evaluations per example, with no scorer training or tunable signal-mixing weight.

4 Experimental Setup

We describe the model–data settings and selection comparisons in §4.1, and the training and evaluation protocols in §4.2.

4.1 Settings and selection arms

Vision-Flan-186K / Qwen2-VL-2B. Vision-Flan-186K (Xu et al., 2024) provides 186,087 eligible multimodal examples across 191 source strata defined by the released task-split field. Qwen2-VL-2B Base (Wang et al., 2024) is both the frozen scorer and the downstream target. Each arm retains 50% of the pool under fixed source quotas.

LLaVA-665K / LLaVA-1.5-7B. LLaVA-665K (Liu et al., 2024a) contains multimodal examples from five sources and 40,688 text-only examples. Each arm retains 50% of the multimodal portion and all text-only examples. LLaVA-1.5-7B stage-1 initialization is both the frozen scorer and downstream target. Appendix B gives the exact pool and subset sizes.

Controlled arms. **Random**, G_I**-only** (rank by p_I), G_Q**-only** (rank by p_Q), and **BIG** (rank by $\min(p_I, p_Q)$) share the eligible pool, source quotas, budget, target model, and training recipe. These comparisons isolate the within-source selection signal.

Cross-method comparison. We also compare complete selectors on Vision-Flan-186K with LLaVA-1.5-7B at a global 50% budget: **Random**, **Loss / Perplexity**, **CLIP-Score** (Radford et al., 2021), **COINCIDE** (Lee et al., 2024), **PreSel** (Safaei et al., 2025), **ScalSelect** (Wu et al., 2026), **IFD** (Li et al., 2024b), **VisNec** (Dong et al., 2026), and **BIG**. Each selector keeps its own scoring and budget-allocation procedure; all subsets and the full-data reference use the same downstream LoRA recipe.

4.2 Training and evaluation

The controlled settings fine-tune the language model and multimodal projector with the vision encoder frozen. The external comparison trains LoRA adapters and the projector. All runs use one epoch; Appendix B lists the training and decoding settings.

We evaluate on seven benchmarks: MMBench (Liu et al., 2025), TextVQA (Singh et al., 2019), ScienceQA (Lu et al., 2022), MME (Fu et al., 2025), MMStar (Chen et al., 2024a), AI2D (Kembhavi et al., 2016), and RealWorldQA (xAI, 2024). Relative performance uses the full-data model trained with the matching recipe as its reference:

$$\mathrm{RP}_b = 100 \times \frac{\text{performance}_b}{\text{full-data performance}_b}. \tag{7}$$

Avg. RP is the arithmetic mean over the seven benchmarks. Benchmark splits and native metrics appear in Appendix B.

5 Results

5.1 Consistent gains across settings

Joint selection consistently delivers stronger aggregate performance across settings, while the relative effectiveness of the two single-direction criteria varies (Table 1).

Table 1: Matched four-arm results at 50% multimodal retention. Entries are RP (%); Avg. RP is the mean over seven benchmarks. Bold marks the best arm within each setting.

Method	MMBench	TextVQA	ScienceQA	MME	MMStar	AI2D	RealWorldQA	Avg. RP
Vision-Flan-186K / Qwen2-VL-2B								
Random	100.24	98.43	100.40	97.89	98.28	98.76	97.12	98.73
G_I-only	**101.80**	**99.27**	**103.60**	97.99	100.56	98.43	94.91	99.51
G_Q-only	100.72	99.11	100.61	98.51	100.13	**100.06**	98.46	99.66
BIG	100.47	99.17	103.06	**98.67**	**101.42**	99.69	**101.12**	**100.51**
LLaVA-665K / LLaVA-1.5-7B								
Random	98.52	97.60	99.04	97.20	96.93	97.86	92.86	97.14
G_I-only	99.15	97.03	98.39	96.98	98.68	98.07	97.78	98.01
G_Q-only	99.71	**98.01**	98.19	**98.27**	98.55	97.52	95.25	97.93
BIG	**100.41**	97.73	**100.16**	97.09	**100.45**	**98.95**	**98.97**	**99.11**

Table 2: Cross-method comparison on Vision-Flan-186K / LLaVA-1.5-7B at 50% global retention. Each method keeps its own selection mechanism; downstream training uses the same shared LoRA recipe. Bold and underlined values indicate the best and second-best results in each column, respectively.

Method	MMBench	TextVQA	ScienceQA	MME	MMStar	AI2D	RealWorldQA	Avg. RP
Random	95.46	95.92	97.36	98.26	94.63	95.31	93.66	95.80
Loss / Perplexity	84.72	96.74	94.80	**100.76**	97.85	88.76	94.25	93.98
CLIP-Score [ICML-2021]	89.40	<u>103.61</u>	97.14	99.20	99.23	91.73	91.63	95.99
COINCIDE [EMNLP-2024]	95.76	101.55	99.02	94.83	**104.54**	90.65	90.07	96.63
IFD [NAACL-2024]	93.34	98.48	98.12	90.38	96.88	95.35	86.14	94.10
PreSel [CVPR-2025]	95.16	99.40	98.64	98.51	97.05	95.59	93.73	96.87
ScalSelect [ACM MM-2026]	<u>99.24</u>	97.91	98.20	<u>100.44</u>	98.82	<u>97.13</u>	90.85	97.51
VisNec [ECCV-2026]	**102.57**	**104.43**	<u>100.68</u>	96.22	92.75	95.89	<u>95.55</u>	<u>98.30</u>
BIG (ours)	98.78	97.69	**101.82**	97.32	<u>103.74</u>	**97.50**	**96.88**	**99.10**

On Vision-Flan / Qwen2-VL-2B, BIG reaches 100.51 Avg. RP, exceeding Random by 1.78 points and the stronger single-direction selector, G_Q-only, by 0.85 points. On LLaVA-665K / LLaVA-1.5-7B, it reaches 99.11 Avg. RP, exceeding Random by 1.97 points and the stronger single-direction selector in this setting, G_I-only, by 1.10 points. At the same 50% retention budget, BIG thus exceeds the full-data aggregate on Vision-Flan and closely approaches it on LLaVA-665K. Its consistent lead supports joint conditional contribution as a selection criterion across different model–data combinations.

5.2 Comparison with existing selection methods

BIG's advantage extends to comparisons with established data-selection methods under a shared LoRA recipe (Table 2).

On Vision-Flan-186K / LLaVA-1.5-7B, BIG achieves the highest Avg. RP at 99.10, ahead of ScalSelect and VisNec by 1.59 and 0.80 points, respectively, and Random by 3.30 points. These comparisons evaluate complete selectors with their own scoring and allocation strategies. Together with the controlled results, they establish the value of joint conditional contribution both relative to its component signals and alongside existing approaches to data selection.

5.3 Ablation of calibration and aggregation

We next examine calibration and aggregation separately on Vision-Flan-186K / Qwen2-VL-2B: **Raw-min** retains minimum aggregation but removes percentile calibration, using $\min(G_I, G_Q)$, while **Percentile-prod** and **Percentile-mean** retain calibration but replace the minimum with $p_I p_Q$ and $(p_I + p_Q)/2$, respectively. All variants share the eligible pool, 50% quotas, initialization, and full-parameter training recipe.

Table 3: Calibration and aggregation comparison on Vision-Flan-186K / Qwen2-VL-2B at 50% retention with fixed selection quotas. Entries are RP (%); Avg. RP averages the seven benchmarks. Bold marks the best result in each column.

Rule	MMBench	TextVQA	ScienceQA	MME	MMStar	AI2D	RealWorldQA	Avg. RP
BIG (Percentile-min)	100.47	99.17	**103.06**	**98.67**	**101.42**	**99.69**	**101.12**	**100.51**
Random	100.24	98.43	100.40	97.89	98.28	98.76	97.12	98.73
Percentile-prod	**100.96**	**99.86**	102.24	87.90	99.70	95.15	91.59	96.77
Raw-min	92.70	92.40	100.89	95.28	92.10	94.33	92.69	94.34
Percentile-mean	91.51	93.34	98.77	98.32	95.39	94.01	88.26	94.23

Table 4: Contribution profiles of examples exchanged between BIG and each ablation under the same source quotas. BIG-only and ablation-only are the examples selected exclusively by each rule in a pairwise comparison. Replaced is the fraction of the selected subset exchanged on each side. Entries are mean within-source percentiles of the weaker and stronger directions, $w = \min(p_I, p_Q)$ and $s = \max(p_I, p_Q)$.

		BIG-only		Ablation-only	
Ablation	Replaced (%)	Weaker	Stronger	Weaker	Stronger
Raw-min	15.56	0.394	0.632	0.210	0.683
Percentile-mean	11.46	0.380	0.494	0.236	0.850
Percentile-prod	6.69	0.366	0.460	0.290	0.821

Table 3 reports downstream performance, while Table 4 examines the examples exchanged between BIG and each ablation to identify how the two design choices affect selection.

Effect of calibration. With minimum aggregation fixed, removing calibration lowers Avg. RP from 100.51 to 94.34. In the exchanged samples, the mean weaker-direction percentile falls from 0.394 for BIG-only to 0.210 for Raw-min-only, while the stronger-direction mean rises only slightly, from 0.632 to 0.683. Calibration therefore prevents scale differences between the two contribution scores from favoring examples with weak support in one direction, even when minimum aggregation is retained.

Effect of aggregation. With calibration fixed, Percentile-mean and Percentile-prod reach Avg. RP of 94.23 and 96.77, both below Random. Percentile-mean-only examples have a much higher stronger-direction mean than BIG-only examples (0.850 versus 0.494), but a lower weaker-direction mean (0.236 versus 0.380). Percentile-prod shows the same trade-off, with a smaller decrease in the weaker-direction percentile. These exchanges demonstrate compensation in the selected data: a high score in one direction offsets substantially weaker support in the other.

Across all three comparisons, BIG preferentially retains examples with stronger support along their weaker direction.

5.4 Complementary supervision in the selected data

Relative to single-direction selectors, BIG recovers examples with stronger contributions in the direction that each selector overlooks. The two gains provide complementary rankings, with a size-weighted within-source Spearman correlation of $\rho = +0.3966$ on Qwen/Vision-Flan. Across the two settings, BIG replaces 18–23% of either single-direction subset while preserving source quotas.

On Qwen/Vision-Flan, we match introduced (BIG-only) and removed (comparator-only) examples on source, answer format, and within-source quartiles of answer length, instruction length, and full-condition NLL. Table 5 shows that the mean omitted-direction percentile rises from 0.194 to 0.584 relative to G_I-only and from 0.186 to 0.613 relative to G_Q-only. BIG therefore recovers stronger instruction constraints than G_I-only and stronger visual support than G_Q-only, even among examples with similar format, length, and prediction difficulty. Joint selection distinguishes how the inputs support the answer within otherwise similar supervision.

Table 5: Matched Qwen/Vision-Flan exchanges. Removed (comparator-only) and introduced (BIG-only) columns report mean within-source percentiles in the direction omitted by the comparator. Matched is the fraction of each exchange side retained by matching.

Comparator	Omitted direction	Removed	Introduced (BIG)	Matched (%)
G_I-only	Instruction (p_Q)	0.194	0.584	48.47
G_Q-only	Image (p_I)	0.186	0.613	51.52

On LLaVA-665K, the exchange also changes the form of supervision: relative to G_Q-only, BIG introduces longer, non-letter answers, with median answer lengths of 23 versus 1 token for introduced and removed examples. Joint selection thus changes both the contribution profile and the answer composition of the retained supervision. Appendix D reports the matching details and full attribute profiles.

5.5 Performance across conditional-contribution groups

Similar overall TextVQA scores conceal distinct strengths across contribution groups, revealing how the selection criterion shapes the model's downstream performance profile. We examine the four models from the Vision-Flan-186K / Qwen2-VL-2B controlled comparison (§5.1), each fine-tuned on a 50% subset. We split TextVQA validation examples into four groups at the median image and instruction gains, using a common pretrained Qwen2-VL-2B scorer to keep the groups fixed across selectors. We report TextVQA multi-reference soft scores on a 0–100 scale, with differences in score points.

The single-direction selectors favor different groups: G_I-only outperforms G_Q-only on high-image/low-instruction examples, while their ordering reverses on low-image/high-instruction examples. BIG achieves the highest score in the latter group, improving over G_I-only by 3.41 soft-score points, while giving up 3.15 points in the high-image/low-instruction group. These opposing differences persist across full-condition NLL quartiles (Appendix E). Joint selection therefore redistributes downstream strengths across contribution groups, with gains and trade-offs that the overall TextVQA score largely obscures.

6 Conclusion

BIG establishes joint conditional contribution as a practical criterion for visual instruction data selection. It evaluates both inputs against the same reference answer and ranks examples by their calibrated weaker contribution. Across model–data settings, the single-direction selectors reverse their aggregate ordering, while joint selection consistently leads in aggregate performance. Its advantage also extends to comparisons with existing selectors under a shared training recipe.

Analysis of the exchanged samples clarifies the distinct roles of the two components: calibration prevents scale differences from favoring examples with weak support in one direction, while minimum aggregation prevents a strong contribution from compensating for a weak one. Selected-data analysis shows that BIG recovers examples with stronger contributions in the direction overlooked by each single-direction criterion. TextVQA diagnostics reveal accompanying changes in downstream strengths, with gains and losses across contribution groups that overall scores obscure. Joint selection therefore changes both the supervision retained and the distribution of downstream performance.

Several limitations remain. The gains depend on the frozen scorer and the input-absence conditions, so rankings reflect that model's predictive capabilities. Source-wise calibration requires a meaningful source partition, and fixed quotas do not adapt source budgets to downstream tasks. Scoring every candidate under three conditions adds preprocessing cost. The observed group-specific trade-offs also leave open how best to balance joint contribution with task-specific learning needs. These limitations motivate cheaper scoring and selection rules that adapt source allocation and contribution balance to the target task.

AI USE STATEMENT

We used AI tools to assist with implementing our research ideas, drafting the manuscript, searching for relevant literature, and polishing the writing.

REPRODUCIBILITY STATEMENT

We provide details of our method and experimental setup in the main text, with additional implementation details and experimental settings included in the appendix.

REFERENCES

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024a. doi: 10.52202/079017-0850.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4156–4172, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.246.

Mingkang Dong, Hongyi Cai, Jie Li, Sifan Zhou, Bin Ren, Kunyu Peng, and Yuqian Fu. VisNec: Measuring and leveraging visual necessity for multimodal instruction tuning. In *Computer Vision – ECCV 2026*, pp. 531–548, 2026. doi: 10.1007/978-3-032-37577-3_29.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4899.

Yuyang Hong, Jinhui Guo, Jiaqi Gu, Lubin Fan, Ruixiang Wang, Kun Ding, Yue Wu, Shiming Xiang, and Jieping Ye. DICS: Exploring data intrinsic consistency for visual instruction selection. *arXiv preprint arXiv:2608.30209*, 2026. doi: 10.48550/arXiv.2608.30209. Accepted at Findings of EMNLP 2026.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision – ECCV 2016*, pp. 235–251, 2016. doi: 10.1007/978-3-319-46493-0_15.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5060–5080, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.291.

Seulbi Lee and Sangheum Hwang. Focusing where vision matters: Selective training for large vision language models via visual information gain. *arXiv preprint arXiv:2602.17186*, 2026. doi: 10.48550/arXiv.2602.17186. Accepted at ICML 2026.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 14255–14273, 2024a. doi: 10.18653/v1/2024.acl-long.769.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 7602–7635, 2024b. doi: 10.18653/v1/2024.naacl-long.421.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. One-shot learning as instruction data prospector for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4586–4601, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.252.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024a. doi: 10.1109/CVPR52733.2024.02484. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://openreview.net/forum?id=BTKAeLqLMw.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pp. 216–233, 2025. doi: 10.1007/978-3-031-72658-3_13.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022. doi: 10.52202/068431-0182.

Adyasha Maharana, Jaehong Yoon, Tianlong Chen, and Mohit Bansal. Adapt-∞: Scalable continual multimodal instruction tuning via dynamic data selection. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=EwFJaXVePU.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, and Shao-Yuan Lo. Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14247–14256, 2025. doi: 10.1109/CVPR52734.2025.01329.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8317–8326, 2019. doi: 10.1109/CVPR.2019.00851.

Peng Sun, Huawen Shen, Yi Ban, Tianfan Fu, Yanbo Wang, and Yuqiang Li. Does the question really matter? training-free data selection for vision-language SFT. *arXiv preprint arXiv:2603.09715v1*, 2026. doi: 10.48550/arXiv.2603.09715. URL https://arxiv.org/abs/2603.09715v1.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. doi: 10.48550/arXiv.2409.12191.

Changti Wu, Jiahuai Mao, Yuzhuo Miao, Shijie Lian, Bin Yu, Xiaopeng Lin, Cong Huang, Lei Zhang, and Kai Chen. ScalSelect: Scalable training-free multimodal data selection for efficient visual instruction tuning. *arXiv preprint arXiv:2602.11636*, 2026. doi: 10.48550/arXiv.2602.11636. Accepted at ACM MM 2026.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. ICONS: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*, 2025. doi: 10.48550/arXiv.2501.00654.

xAI. RealWorldQA. Dataset release, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA. Introduced in Grok-1.5 Vision Preview, 12 April 2024.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-Flan: Scaling human-labeled tasks in visual instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15271–15342, 2024. doi: 10.18653/v1/2024.findings-acl.905.

Yichen Yan, Ming Zhong, Qi Zhu, Xiaoling Gu, Jinpeng Chen, and Huan Li. CoIDO: Efficient data selection for visual instruction tuning via coupled importance-diversity optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167045–167073, 2025. doi: 10.52202/085713-5569.

Shijie Zhou, Ruiyi Zhang, Yufan Zhou, and Changyou Chen. A high-quality text-rich image instruction tuning dataset via hybrid instruction generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10091–10110, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.674/.

A Additional method details

A.1 Likelihood, cross-entropy, and perplexity

BIG compares conditional prediction difficulty using the answer-token NLL in Eq. 1. The following derivation connects this score to autoregressive likelihood, maximum-likelihood training, and perplexity.

Reference-answer likelihood. Let $A = (a_1, \ldots, a_T)$ be a reference answer with T tokens, and let C denote the conditioning inputs available to a vision–language model with parameters θ. An autoregressive model assigns a probability to the answer by predicting its tokens in sequence:

$$p_\theta(A \mid C) = \prod_{t=1}^{T} p_\theta(a_t \mid a_{<t}, C). \tag{8}$$

Here $a_{<t}$ is the reference-answer prefix preceding token a_t. During scoring, this prefix is supplied to the model at each position, a procedure known as teacher forcing. Thus, every conditional probability evaluates the next reference token in the context of the same reference answer.

The mean NLL in Eq. 1 connects to the information-theoretic notion of cross-entropy. For a fixed prediction context $c = (a_{<t}, C)$, let $q(\cdot \mid c)$ denote the true next-token distribution over vocabulary $\mathcal{V}$. Its cross-entropy with the model distribution satisfies

$$\begin{aligned} H(q, p_\theta \mid c) &= -\sum_{v \in \mathcal{V}} q(v \mid c) \log p_\theta(v \mid c) \\ &= H(q \mid c) + D_{\mathrm{KL}}(q(\cdot \mid c) \,\|\, p_\theta(\cdot \mid c)) \,. \end{aligned} \tag{9}$$

The entropy $H(q \mid c)$ captures uncertainty in the reference distribution, while the Kullback–Leibler divergence D_{KL} measures the mismatch between that distribution and the model. Averaging the observed token losses in Eq. 1 gives an empirical cross-entropy score for the reference answer.

The same token losses underlie maximum-likelihood estimation. For training examples (A_j, C_j) with answer lengths T_j, maximizing answer likelihood is equivalent to minimizing total negative log-likelihood:

$$\arg\max_{\theta} \sum_{j} \log p_\theta(A_j \mid C_j) = \arg\min_{\theta} \sum_{j} T_j L_\theta(A_j \mid C_j). \tag{10}$$

This equivalence explains why cross-entropy is a natural measure of how well a model predicts reference answers. BIG uses this predictive objective to compare the information supplied by different input conditions.

Perplexity as prediction difficulty. Perplexity (PPL), widely used to evaluate language models, expresses the same average token loss on an exponential scale:

$$\mathrm{PPL}_\theta(A \mid C) = \exp(L_\theta(A \mid C)) = p_\theta(A \mid C)^{-1/T}. \tag{11}$$

It is the inverse geometric mean of the probabilities assigned to the reference-answer tokens. For example, if the model assigns probability $1/4$ to every reference token, the answer has perplexity 4. Lower perplexity indicates that the model finds the answer easier to predict. Comparing the same answer under different input conditions therefore reveals how those inputs change its prediction difficulty.

A.2 Conditional scoring

Answer-token losses. Each condition teacher-forces the same reference answer and averages negative log-likelihood over answer tokens, excluding prompt tokens. For multi-turn records, scoring uses the first user instruction and its first assistant response. Training uses the selected conversations up to the sequence-length limit.

Image-absent condition. Following the blind-forward construction of VisNec (Dong et al., 2026), the full prompt is retained after the processor expands the image placeholder. At visual-token positions, input embeddings are zeroed and attention is suppressed, preserving the text, answer labels, and sequence layout. For Qwen2-VL, this pass also preserves the full pass's M-RoPE position IDs.

Instruction-absent condition. The image is retained while the sample-specific instruction is replaced with a fixed scorer-specific turn. For LLaVA-1.5, the user content is `"<image>\nAnswer based on the image."`; for Qwen2-VL, it is the image block `<|vision_start|><|image_pad|><|vision_end|>` inside the standard ChatML frame, without an instruction sentence.

B Training and evaluation details

Pool sizes and selection budgets. Vision-Flan has 186,087 eligible examples across 191 strata; each controlled 50% subset contains 93,047 examples. The LLaVA mixture has 665,297 examples: 624,609 multimodal and 40,688 text-only. Each controlled arm selects 312,304 multimodal examples and retains all text-only data, yielding 352,992 training examples.

Training recipes. Table 6 lists the full-parameter and LoRA configurations. All runs train for one epoch.

Evaluation. We use VLMEvalKit with MMBench's English development split, the TextVQA and ScienceQA validation splits, AI2D's test split, and the released MME, MMStar, and RealWorldQA evaluation sets. Native metrics are accuracy for MMBench, ScienceQA, MMStar, AI2D, and RealWorldQA; normalized multi-reference soft accuracy for TextVQA; and the MME perception score, summing single-question and paired-question accuracy across perception categories. MMBench uses circular evaluation. Answers are scored by benchmark-specific rules without an LLM judge. Generation is capped at 1,024 new tokens, with sampling disabled for LLaVA and top-$k = 1$ for Qwen.

Table 6: Training recipes used by the controlled four-arm settings and the external comparison.

Field	Vision-Flan / Qwen2-VL	LLaVA-665K / LLaVA-1.5	External / LLaVA + Vision-Flan
Tuning scheme	full parameters	full parameters	shared LoRA
Initialization / template	Qwen2-VL-2B Base / standard ChatML	LLaVA-1.5-7B stage-1 / LLaVA-1.5	LLaVA-1.5-7B stage-1 / official LLaVA-1.5
Trainable modules	language model + projector	language model + projector	LoRA adapters + projector
Frozen module	vision encoder	vision encoder	vision encoder
Effective batch	128	128	128
Learning rate	1×10^{-5}	2×10^{-5}	adapters: 2×10^{-4}; projector: 2×10^{-5}
Sequence length	2048	2048	2048
Optimizer / schedule	AdamW / cosine	AdamW / cosine	AdamW / cosine
Warmup / weight decay	0.03 / 0	0.03 / 0	0.03 / 0
Gradient clipping	1	1	1
Precision / sharding	bf16 / ZeRO-2	bf16 / ZeRO-2	bf16 / ZeRO-3
Seed / data seed	42 / 42	42 / 42	42 / 42
Optimizer steps	727	2,758	727
LoRA fields	–	–	$r = 128, \alpha = 256$, dropout $= 0.05$, target $=$ `all`

C CALIBRATED-SPACE RETENTION

Figure 4 shows retention in calibrated (p_I, p_Q) space. Percentiles are computed within the full eligible source pools; each cell gives the retained fraction in a 10×10 percentile grid.

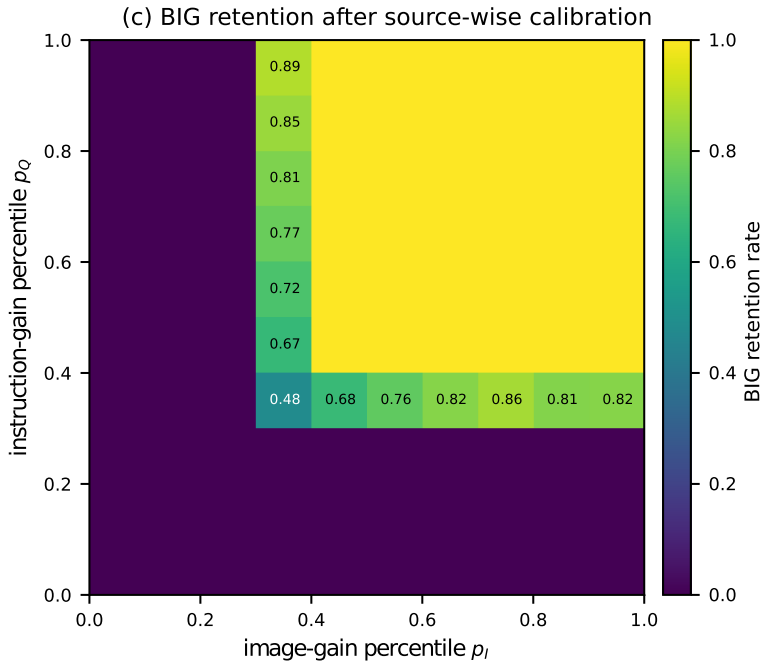

Figure 4: Source-wise calibration and BIG retention on a subset of LLaVA-665K. **(a)** Raw image and instruction gains, colored by source. **(b)** BIG-selected examples in the same raw-gain space. **(c)** Retention after within-source percentile calibration: each cell in the 10×10 grid shows the fraction of examples retained, with color indicating this fraction.

D SELECTED-DATA ATTRIBUTES AND MATCHING

Rankings and exchanged examples. On Qwen/Vision-Flan, the size-weighted within-source Spearman correlation between the gains is $\rho = +0.3966$ across 189 strata with at least 100 ex-

Table 7: Training members exchanged by BIG relative to each single-direction selector. Each exchange side counts E^+ or E^- separately; replaced share is $|E^-|/|\mathcal{S}_X|$.

Setting	Comparator	Each exchange side	Replaced share
Qwen / Vision-Flan	G_I-only	19,697	21.17%
Qwen / Vision-Flan	G_Q-only	19,674	21.14%
LLaVA / 665K	G_I-only	56,688	18.15%
LLaVA / 665K	G_Q-only	71,424	22.87%

amples. At 50% retention, the single-direction subsets have Jaccard overlap $J = 0.4700$, above the independent-subset reference $r/(2-r) = 0.3333$. BIG overlaps with them at $J = 0.6506$ (G_I) and 0.6509 (G_Q). For comparator X, let $E^+(X) = \mathcal{S}_{\mathrm{BIG}} \setminus \mathcal{S}_X$ denote introduced examples and $E^-(X) = \mathcal{S}_X \setminus \mathcal{S}_{\mathrm{BIG}}$ removed examples. BIG replaces 18–23% of either single-direction subset (Table 7) without changing source quotas.

Matched contribution profiles. Matching uses source, lexical answer format, and within-source quartiles of answer-token count, instruction character count, and L_{full}, computed over the full eligible source pools. Answer formats are single letters A/B/C/D, yes/no, and other. Matching covers 48.47% and 51.52% of the exchanges against G_I-only and G_Q-only, respectively. Among these matched examples, the mean omitted-direction percentile is 0.584 versus 0.194 for G_I-only and 0.613 versus 0.186 for G_Q-only (Table 8). The difference persists among examples with similar measured attributes.

Table 8: Matched Qwen / Vision-Flan exchange attributes (means). Answer length is a token count; instruction length is in characters. Omitted pct. is the percentile in the direction unused by the comparator.

Comparator	Side	n	Answer length	Instr. chars	Full NLL	Omitted pct.
G_I-only	E^+	9,547	18.05	229.57	3.944	0.584
G_I-only	E^-	9,547	17.87	228.68	3.970	0.194
G_Q-only	E^+	10,137	18.82	243.74	3.794	0.613
G_Q-only	E^-	10,137	19.79	241.29	3.794	0.186

Answer composition on LLaVA-665K. Relative to G_Q-only, BIG introduces longer, non-letter answers within the fixed source quotas. Introduced versus removed examples have median answer lengths of 23 versus 1 token, and letter-only shares of 0.024% versus 64.88%.

E Conditional-contribution diagnostic

Reference answers and contribution groups. The pretrained Qwen2-VL-2B Base scorer evaluates all 5,000 TextVQA validation examples (3,166 distinct RGB images). We score the normalized modal reference answer and evaluate predictions with the standard multi-reference TextVQA metric. The full condition retains the instruction and short-answer suffix; scoring uses the three conditions in Appendix A, including template terminators. Thirteen references normalize to empty strings, yielding terminator-only targets in the low/low group. Negative gains are retained. Each gain is converted to a diagnostic-set percentile and split at .5 to define four groups, shared across models and independent of their predictions.

Group performance. The single-direction models differ most clearly in the asymmetric groups: G_I-only leads G_Q-only on high-image/low-instruction examples (72.25 versus 69.80), while G_Q-only leads on low-image/high-instruction examples (72.07 versus 69.79). BIG reaches 73.20 in the latter group, improving over G_I-only by 3.41 points and G_Q-only by 1.12 points. BIG falls below G_I-only by 3.15 points in high-image/low-instruction and by 1.10 in high/high; Random leads in low/low. Overall scores are 70.98, 71.58, 71.47, and 71.51 for Random, G_I-only, G_Q-only, and BIG, respectively.

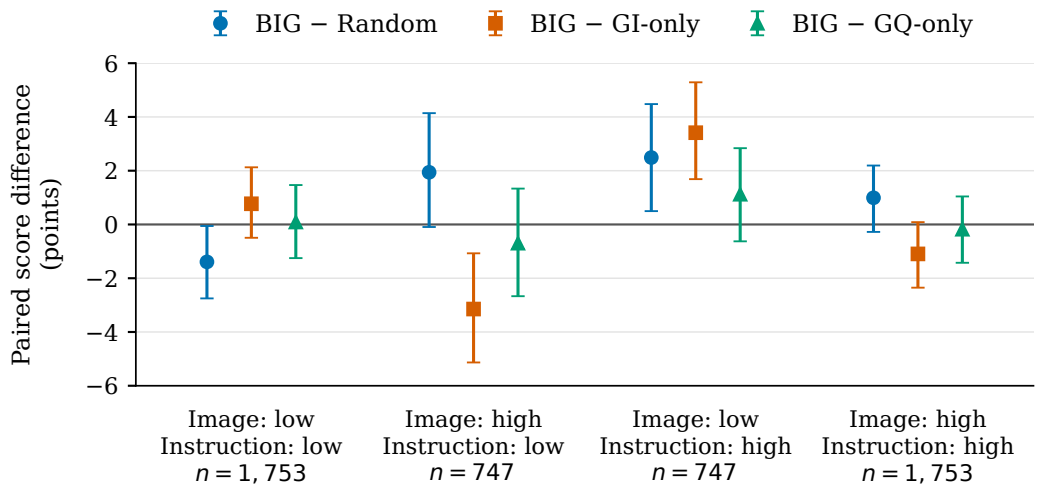

Figure 5: Performance trade-offs across conditional-contribution groups on TextVQA. Points show BIG minus each comparator in original soft-score points ($\times 100$), with all four groups fixed by the common pretrained scorer. I and Q denote image and instruction contribution. Bars are 95% percentile intervals from 2,000 paired bootstrap resamples of shared-image clusters within each group.

Table 9: TextVQA contribution groups defined by the common frozen scorer. Scores are original multi-reference soft scores $\times 100$; bold marks the largest mean in each group.

Group (I, Q)	n	Images	Random	G_I-only	G_Q-only	BIG
low, low	1,753	1,474	**62.95**	60.79	61.47	61.56
high, low	747	688	67.16	**72.25**	69.80	69.10
low, high	747	698	70.71	69.79	72.07	**73.20**
high, high	1,753	1,490	80.76	**82.85**	81.93	81.76

The 95% intervals in Figure 5 use 2,000 paired bootstrap resamples of shared-image clusters within each group (seed 20260919), averaging over resampled questions. Images can occur in multiple groups. Training-image and exact-QA overlap remain unknown from the available identity information.

Full-condition difficulty. Mean L_{full} is 6.225, 4.324, 4.416, and 3.215 in low/low, high/low, low/high, and high/high; median target lengths are 4, 5, 4, and 5 tokens, including terminators. To examine difficulty effects, we split the diagnostic set into four full-NLL quartiles while keeping contribution groups fixed. BIG minus G_I-only remains positive in all four low/high cells and negative in all four high/low and high/high cells (Table 10).

Table 10: Fixed full-NLL sensitivity analysis. Entries are paired BIG minus comparator differences in TextVQA soft-score points. Quartiles increase from lowest (1) to highest (4) full NLL.

Group (I, Q)	NLL quartile	n	Δ Random	ΔG_I-only	ΔG_Q-only
low, low	1	30	+1.00	-4.33	+0.00
low, low	2	211	+1.52	+1.85	-0.38
low, low	3	539	-1.04	-0.89	+0.19
low, low	4	973	-2.29	+1.61	+0.14
high, low	1	140	+5.57	-2.57	-3.50
high, low	2	267	+0.30	-4.08	-0.67
high, low	3	226	+3.45	-2.21	-0.13
high, low	4	114	-1.67	-3.51	+1.58
low, high	1	135	+2.89	+8.00	+2.89
low, high	2	236	+2.58	+1.14	+0.13
low, high	3	269	+1.19	+3.49	+1.64
low, high	4	107	+5.05	+2.43	-0.19
high, high	1	944	+1.95	-0.85	-0.88
high, high	2	536	-0.06	-1.47	+0.17
high, high	3	216	+0.00	-0.88	+2.36
high, high	4	57	-1.23	-2.46	-1.40