Bootstrapping Multi-view Learning for Test-time Noisy Correspondence

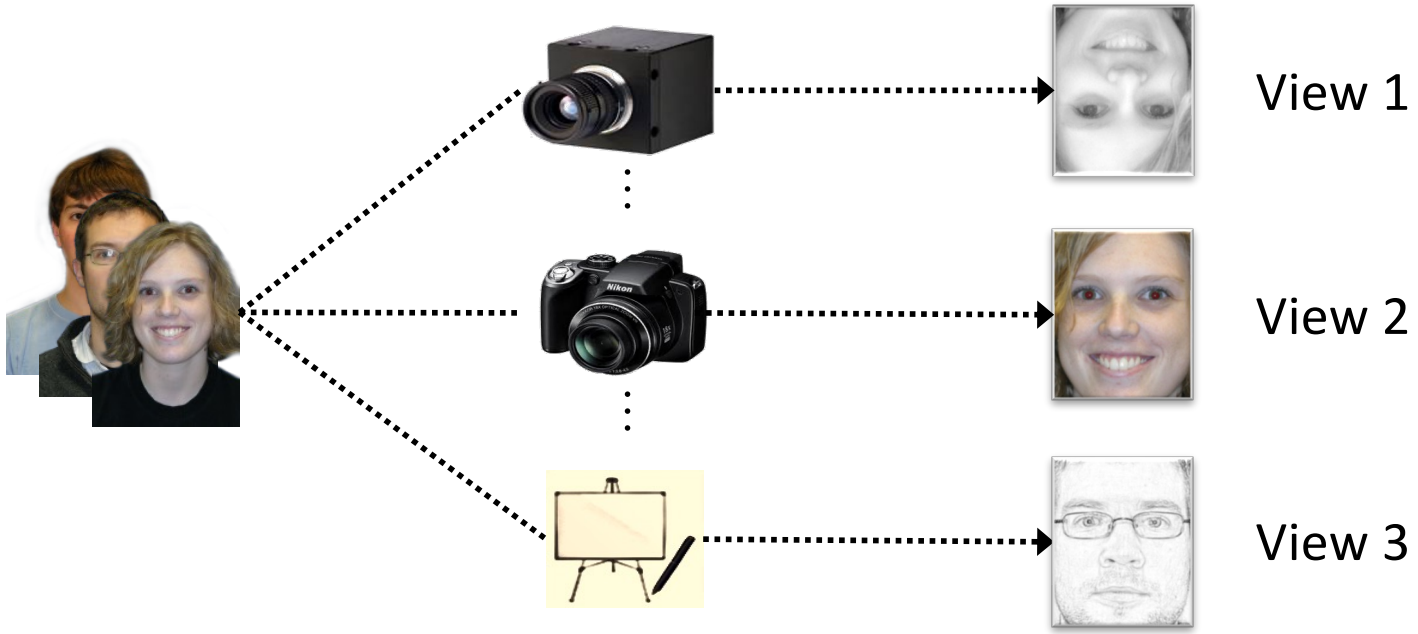
Changhao He, Di Xue, Shuxian Li, Yanji Hao, Xi Peng, Peng Hu#

Sichuan University，AVIC Chengdu Aircraft Design & Research Institute

CVPR 2026

Background

- **Data may consist of multiple modalities or views**



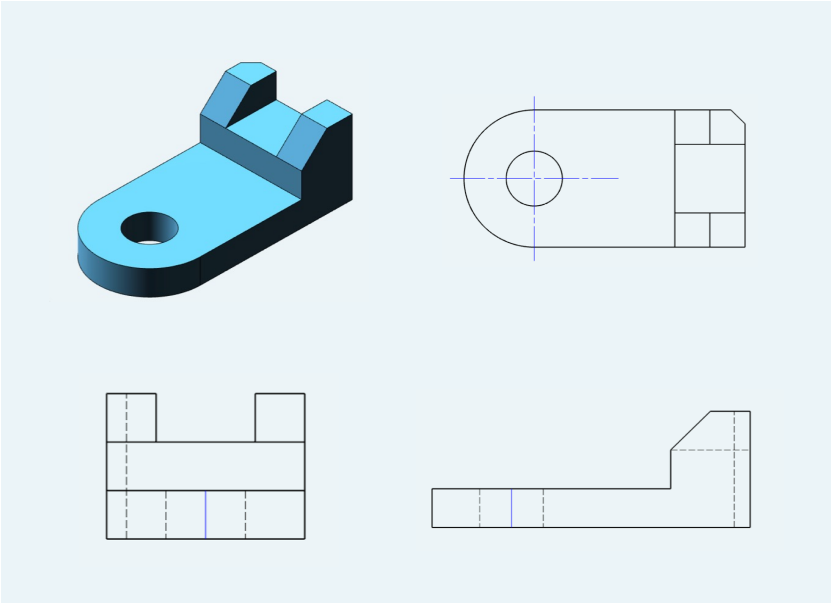
Image

A young man in a yellow short shirt is sitting and playing the guitar on the beach.

Text

Background

- **Single-view/modal methods cannot utilize the information from multiple views/modals**
  - *Examples:*



With all three-view drawing, the objects can be completely modeled

By combining video and audio, we can understand what they are talking about

Thus it is highly expected to develop the multi-view/modals analysis methods!

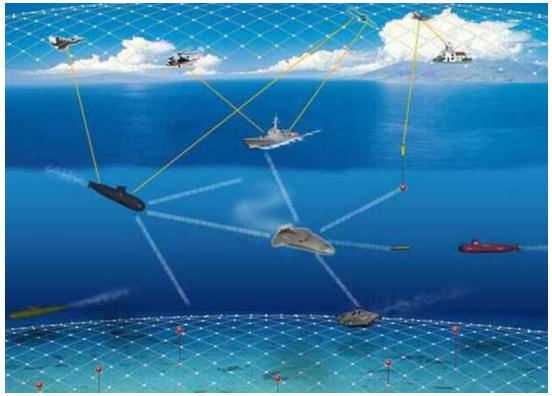
Observation

❖ **Open World: Viewpoint shifts and sensor delays lead to inevitable misalignment.**

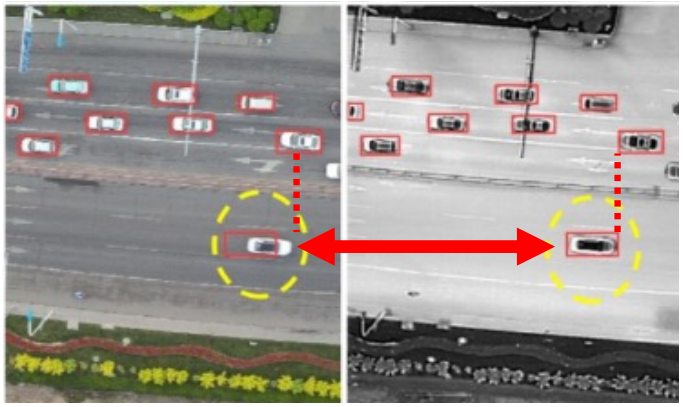
Varying Viewpoints

RGB **Infrared**

Global Shift



Sensor Asynchrony



RGB **Infrared**

Relative Shift

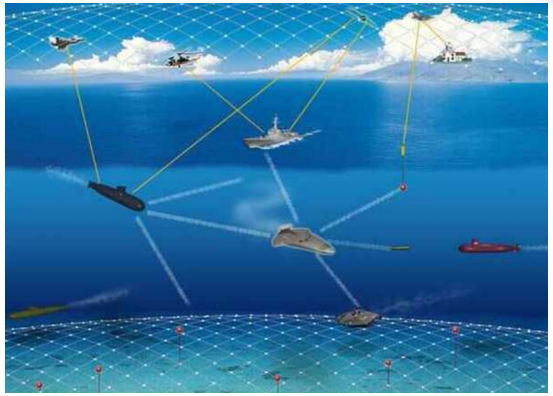
Observation

❖ **Open World: Viewpoint shifts and sensor delays lead to inevitable misalignment.**

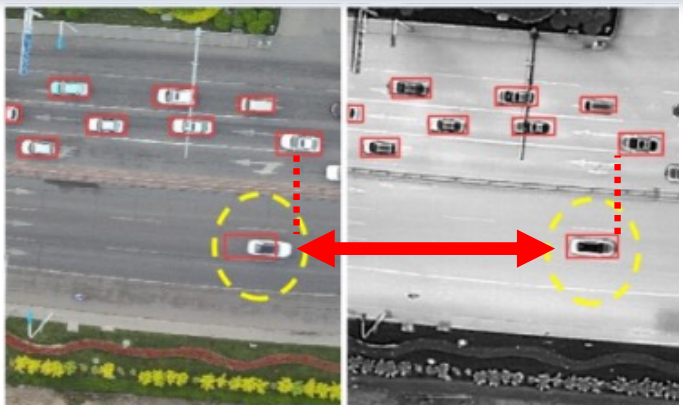
Varying Viewpoints

Global Shift

Test-time Noisy Correspondence



Sensor Asynchrony

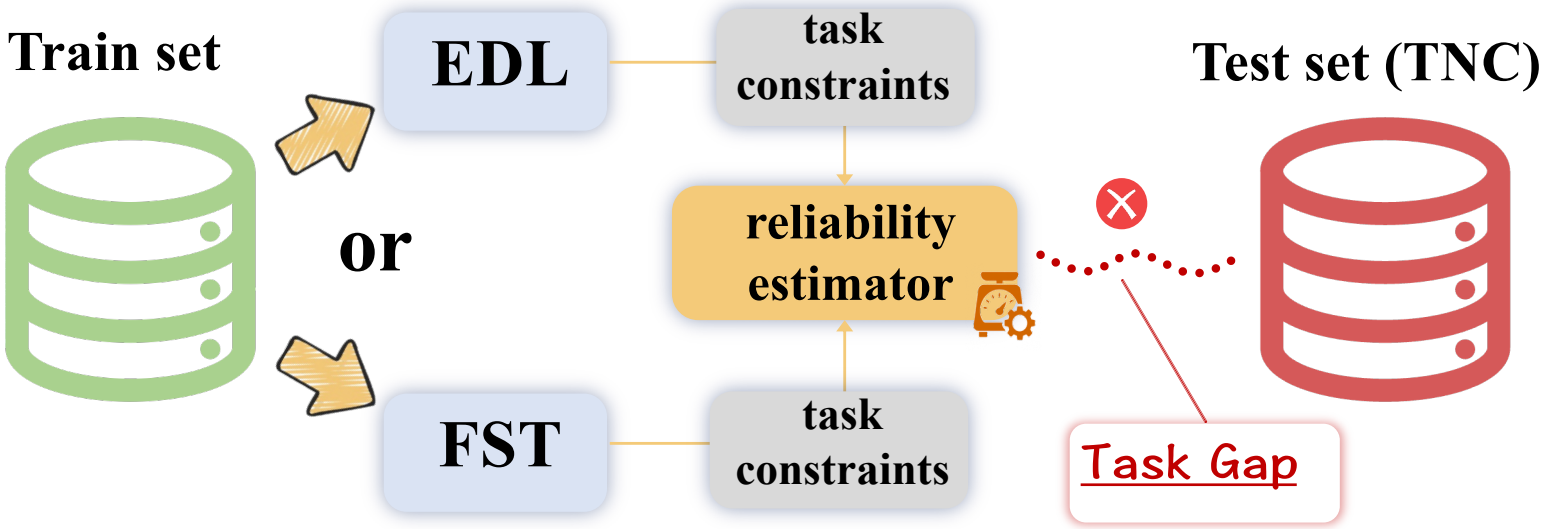


RGB **Infrared**

Relative Shift

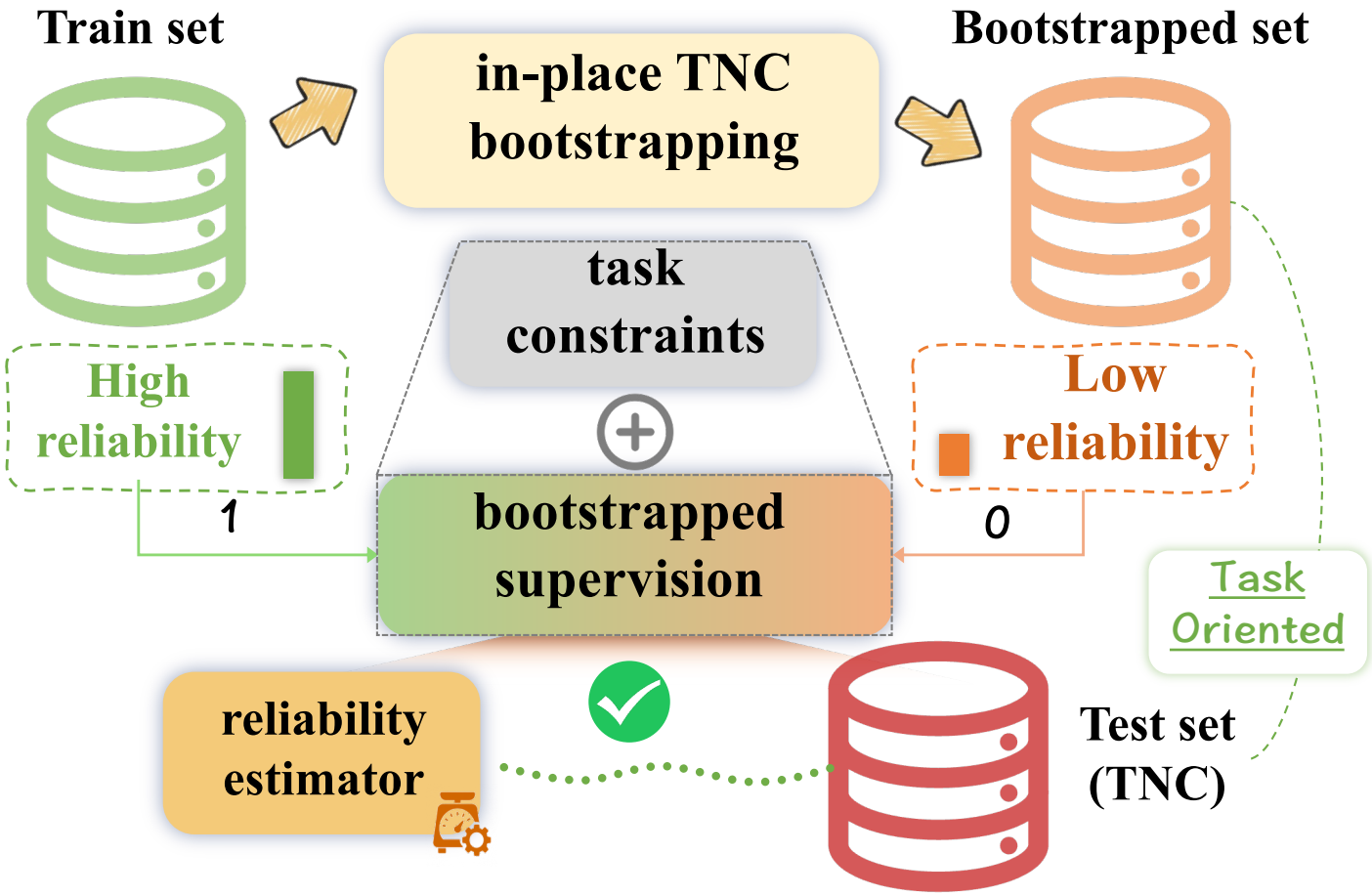
Observation & Key idea

❖ Existing works



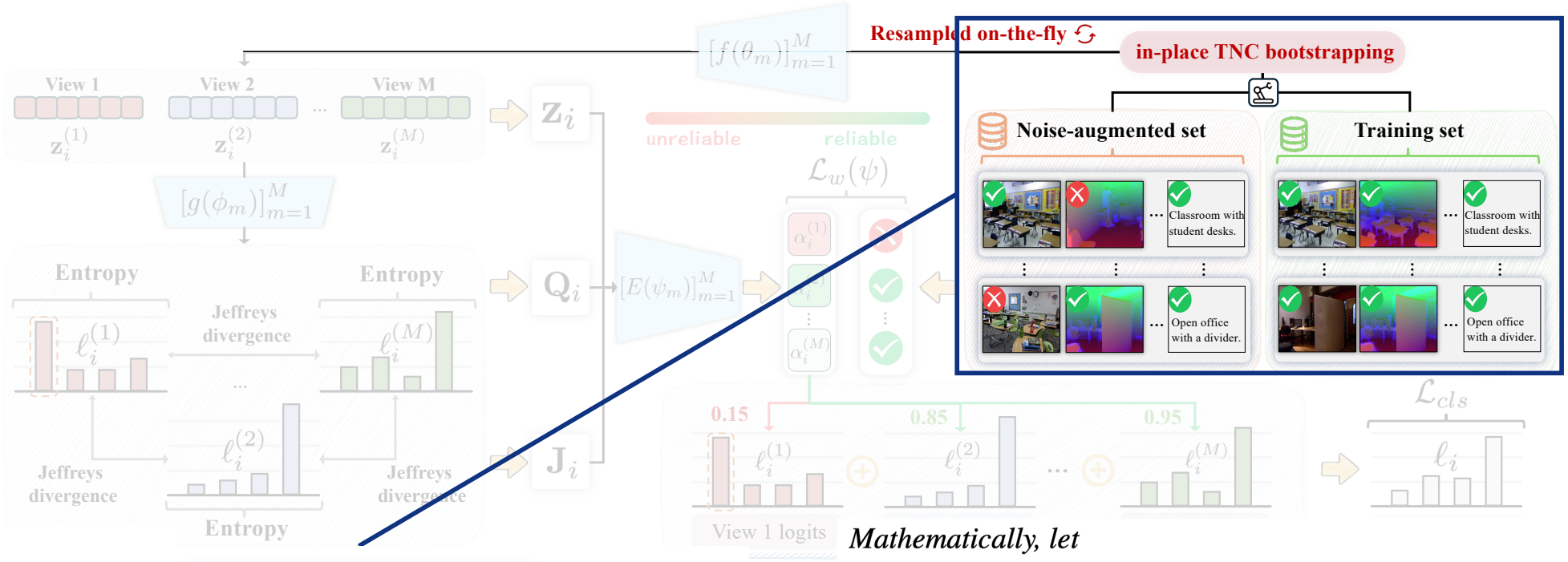
- Blind Estimation: Reliability is learned exclusively from clean, well-aligned training data.
- The Consequence: A severe Train-Test Task Gap. Models cannot extrapolate to handle unseen misalignments at deployment.

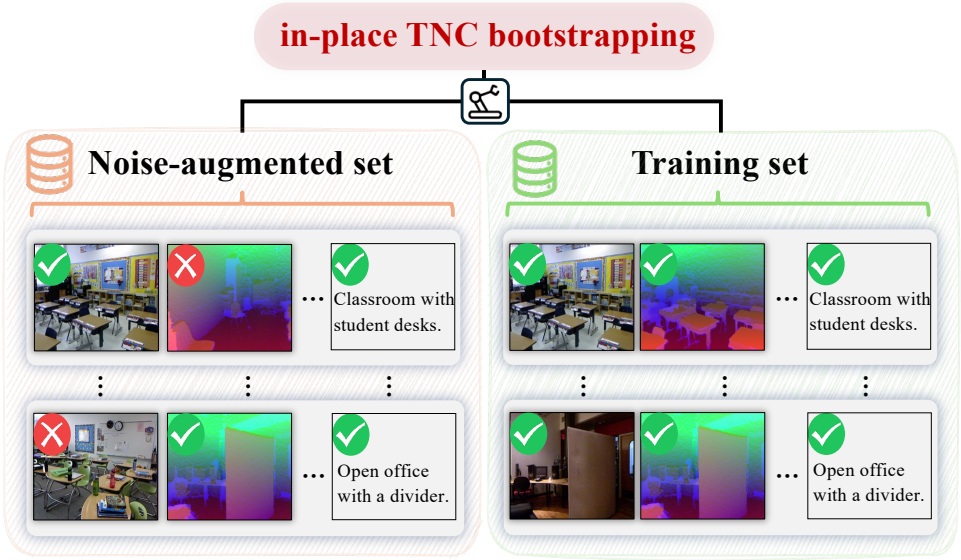
Observation & Key idea



- Data Perspective: Simulating test-time misalignments via in-place TNC bootstrapping.
- Model Perspective: Adopting a Reveal-Supervised Paradigm to provide explicit, task-oriented guidance for reliability estimator.

Method





Mathematically, let

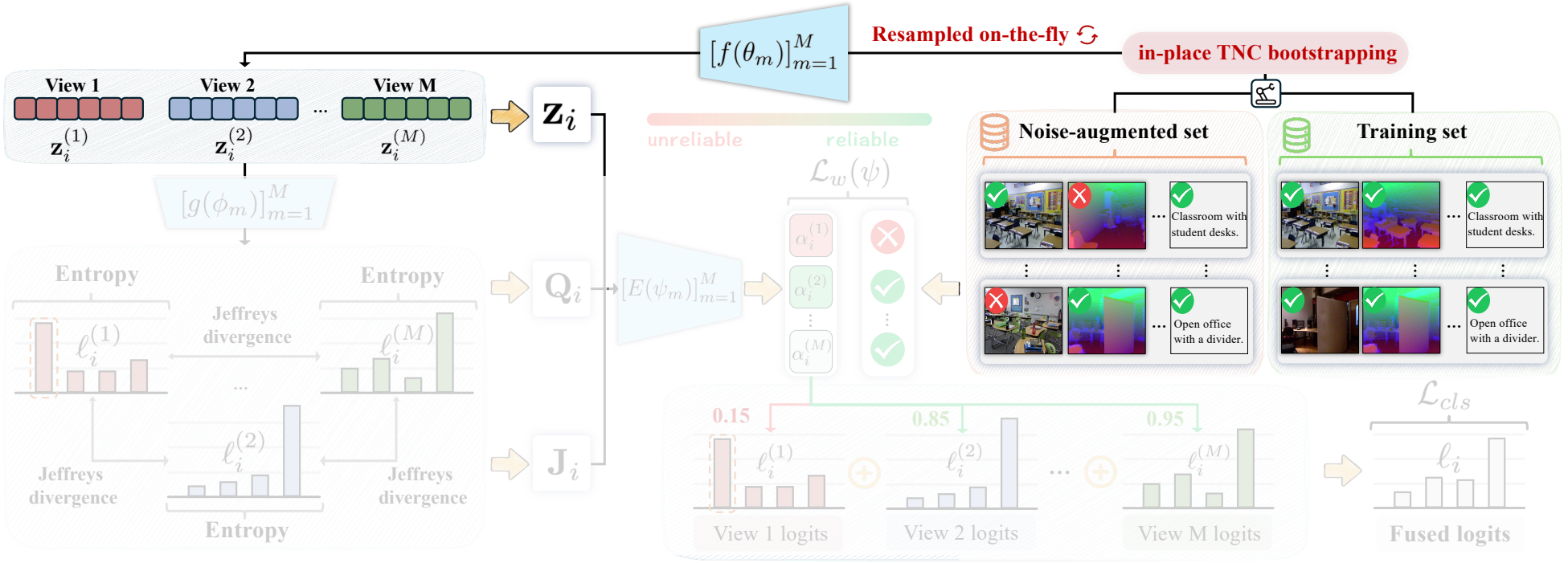
$$k_i = \sum_{m=1}^{M} \mathbb{I}\left[\mathbf{x}_i^{(m)} \not\leftrightarrow y_i\right]$$

denote the number of misaligned views, where $\mathbb{I}[\cdot]$ *is the indicator function and* $\mathbf{x}_i^{(m)} \not\leftrightarrow y_i$ *indicates that view* m *does not correspond to the ground-truth label* y_i*. We focus on the regime:*

$$1 \le k_i \le \left\lfloor \frac{M}{2} \right\rfloor,$$

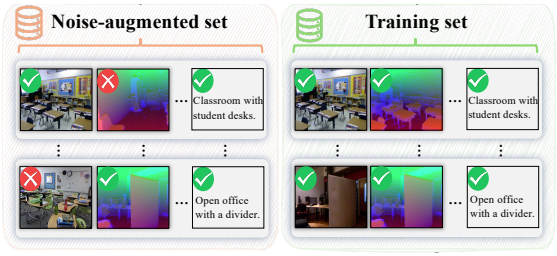
so that a majority of views remain correctly aligned.

Method

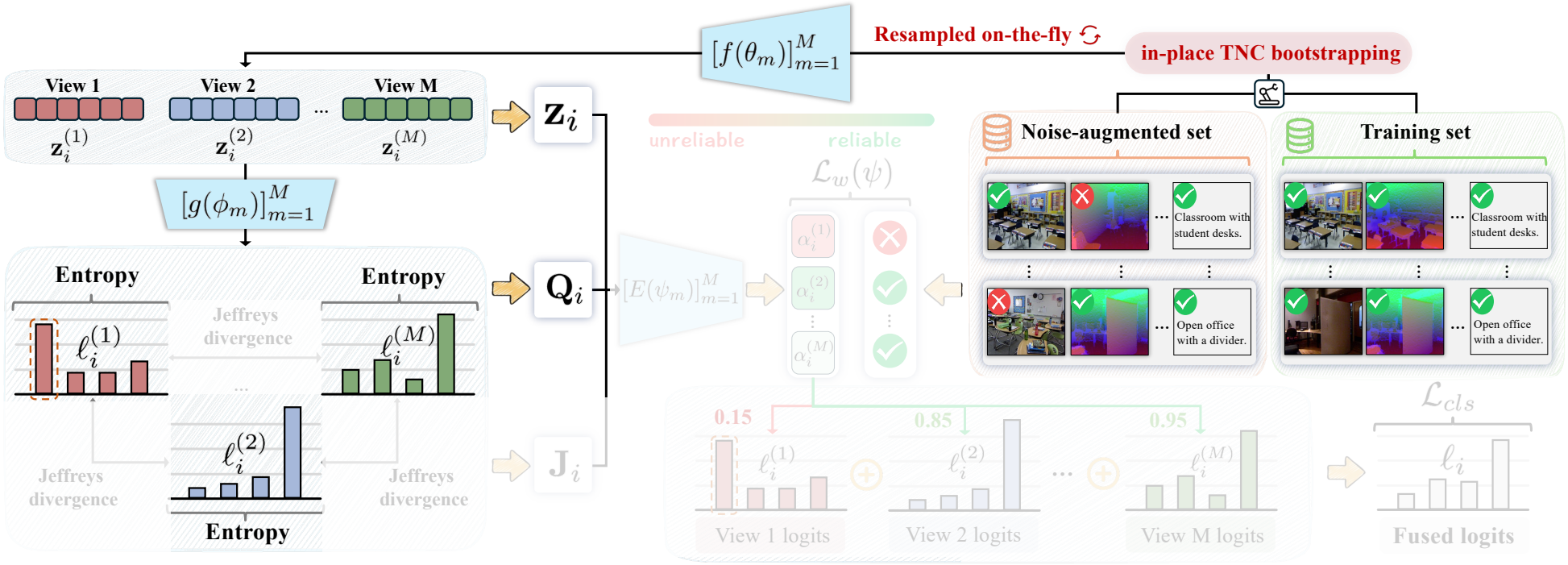


❖ Signal 1: View-specific Feature

$$\breve{\mathbf{z}}_i^{(m)} = f\left(\breve{\mathbf{x}}_i^{(m)}; \theta_m\right) \in \mathbb{R}^d$$

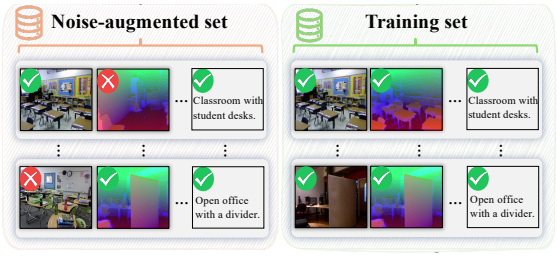


Method



❖ Signal 1: View-specific Feature

$$\breve{\mathbf{z}}_i^{(m)} = f\left(\breve{\mathbf{x}}_i^{(m)}; \theta_m\right) \in \mathbb{R}^d$$



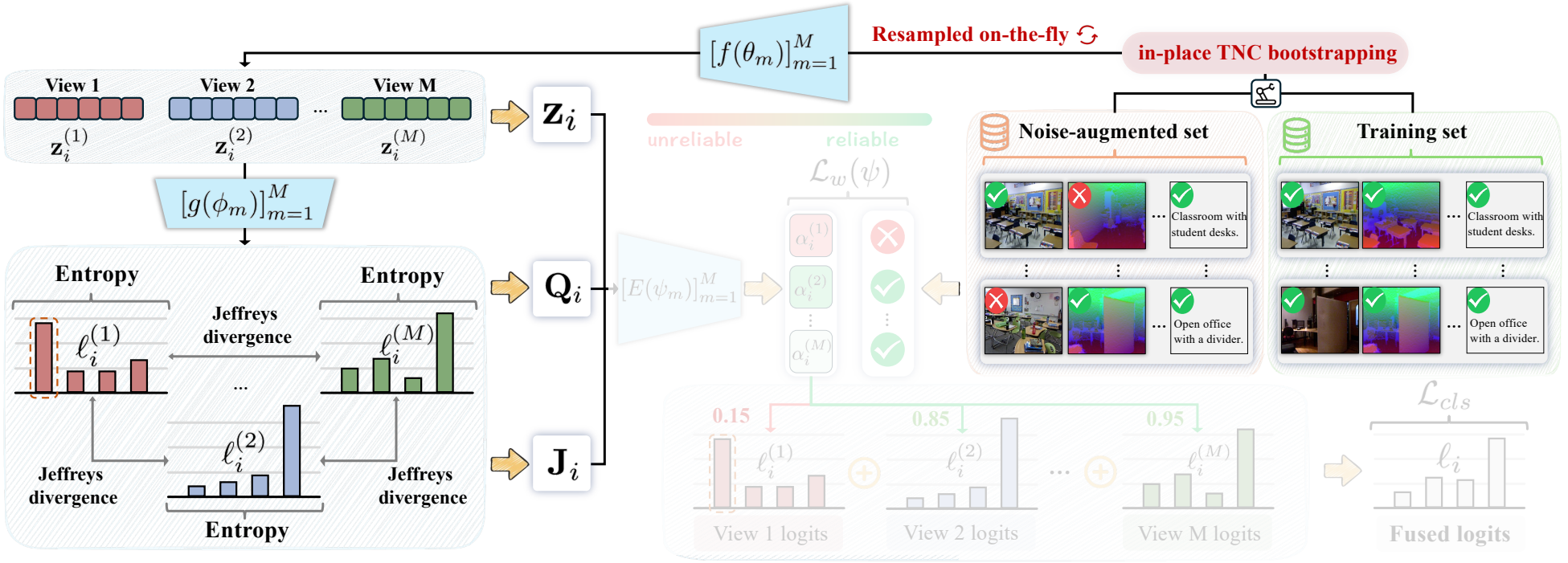
❖ Signal 2: Intra-view prediction uncertainty

Proposition 2 (Intra-view prediction uncertainty). *Given logits $\ell_i^{(m)}$ for view m, we simply use Shannon entropy to measure the view quality as follows:*

$$\begin{aligned} Q_i^{(m)} &= -\log\left[1 - \frac{H_i^{(m)}}{\log C}\right] \\ &= -\log\left[1 + \frac{\sum_{c=1}^{C} p_{i,c}^{(m)} \log p_{i,c}^{(m)}}{\log C}\right], \end{aligned} \tag{11}$$

which is small for confident predictions and large for ambiguous ones.

Method



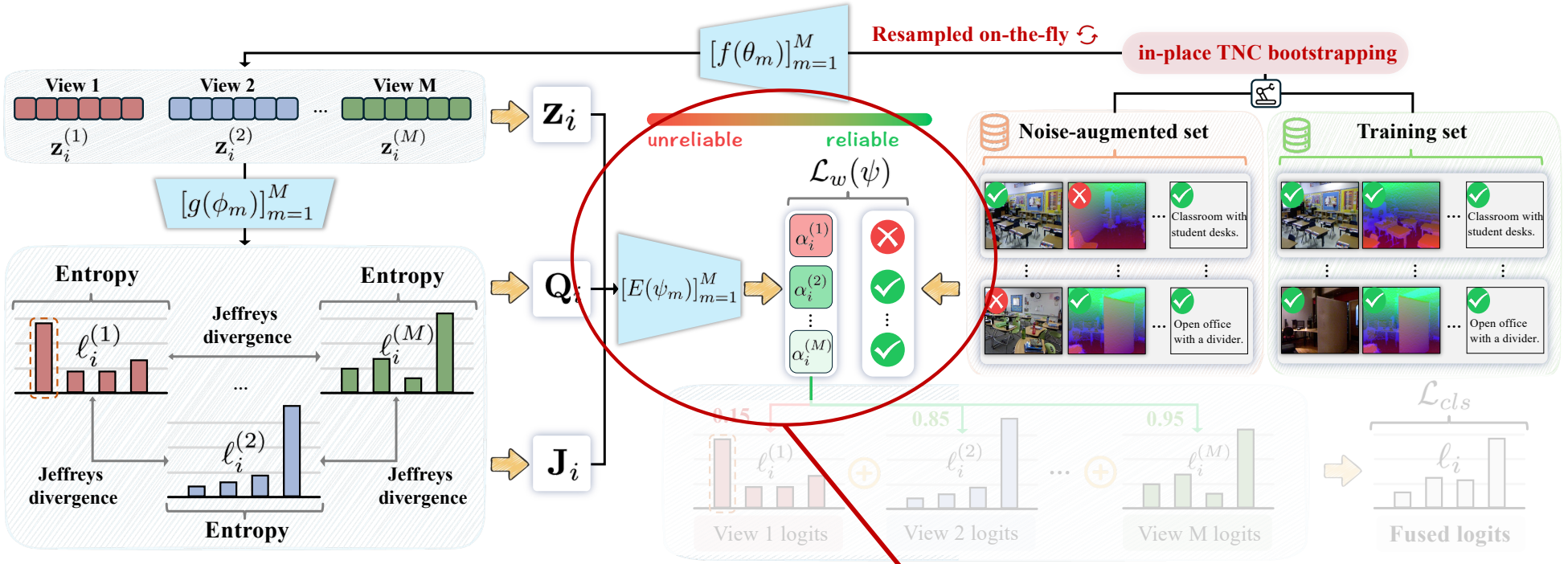
❖ Signal 3: Inter-view prediction discrepancy

Proposition 1 (Inter-view prediction discrepancy). *Given prediction distribution $\mathbf{p}_i^{(m)}$ for view m, we quantify disagreement with other views via the averaged Jeffreys divergence [25] as follows:*

$$J_i^{(m)} = \frac{1}{M-1} \sum_{n=1, n\neq m}^{M} \left[D_{KL}(\mathbf{p}_i^{(m)} \,\|\, \mathbf{p}_i^{(n)}) + D_{KL}(\mathbf{p}_i^{(n)} \,\|\, \mathbf{p}_i^{(m)}) \right], \tag{10}$$

where D_{KL} denotes the Kullback–Leibler (KL) divergence operator and a larger $J_i^{(m)}$ indicates stronger cross-view disagreement and potential mismatch.

Method



❖ Signal 3: Inter-view prediction discrepancy

Proposition 1 (Inter-view prediction discrepancy). *Given prediction distribution $\mathbf{p}_i^{(m)}$ for view m, we quantify disagreement with other views via the averaged Jeffreys divergence [25] as follows:*

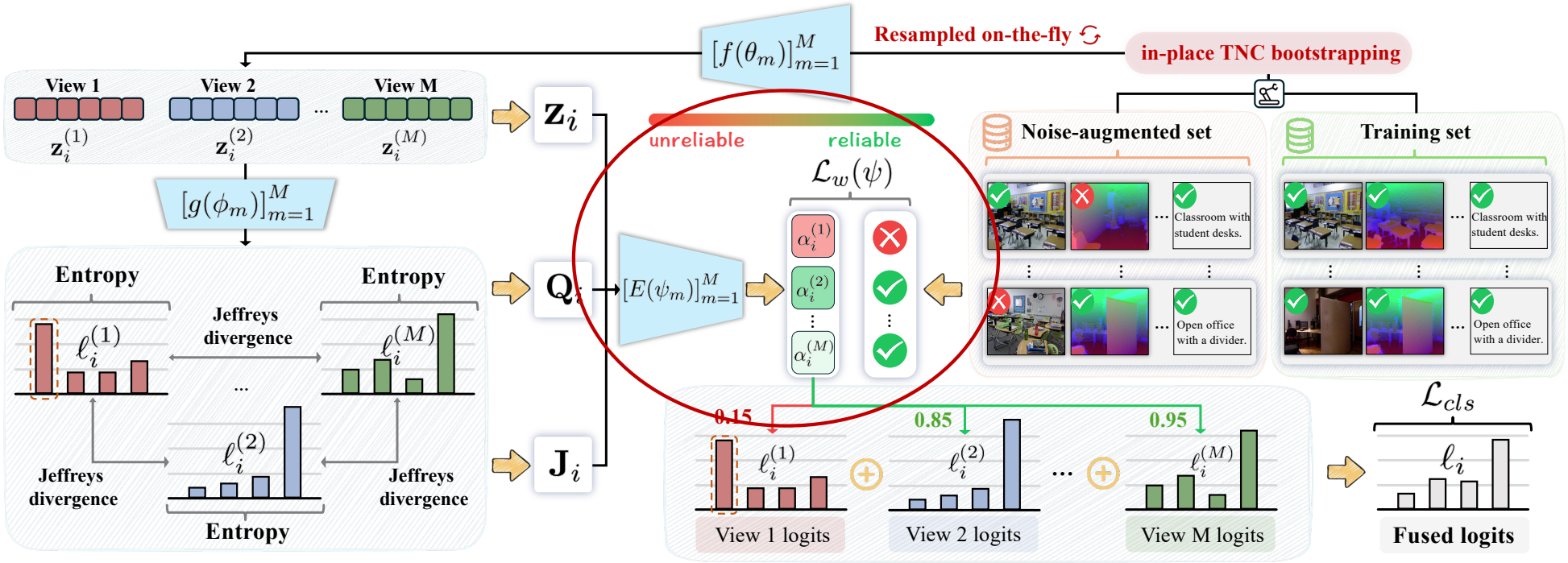
$$J_i^{(m)} = \frac{1}{M-1} \sum_{n=1, n\neq m}^{M} \left[D_{KL}(\mathbf{p}_i^{(m)} \,\|\, \mathbf{p}_i^{(n)}) + D_{KL}(\mathbf{p}_i^{(n)} \,\|\, \mathbf{p}_i^{(m)}) \right], \tag{10}$$

where D_{KL} denotes the Kullback–Leibler (KL) divergence operator and a larger $J_i^{(m)}$ indicates stronger cross-view disagreement and potential mismatch.

❖ Bootstrapped Supervision Loss

$$\mathcal{L}_w(\psi) = -\frac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} \left[\left(1 - s_i^{(m)}\right) \log \alpha_i^{(m)} + s_i^{(m)} \log \left(1 - \alpha_i^{(m)}\right) \right]$$

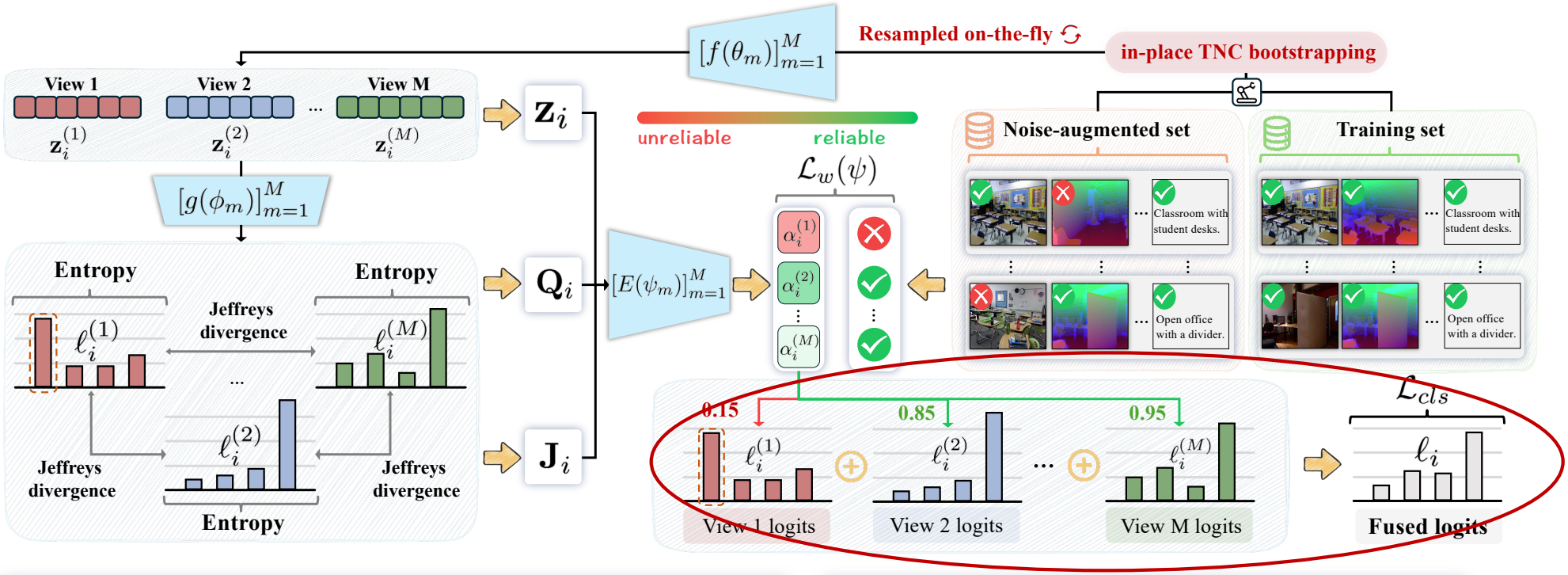
Method



❖ Bootstrapped Supervision Loss

$$\mathcal{L}_w(\psi) = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M}\left[\left(1-s_i^{(m)}\right)\log\alpha_i^{(m)} + s_i^{(m)}\log\left(1-\alpha_i^{(m)}\right)\right]$$

Method



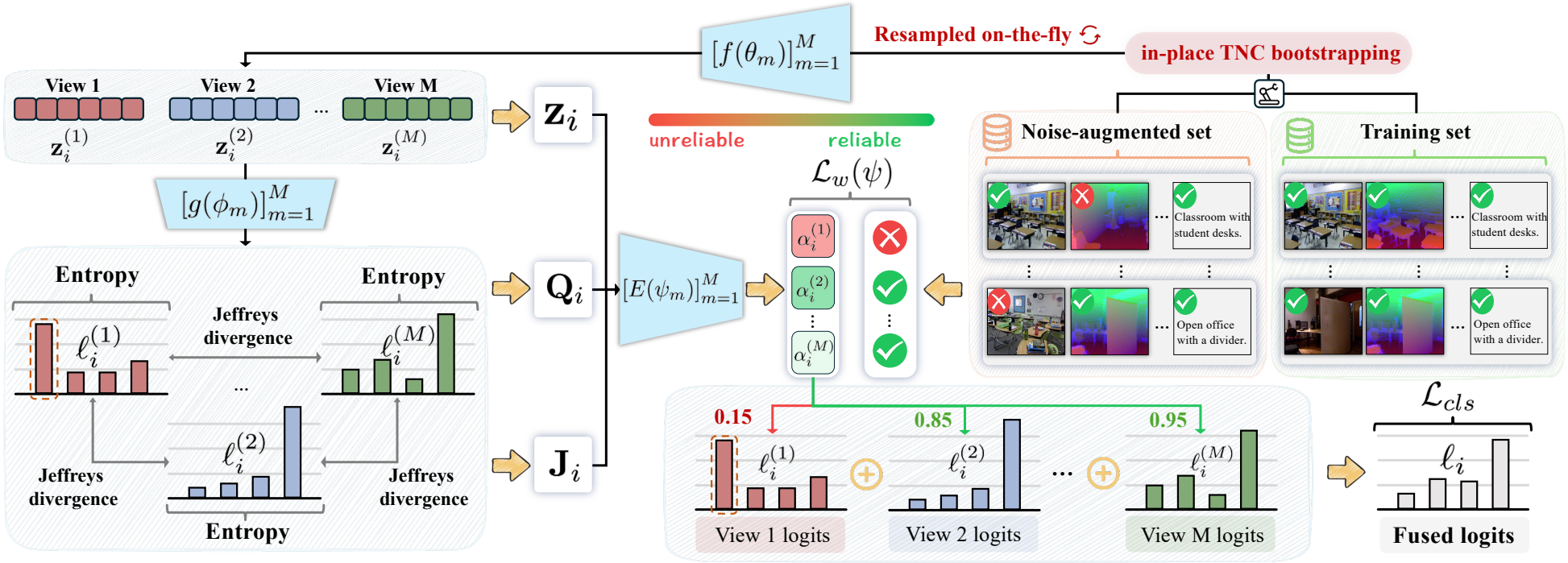
❖ Bootstrapped Supervision Loss

$$\mathcal{L}_w(\psi) = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M}$$
$$\left[\left(1-s_i^{(m)}\right)\log\alpha_i^{(m)} + s_i^{(m)}\log\left(1-\alpha_i^{(m)}\right)\right]$$

❖ Task Loss

$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\hat{p}_{i,y_i}$$

Method



❖ Bootstrapped Supervision Loss

$$\mathcal{L}_w(\psi) = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M}\left[\left(1-s_i^{(m)}\right)\log\alpha_i^{(m)} + s_i^{(m)}\log\left(1-\alpha_i^{(m)}\right)\right]$$

❖ Task Loss

$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\hat{p}_{i,y_i}$$

❖ Overall Loss

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\,\mathcal{L}_w$$

Experiment

❖ Classification Performance under Test-time Noisy Correspondence

Noise	Method	Caltech	Leaves	HW	LandUse	Scene	AVG.	
0%	TMC [ICLR'21]	95.96±1.12	95.59±0.85	97.55±0.56	46.14±1.82	72.47±0.91	81.54	
	UIMC [CVPR'23]	97.29±0.61	93.94±1.50	98.10±0.56	49.90±1.66	73.61±1.15	82.57	
	ECML [AAAI'24]	96.25±0.79	93.47±1.30	97.22±0.56	44.64±1.85	70.45±1.00	80.41	
	CCML [ACM MM'24]	95.79±1.09	96.81±0.69	97.38±0.81	41.38±2.11	71.56±1.39	80.58	
	MAMC [ICLR'25]	97.82±0.72	89.88±1.55	98.90±0.58	71.67±1.60	80.71±1.28	87.80	
	ETF [ICML'25]	96.47±0.91	98.44±0.31	97.40±0.60	46.09±1.71	75.27±0.99	82.73	
	TMCEK [ICML'25]	96.18±0.83	93.53±1.28	97.30±0.63	45.10±2.29	71.04±1.22	80.63	
	FUML [ICML'25]	95.29±0.93	99.28±0.63	98.45±0.81	74.95±1.71	78.92±1.68	89.38	
	RML [ICCV'25]	95.54±1.58	97.03±1.06	98.20±0.82	75.40±2.43	78.06±1.13	88.85	
	BML	**98.11±0.75**	**99.75±0.27**	**99.03±0.59**	**81.43±1.06**	**83.92±1.39**	**92.45**	↑ 3.07
50%	TMC [ICLR'21]	87.07±1.08	79.69±1.63	83.42±1.34	39.62±2.07	61.83±1.54	70.33	
	UIMC [CVPR'23]	92.96±1.21	78.09±1.45	88.75±1.29	42.43±1.18	63.55±1.27	73.16	
	ECML [AAAI'24]	83.25±2.16	77.69±1.94	81.20±1.48	38.69±1.92	60.33±1.65	68.23	
	CCML [ACM MM'24]	82.14±2.23	78.88±1.54	79.12±1.38	34.88±1.97	60.80±1.15	67.16	
	MAMC [ICLR'25]	96.36±0.75	73.69±2.30	96.47±0.90	62.93±1.19	74.06±1.49	80.70	
	ETF [ICML'25]	86.46±2.72	78.72±1.58	81.57±1.63	38.98±1.51	63.23±1.29	69.79	
	TMCEK [ICML'25]	83.54±1.92	77.62±2.17	79.53±1.93	38.86±2.48	60.70±1.60	68.05	
	FUML [ICML'25]	85.46±5.15	94.25±1.31	95.35±1.52	68.29±2.34	74.10±1.35	83.49	
	RML [ICCV'25]	86.00±1.84	86.16±1.02	87.97±1.26	66.26±2.46	69.25±1.16	79.13	
	BML	**96.79±0.83**	**96.16±0.84**	**97.28±0.66**	**75.62±1.29**	**79.25±1.53**	**89.02**	↑ 5.53
100%	TMC [ICLR'21]	77.43±3.50	61.78±3.10	67.35±2.08	33.50±1.51	51.71±1.11	58.35	
	UIMC [CVPR'23]	88.71±1.47	60.78±3.49	78.05±1.55	34.90±1.64	53.60±1.35	63.21	
	ECML [AAAI'24]	68.11±2.23	59.91±3.60	64.28±1.73	32.48±1.92	51.01±0.95	55.16	
	CCML [ACM MM'24]	67.71±2.47	60.09±3.21	59.45±2.10	29.52±1.89	50.08±1.08	53.37	
	MAMC [ICLR'25]	95.21±1.50	58.81±2.57	93.28±0.49	53.33±2.35	67.92±1.30	73.71	
	ETF [ICML'25]	76.14±4.71	59.53±2.27	64.33±1.92	31.52±1.05	51.80±0.90	56.67	
	TMCEK [ICML'25]	69.11±2.78	59.91±3.71	61.80±2.26	33.02±1.50	51.65±0.84	55.10	
	FUML [ICML'25]	75.79±9.43	89.88±1.93	92.10±3.13	63.74±2.49	69.05±1.22	78.11	
	RML [ICCV'25]	75.29±3.36	76.81±2.89	77.05±1.06	55.43±1.99	60.45±1.01	69.01	
	BML	**95.36±1.13**	**93.56±0.73**	**94.97±1.02**	**69.02±1.50**	**74.04±1.10**	**85.39**	↑ 7.28

Consistent robustness across noise ratios !

Experiment

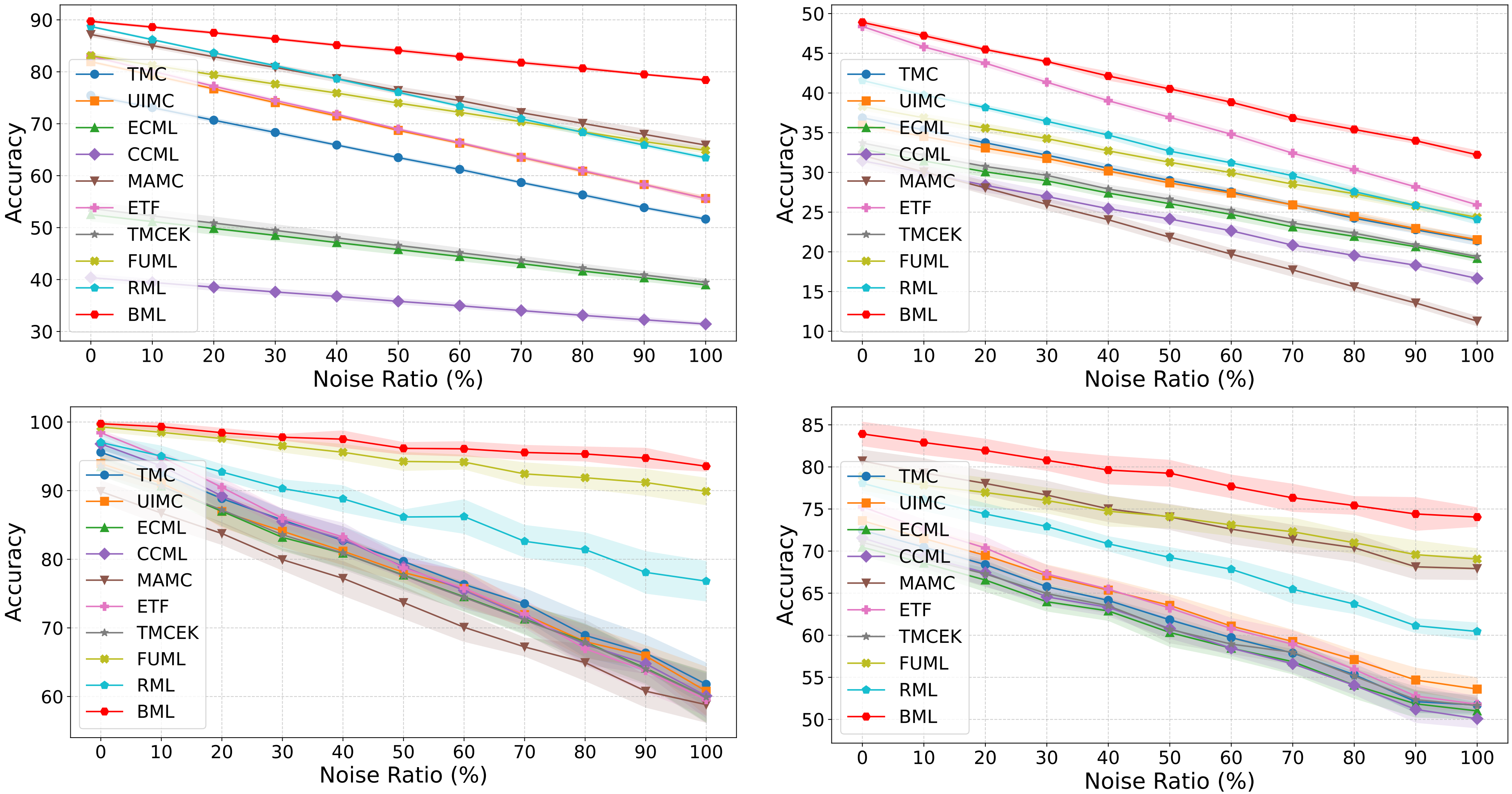
❖ Classification Performance under Test-time Noisy Correspondence

Noise	Method	CCV	Fashion	NUS-OBJ	AWA	YouTubeFace	AVG.	
0%	TMC [ICLR'21]	44.58±0.81	96.07±0.31	39.72±0.51	36.87±0.44	75.44±0.36	58.54	
	UIMC [CVPR'23]	45.48±1.29	98.25±0.22	41.27±0.68	35.95±0.42	81.95±0.24	60.58	
	ECML [AAAI'24]	42.59±1.09	95.43±0.38	35.34±0.49	32.93±0.56	52.52±1.22	51.76	
	CCML [ACM MM'24]	43.62±0.67	95.66±0.52	38.04±0.58	31.46±0.92	40.36±0.63	49.83	
	MAMC [ICLR'25]	54.66±1.26	98.53±0.28	45.84±0.57	32.15±0.83	87.18±0.39	63.67	
	ETF [ICML'25]	48.40±0.79	96.78±0.26	45.98±0.41	48.38±0.35	83.02±0.25	64.51	
	TMCEK [ICML'25]	42.90±0.97	94.53±0.33	35.40±0.59	33.70±0.47	53.69±1.26	52.04	
	FUML [ICML'25]	47.38±2.43	98.09±0.33	47.14±0.56	38.30±0.43	83.10±0.43	62.80	
	RML [ICCV'25]	47.50±1.03	98.66±0.21	44.00±0.52	41.58±0.38	88.72±0.13	64.09	
	BML	**59.56±0.99**	**98.85±0.16**	**51.91±0.40**	**48.90±0.27**	**89.72±0.22**	**69.79**	↑ 5.28
50%	TMC [ICLR'21]	38.92±0.79	84.11±0.61	34.51±0.29	28.98±0.42	63.48±0.28	50.00	
	UIMC [CVPR'23]	39.68±1.02	91.43±0.49	35.62±0.71	28.68±0.53	68.72±0.24	52.83	
	ECML [AAAI'24]	37.62±1.00	82.34±0.50	31.26±0.55	26.06±0.51	45.79±0.91	44.61	
	CCML [ACM MM'24]	37.23±1.05	83.07±0.52	32.45±0.56	24.14±0.68	35.81±0.51	42.54	
	MAMC [ICLR'25]	46.87±1.10	95.66±0.75	38.58±0.58	21.84±0.70	76.39±0.69	55.87	
	ETF [ICML'25]	41.67±1.04	88.28±0.45	38.26±0.45	36.94±0.36	68.97±0.24	54.82	
	TMCEK [ICML'25]	37.70±0.72	82.03±0.32	31.27±0.59	26.63±0.56	46.57±0.98	44.84	
	FUML [ICML'25]	41.97±2.42	95.25±0.48	42.72±0.33	31.28±0.42	73.98±0.50	57.04	
	RML [ICCV'25]	41.47±1.23	90.87±0.69	37.48±0.59	32.68±0.58	76.06±0.17	55.71	
	BML	**51.73±1.10**	**96.24±0.29**	**45.99±0.44**	**40.52±0.31**	**84.12±0.32**	**63.72**	↑ 6.68
100%	TMC [ICLR'21]	33.12±1.42	72.00±0.94	29.56±0.45	21.39±0.55	51.64±0.33	41.54	
	UIMC [CVPR'23]	34.04±1.22	84.99±0.85	29.86±0.47	21.53±0.27	55.61±0.41	45.21	
	ECML [AAAI'24]	31.99±0.94	68.60±0.90	27.44±0.41	19.17±0.41	38.99±0.63	37.24	
	CCML [ACM MM'24]	30.67±1.08	69.91±1.05	27.23±0.36	16.65±0.67	31.43±0.33	35.18	
	MAMC [ICLR'25]	38.09±1.54	93.47±0.51	31.28±0.48	11.28±0.59	65.89±1.02	48.00	
	ETF [ICML'25]	33.30±1.45	80.55±0.59	31.56±0.29	25.92±0.51	55.57±0.45	45.38	
	TMCEK [ICML'25]	32.18±0.88	69.21±1.06	27.40±0.39	19.40±0.33	39.46±0.65	37.53	
	FUML [ICML'25]	36.66±2.29	92.75±0.52	37.96±0.52	24.34±0.41	64.87±0.50	51.32	
	RML [ICCV'25]	35.09±1.04	83.31±1.71	31.17±0.48	24.08±0.59	63.45±0.34	47.42	
	BML	**44.16±0.97**	**94.37±0.35**	**39.98±0.53**	**32.22±0.51**	**78.42±0.26**	**57.83**	↑ 6.51

Consistent robustness across noise ratios !

Experiment

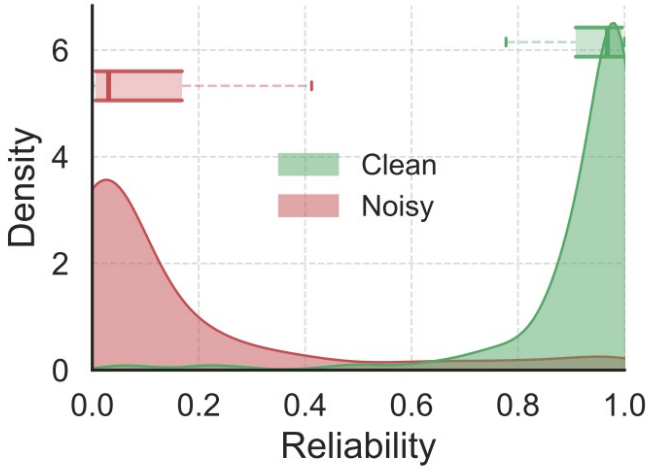
❖ Classification Performance under varying noise levels



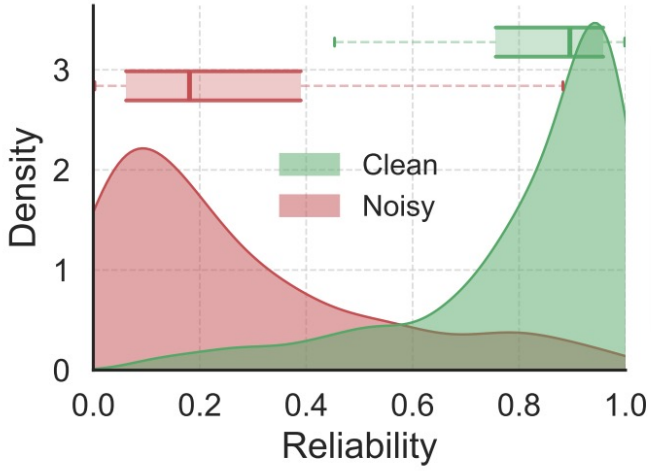
Performance advantage becomes more pronounced as TNC noise increases

Experiment

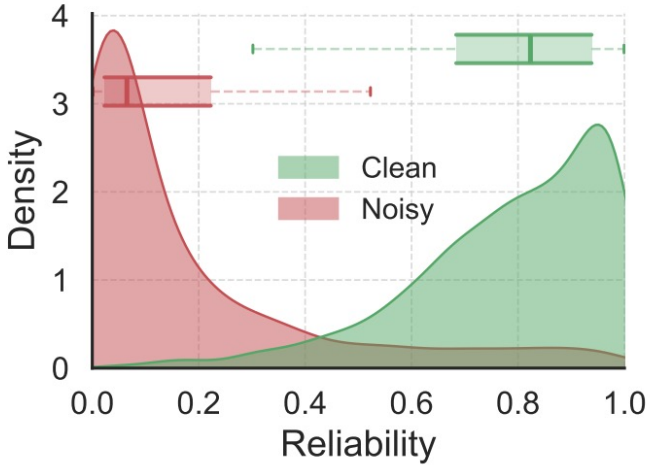
❖ **Density plots and box plots of the estimated view reliability**



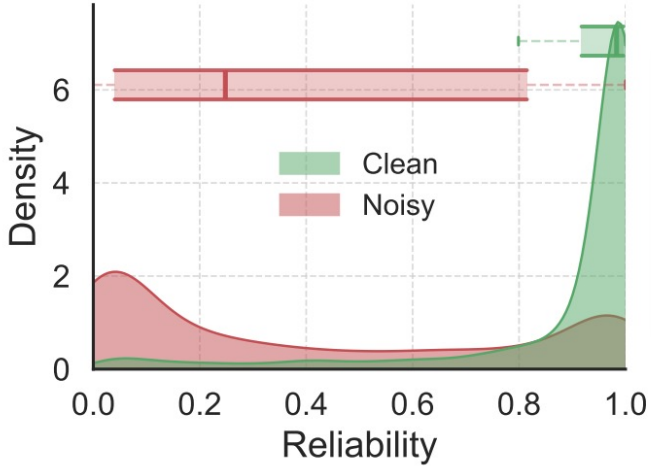
(a) Caltech



(b) Leaves



(c) HW



(d) SUN R-D-T

Clean views:
Accurately assigned high reliability (~1.0)

Noisy views:
Effectively suppressed with low reliability (~0.0).

Experiment

❖ t-SNE visualization comparison W/ & W/O view reliability

(a) W/O $\alpha_i^{(m)}$ $(Acc = 90.75)$ (b) W/ $\alpha_i^{(m)}$ $(Acc = 94.15)$

Reliability weighting effectively clusters aligned views and separates misaligned ones.

Experiment

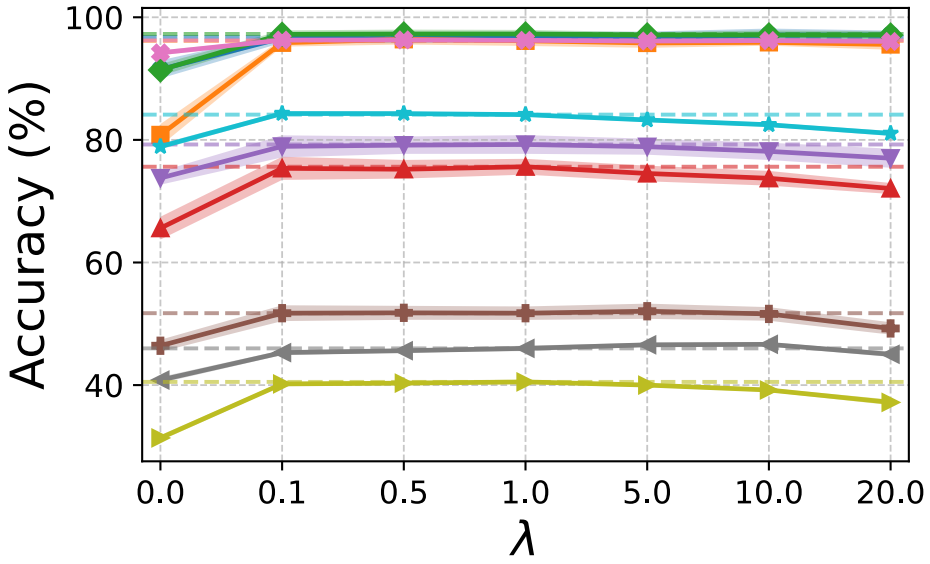
❖ Ablation results at various noise ratios

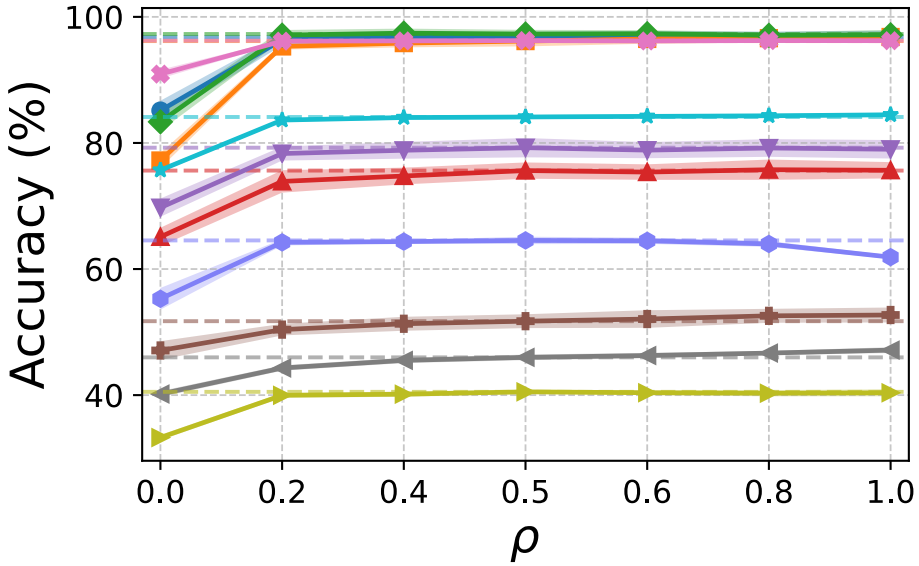
Noise	W/O $\mathcal{L}_w$	W/O J	W/O Q	W/O on-the-fly	FULL
0%	64.51±0.48	67.02±0.50	67.57±0.27	63.30±0.48	**68.15±0.28**
10%	63.28±0.74	66.01±0.53	66.79±0.28	61.97±0.54	**67.43±0.34**
20%	62.15±0.49	65.03±0.57	66.18±0.25	60.44±0.54	**66.68±0.42**
30%	60.72±0.69	63.82±0.30	65.39±0.29	58.79±0.42	**66.07±0.46**
40%	59.65±0.42	62.72±0.58	64.53±0.31	57.63±0.36	**65.11±0.25**
50%	58.50±0.93	61.76±0.22	64.20±0.55	56.27±0.93	**64.54±0.59**
60%	56.93±0.63	60.46±0.44	63.10±0.51	54.90±0.34	**63.83±0.11**
70%	55.14±0.78	59.14±0.18	62.26±0.31	53.00±0.69	**62.80±0.45**
80%	54.43±0.97	58.22±0.77	61.78±0.47	51.75±0.58	**62.12±0.61**
90%	53.17±0.72	57.45±0.42	60.96±0.54	50.25±0.58	**62.05±0.45**
100%	52.07±1.00	56.80±0.46	60.78±0.61	48.95±0.94	**60.97±0.60**

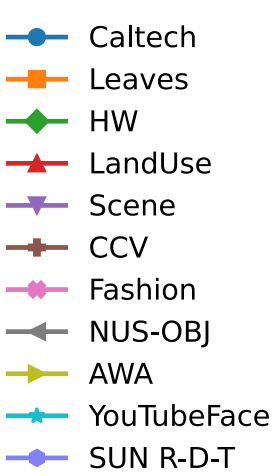
Dynamic resampling and discrepancy modeling are the core drivers of BML's robustness.

Experiment

❖ **Parameter Analysis of the loss coefficient and noise-augmented proportion**

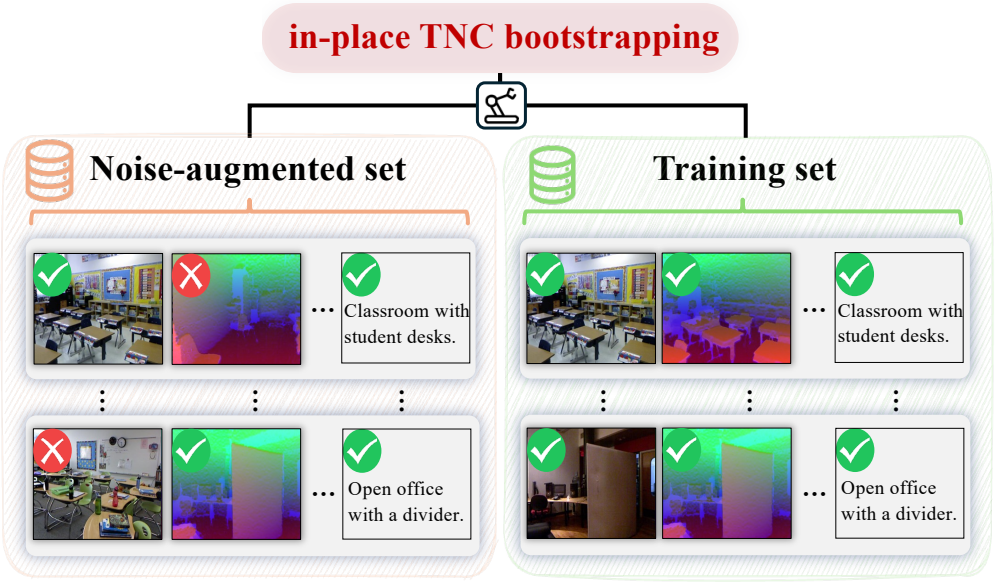






$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_w$$

strong stability across a wide range of hyperparameters



Experiment

❖ SUN R-D-T Benchmark: Structured Prompt for Text Generation

You are generating strictly content-focused image descriptions for research on multimodal classification.

TASK
Describe only visible objects, attributes, spatial relations (left/right/near/behind/under), counts, and human actions if plainly observable. Base every token on visual evidence.

STRICT PROHIBITIONS (to prevent label leakage)
Do NOT name, imply, or hint at any place or scene type. The following words/phrases and their plurals, synonyms, or morphological variants are FORBIDDEN and must not appear: bathroom, bedroom, classroom, computer room, conference room, corridor, dining area, dining room, discussion area, furniture store, home office, kitchen, lab, laboratory, lecture theatre, lecture theater, library, living room, office, rest space, study space, interior, exterior, indoors, outdoors.

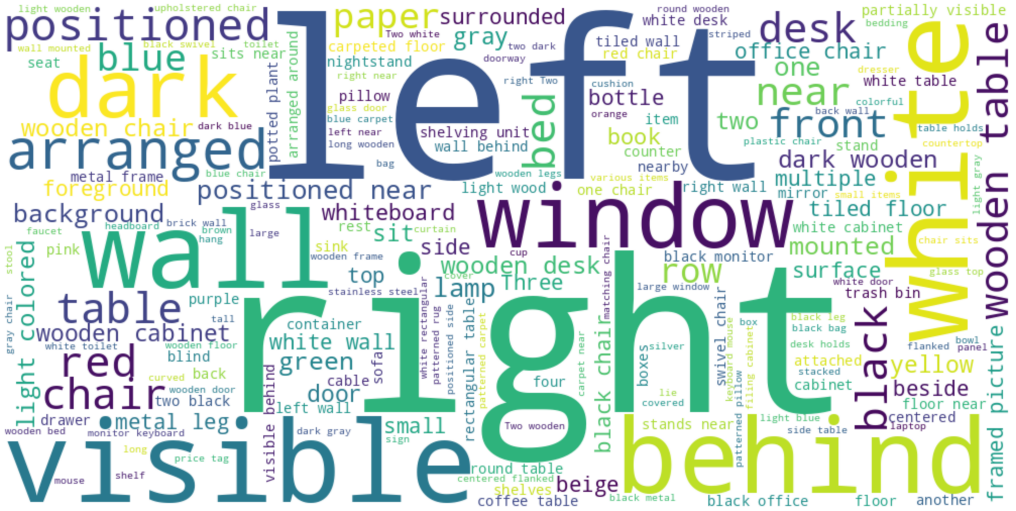
STYLE AND LENGTH
1. No scene/type labels, no brand/model guesses, no value judgments, no speculation.
2. One sentence in English, less than 20 words.

UNCERTAINTY POLICY
If a detail is unclear, omit it rather than guessing.

OUTPUT
Return only the single sentence (no prefixes, no metadata).

Experiment

❖ Word Cloud Visualization



High-frequency terms are strictly spatial relations (e.g., left, right, behind) and object nouns.

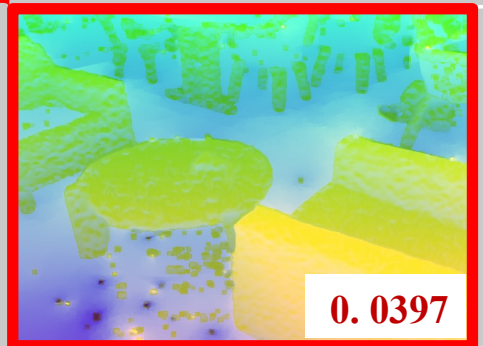
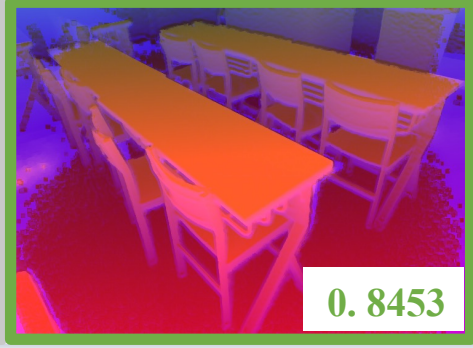
Contains 2,664 unique words despite concise caption lengths.

❖ Length Statistics & Vocabulary

Metric	chars	words
Avg text length	188.90	34.64
Median text length	179.00	33.00
Std dev of length	45.11	8.58
Vocabulary size	-	2664.00

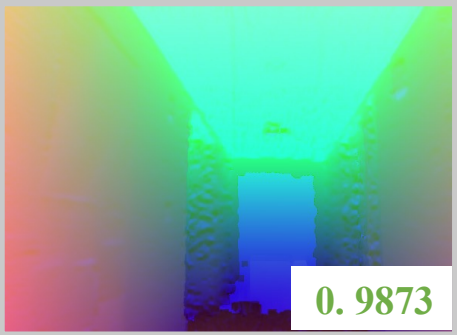
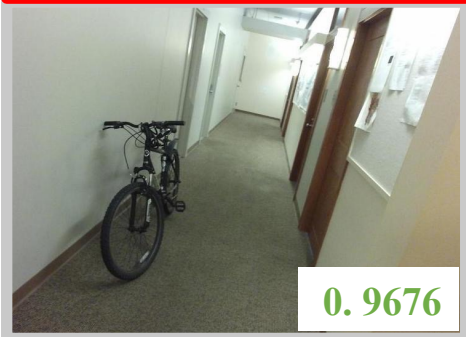
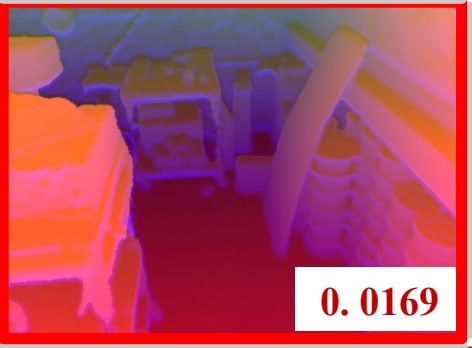
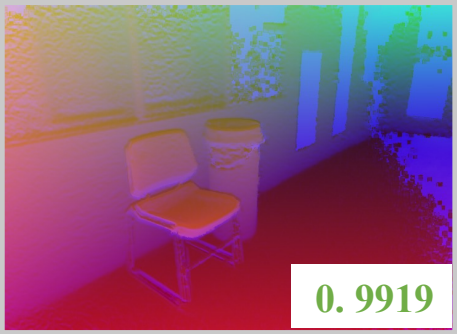
Experiment

❖ **Qualitative Case Study: Reliability Estimation under TNC *(Classroom)***

RGB view	Depth view	Text view
 0. 0802	 0. 9970	**Six white rectangular tables with gray legs are arranged in two rows; each table has a maroon seat with black frame; white cables lie on the gray patterned floor; two whiteboards are on the back wall.** 0. 9941
 0. 9962	 0. 0397	Two purple chairs with metal legs are positioned against a white wall with brown baseboard; to the left, three light blue chairs are arranged near a dark table, with a closed door visible in the background. 0. 9911
 0. 9181	 0. 8453	Three cardboard boxes sit atop a white desk with a light blue drawer unit; a gray swivel chair with a green backrest is positioned in front, and a beige filing cabinet is to the right. 0. 0015

Experiment

❖ **Qualitative Case Study: Reliability Estimation under TNC *(Corridor)***

RGB view	Depth view	Text view
 0. 0033	 0. 9873	A long hallway with red walls on the left and white walls on the right, lined with papers, signs, and doors; ceiling lights illuminate the shiny floor, and an exit sign is visible ahead. 0. 9606
 0. 9676	 0. 0169	A black bicycle with front suspension is parked against the left wall in a narrow space with beige carpet, white walls, and multiple closed doors on the right side. 0. 9916
 0. 9687	 0. 9919	A black mesh office chair sits on blue carpet near a gray filing cabinet, with a desk holding papers, a monitor, and a telephone to the right; shelves with binders and books are mounted on the wall behind. 0. 0116

Thanks for your attention!

College of Computer Science
Sichuan University
XLearning Lab

Code → **XLearning-SCU/2026-CVPR-BML**

Dataset → **XLearning-SCU/SUN-R-D-T**