CoIDO: Efficient Data Selection for Visual Instruction Tuning via Coupled Importance-Diversity Optimization

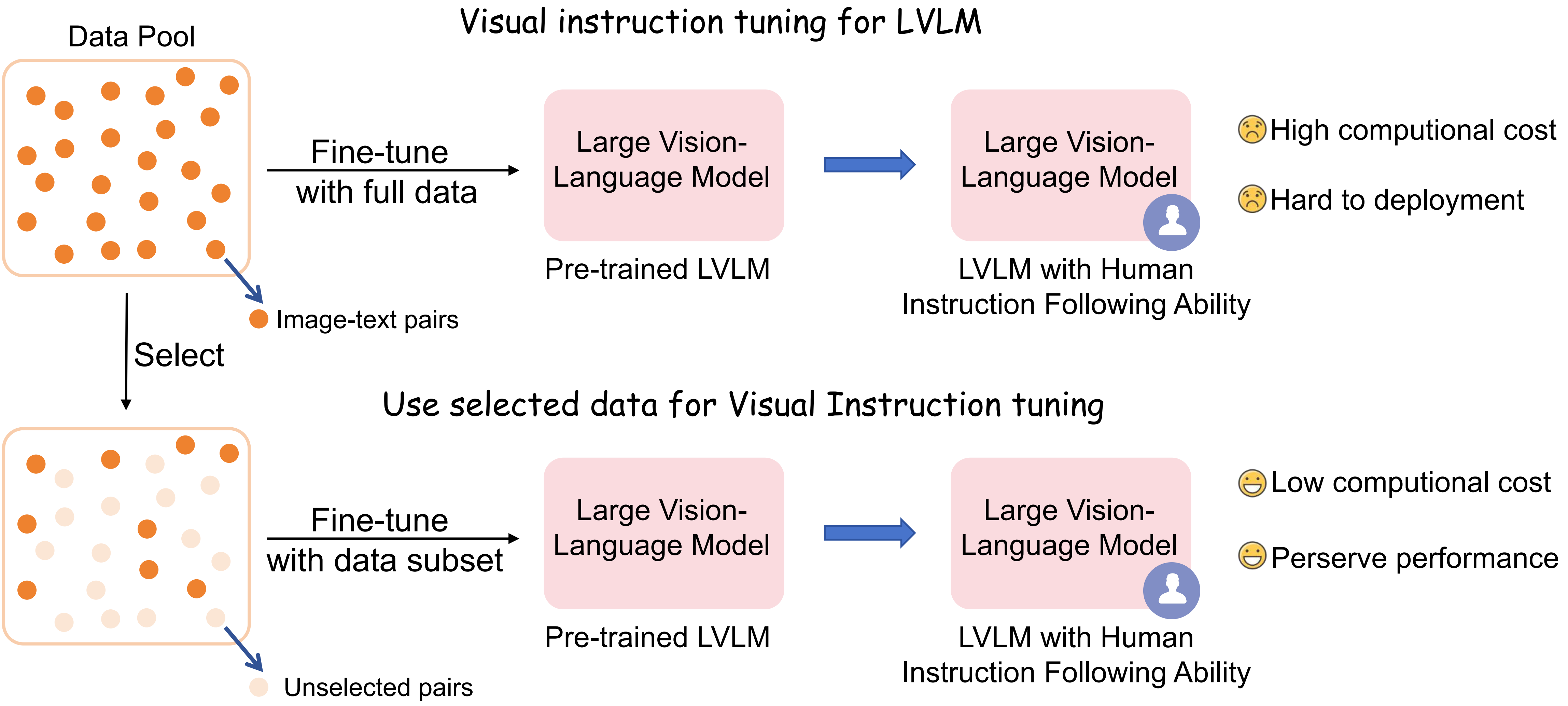
Yichen Yan, Ming Zhong, Qi Zhu, Xiaoling Gu, Jinpeng Chen, Huan Li*

浙江大學 ZHEJIANG UNIVERSITY

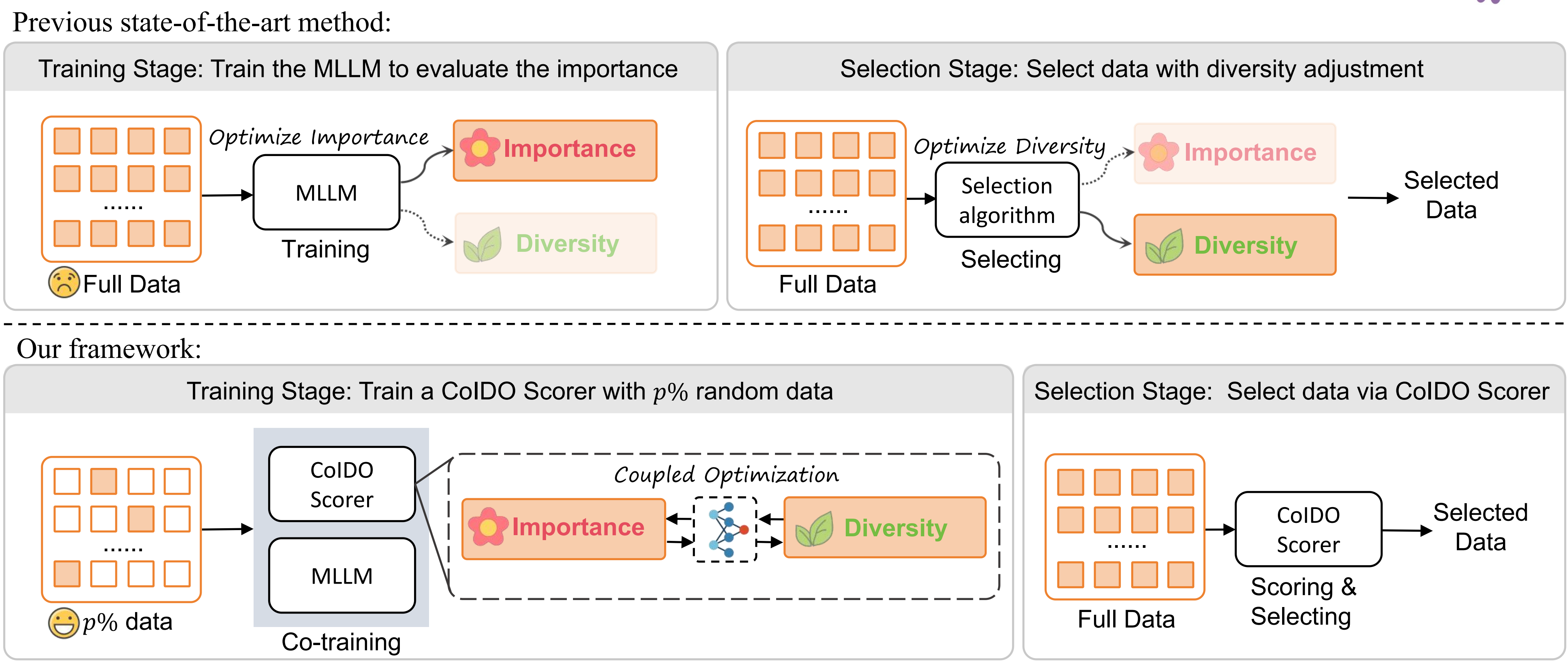
杭州電子科技大学 HANGZHOU DIANZI UNIVERSITY

北京郵電大學 BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Introduction



Motivation

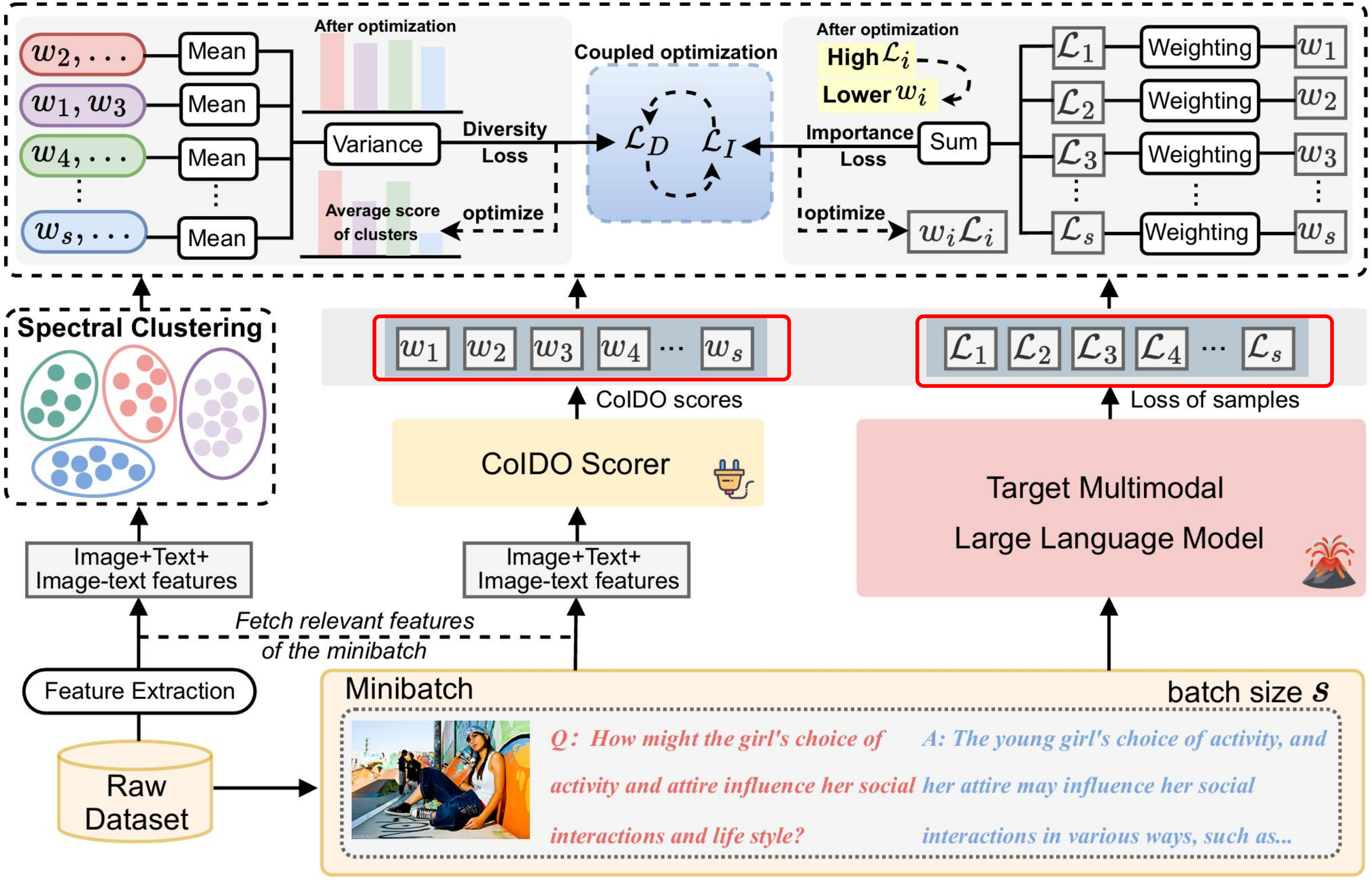


Data Importance: Whether data sample is valuable and informative (e.g., detailed caption, high resolution)

Data Diversity: Whether subset is diverse (including various tasks and domains)

Method (framework)

NEURAL INFORMATION PROCESSING SYSTEMS



Importance Loss

For per sample:

Higher CE loss → More challenging → $\boxed{\text{More Important}}$

$$\mathcal{L}_I = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w_{ik}\cdot \boxed{\mathrm{CE}(y_{ik},\hat{y}_{ik})},$$

Lower w_{ik} → Higher $\mathcal{L}_{ik}$

Diversity Loss

For per batch:

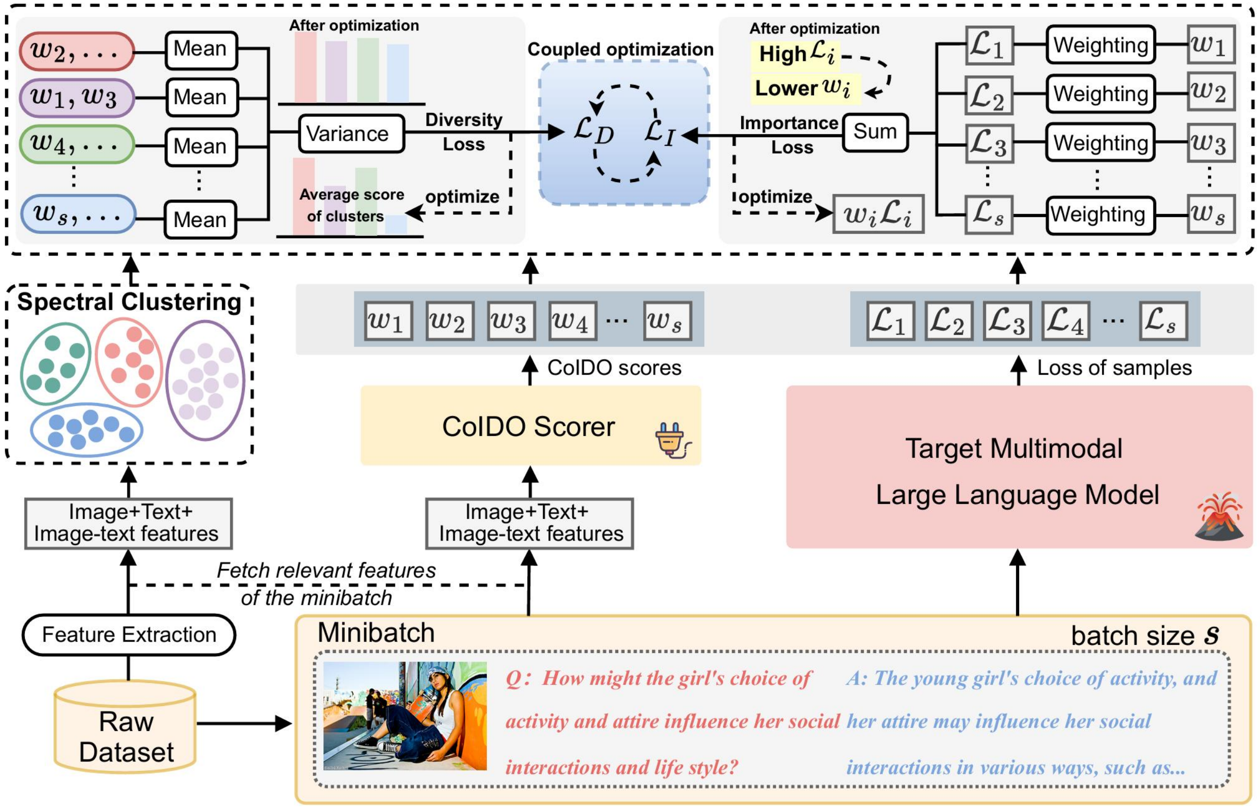
$$\mathcal{L}_D = \mathrm{Var}(\{\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_m\}),$$

$$\bar{w}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} w_{ik},$$

Balanced weight distribution across clusters

Method (loss function)

NEURAL INFORMATION PROCESSING SYSTEMS



Maximum Likelihood Estimation by **homoscedastic uncertainty**

$$\log p(\mathbf{y}, \mathbf{w} \mid \theta, \sigma_I, \sigma_D) = \sum_{i,k} \log p(y_{ik} \mid x_{ik}, \theta, \sigma_I) + \sum_i \log p(\bar{w}_i \mid \theta, \sigma_D).$$

For importance objective:

$$p(y_{ik} \mid x_{ik}, \theta, \sigma_I, w_{ik}) = \text{Softmax}\left(\frac{w_{ik}}{\sigma_I^2} f_\theta(x_{ik})\right),$$

we use **Taylor expansion**:

$$-\sum_{i,k} \log p(y_{ik} \mid x_{ik}, \theta, \sigma_I) = \frac{1}{\sigma_I^2}\mathcal{L}_I + \log \sigma_I.$$

For diversity objective:

$$p(\bar{w}_i|\theta, \sigma_D) = \mathcal{N}(\bar{w}_i; \mu, \sigma_D^2).$$

$$-\sum_i \log p(\bar{w}_i \mid \theta, \sigma_D) = \frac{1}{2\sigma_D^2}\mathcal{L}_D + \log \sigma_D,$$

Coupled Optimization: $\mathcal{L}_{\text{total}} = \frac{1}{\sigma_I^2}\mathcal{L}_I + \frac{1}{2\sigma_D^2}\mathcal{L}_D + \log \sigma_I + \log \sigma_D.$

Experiments (overall performance)

MLLM Training Data Cost: the proportion of data used to train the selection model relative to model fine-tuning.

Table 1: Overall performance and efficiency comparison of selection approaches across various multimodal evaluation benchmarks, with the best measures in bold and the second-best underlined.

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench en	MMBench cn	LLaVA-Bench	*Rel.* (%)	MLLM Training Data Cost (%)	Total FLOPs
Full Data	79.1	63.0	47.8	68.4	58.2	86.4	1476.9	66.1	58.9	67.9	100	\	10.2E
Model-free Methods													
RANDOM	75.9	59.3	43.6	68.6	55.3	85.9	1461.0	60.3	53.3	64.5	95.1	\	\
CLIP-SCORE [29]	73.4	51.4	43.0	65.0	54.7	85.3	1331.6	55.2	52.0	66.2	91.2	\	\
EL2N [32]	76.2	58.7	43.7	65.5	53.0	84.3	1439.5	53.2	47.4	64.9	92.0	\	\
PERPLEXITY [33]	75.8	57.0	47.8	65.1	52.8	82.6	1341.4	52.0	45.8	68.3	91.6	\	\
SEMDEDUP [30]	74.2	54.5	46.9	65.8	55.5	84.7	1376.9	52.2	48.5	70.0	92.6	\	\
D2-PRUNING [31]	73.0	58.4	41.9	69.3	51.8	85.7	1391.2	65.7	57.6	63.9	94.8	\	\
SELF-SUP [30]	74.9	59.5	46.0	67.8	49.3	83.5	1335.9	61.4	53.8	63.3	93.4	\	\
Model-involved Methods													
SELF-FILTER [20]	73.7	58.3	**53.2**	61.4	52.9	83.8	1306.2	48.8	45.3	64.9	90.9	100	31.2E
TIVE♣ ♦ [17]	76.0	58.4	44.6	69.8	53.3	85.7	1448.4	**66.9**	**58.7**	63.4	96.7	100+8	11.7E
ICONS♣ ♢ [19]	77.0	**60.4**	45.5	**70.4**	54.5	**86.1**	1447.7	64.6	54.0	66.9	97.1	100+5+2.2	12.6E
COINCIDE [21]	76.5	59.8	46.8	69.2	**55.6**	**86.1**	**1495.6**	63.1	54.5	67.3	97.4	100	4.9E
CoIDO (Ours)	**77.2**	**60.4**	47.1	69.4	**55.6**	85.4	1450.2	63.8	56.7	**70.1**	**98.2**	**20**	**4.2E**

COIDO outperforms all competitors in terms of both efficiency (lowest training FLOPs) and aggregated accuracy

Experiments (ablation and parameter study)

NEURAL INFORMATION PROCESSING SYSTEMS

Table 2: Ablations of optimization methods (the best in bold and the second-best underlined).

Loss Function	VQAv2	GQA	Vizwiz	SQA-I	TextVQA	POPE	MME	MMBench(en)	MMBench(cn)	LLAVA-B	*Rel.* (%)
$\mathcal{L}_I$	**77.9**	48.9	44.6	59.7	52.5	**86.2**	1393.5	51.1	44.9	<u>64.9</u>	89.0
$\mathcal{L}_I + \mathcal{L}_D$	74.5	55.8	46.4	67.3	52.6	83.5	1339.7	57.0	50.9	62.3	92.0
$\lambda\mathcal{L}_I + (1-\lambda)\mathcal{L}_D$	76.1	<u>59.4</u>	<u>46.8</u>	<u>68.7</u>	<u>54.4</u>	85.2	**1465.6**	<u>60.5</u>	<u>54.0</u>	64.6	<u>95.9</u>
Ours	<u>77.2</u>	**60.4**	**47.1**	**69.4**	**55.6**	<u>85.4</u>	<u>1450.2</u>	**63.8**	**56.7**	**70.1**	**98.2**

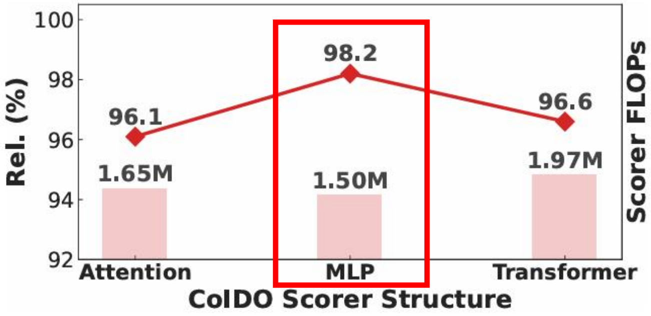


Figure 3: Ablations of different COIDO Scorer architectures.

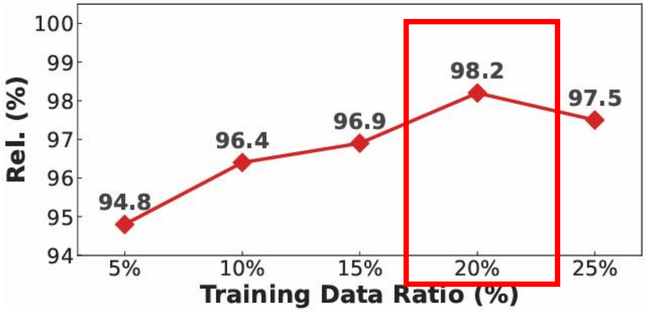


Figure 4: Comparison of different training data ratios $p\%$.

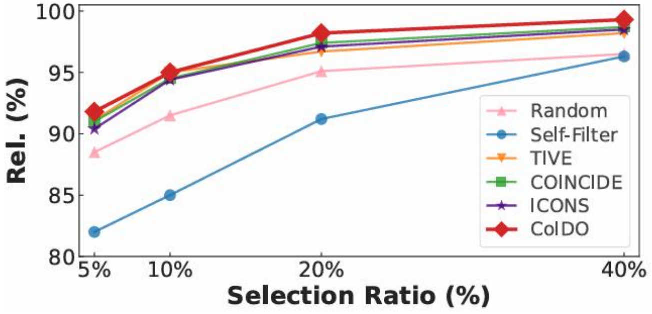


Figure 5: Performance vs. selection ratios γ.

Experiments (generalizability and transferability)

NEURAL INFORMATION PROCESSING SYSTEMS

Table 3: Performance of CoIDO on the `Vision-Flan` dataset (20% data selection). ‡ CoIDO scorer trained on `LLaVA-665K` and applied to `Vision-Flan` (out-of-domain transfer).

Model / Setting	**VQAv2**	**GQA**	**VizWiz**	**SQA**	**POPE**	**TextVQA**	**MME**	**MMBench(en)**	**MMBench(cn)**	**LLaVA-B**	***Rel.* (%)**
Full Fine-tune	74.5	47.1	52.8	61.8	46.4	85.7	1480.6	40.2	46.2	38.2	100.0
Random	74.6	44.3	50.0	59.8	40.9	81.3	1407.1	49.2	**48.3**	33.6	97.8
CoIDO	**75.7**	45.1	**53.5**	62.3	**45.3**	82.8	1452.9	**52.0**	46.8	37.6	102.1
CoIDO‡	**75.7**	**46.8**	53.3	**66.2**	42.1	**85.5**	**1486.1**	51.4	47.3	**40.8**	**103.7**

Generalizability: the ability of the proposed data selection frame work to be directly applied to other models or datasets.

Transferability: whether a CoIDO scorer trained on one domain can be reused to select informative data in another, out-of-domain corpus.

Thanks