Code is Open

Give a Star

TIME
Towards Intelligence MEchanism

山东大学
SHANDONG UNIVERSITY

深圳河套學院
Shenzhen Loop Area Institute

NEURAL INFORMATION PROCESSING SYSTEMS

VT-FSL: Bridging Vision and Text with LLMs for Few-Shot Learning

Wenhao Li[1, 2], Qiangchang Wang[1]*, Xianjing Meng[3], Zhibin Wu[1], Yilong Yin[1]*

[1]School of Software, Shandong University [2]Shenzhen Loop Area Institute

[3]School of Computing and Artificial Intelligence, Shandong University of Finance and Economics

{wenhao.li, zhibinwu}@mail.sdu.edu.cn,

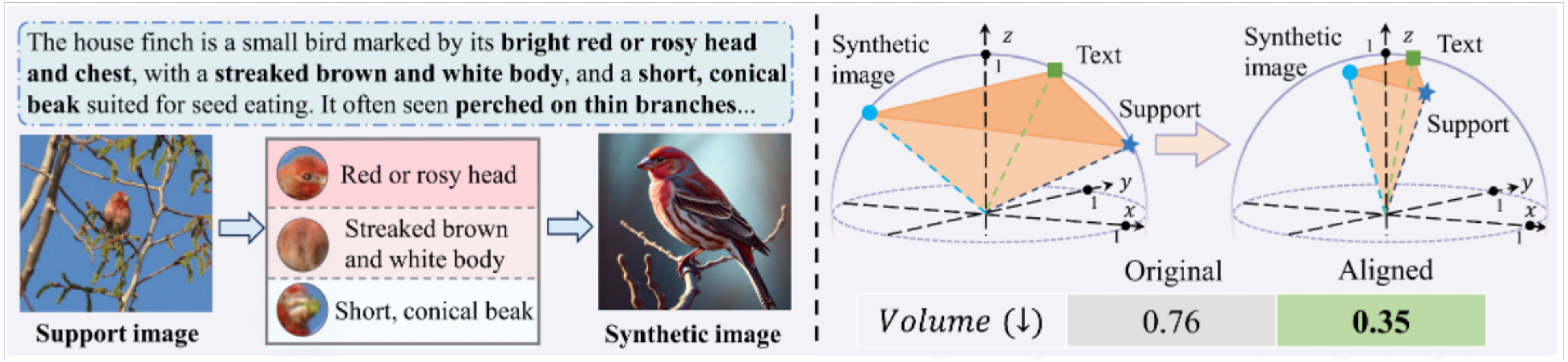
rongmengyuan@gmail.com, {qiangchang.wang, ylyin}@sdu.edu.cn

*Corresponding authors

FSL Task and Motivation
Introduction

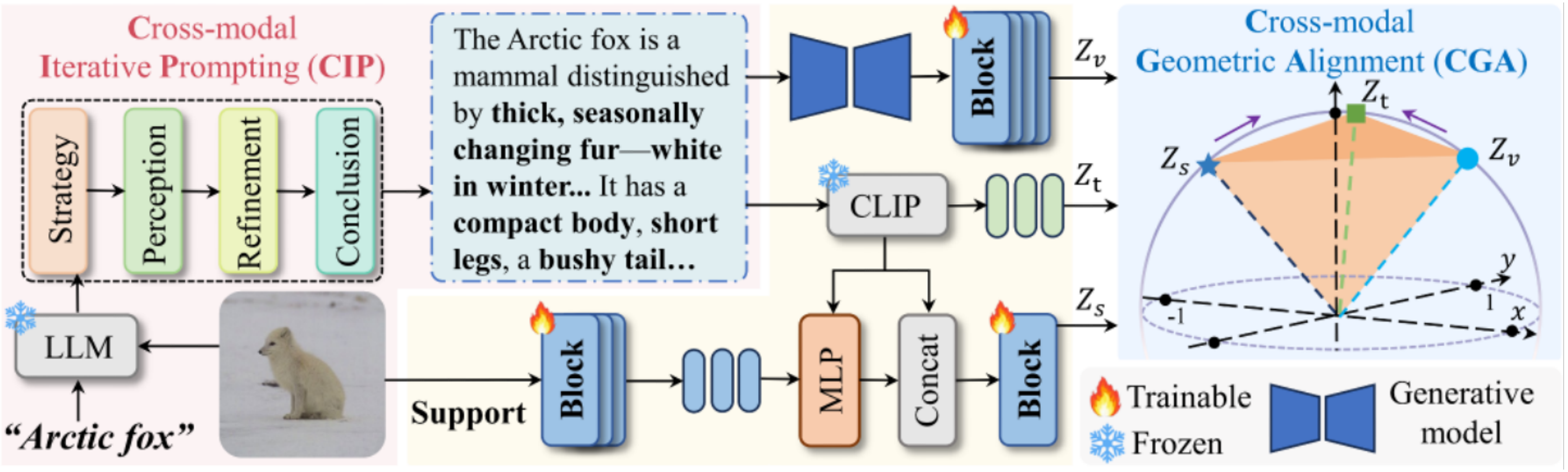
Imagine the challenge of asking a model to recognize an entirely novel category from only a few sample images. Existing LLM-based methods are prone to **semantic hallucinations**, i.e., generating descriptions that contradict the visual evidence due to the lack of grounding in actual instance,requiring significant effort to correct. So the question arises:

- How can we ensure text descriptions generated by large models truly align with visual features?
- How can we simultaneously capture high-level semantic information and low-level visual diversity with limited samples?
- Compared to traditional CLIP pairwise contrastive learning, how can we integrate multimodal information more comprehensively and structurally?



VT-FSL Framework

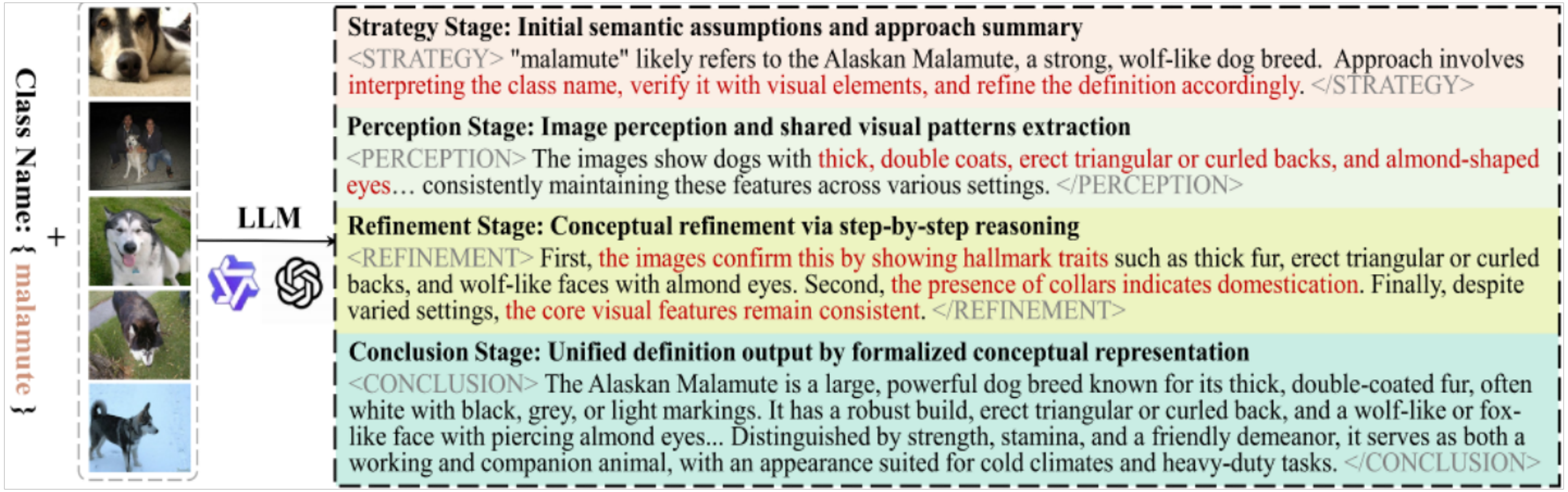
- **Cross-modal Iterative Prompting (CIP):** Beyond mere class names, generate precise and visual-grounding descriptions conditioned on LLMs and support images in a single structured iterative reasoning process.
- Semantically Consistent Images Generation: Zero-shot generation based on textual descriptions to enrich limited samples with diverse yet semantically aligned variations.
- **Cross-modal Geometric Alignment (CGA):** Minimize the kernelized volume of a parallelotope to enforce global consistency among textual descriptions, synthetic images, and supports.



Cross-modal Iterative Prompting (CIP)

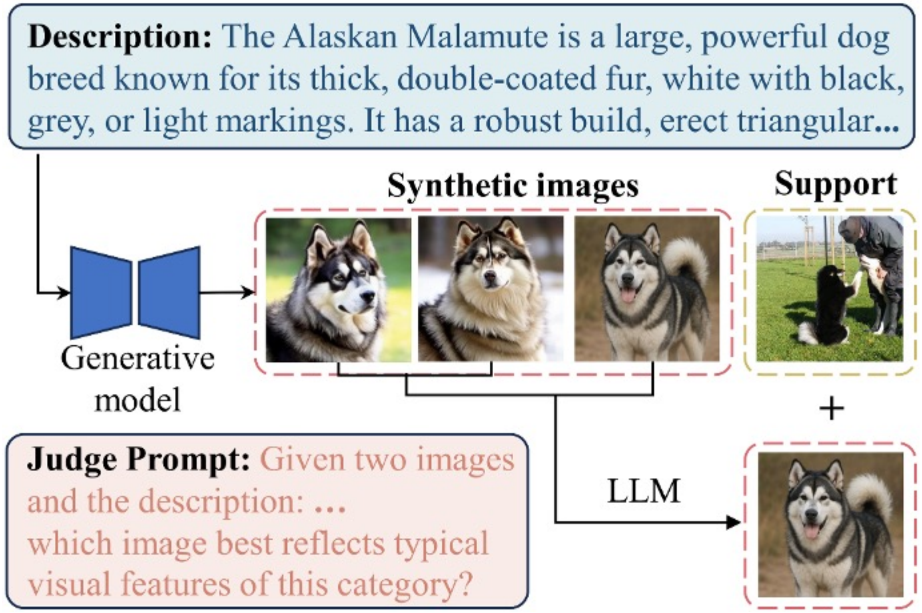
➢ Chain-of-Thought (CoT)-based Four Stage Structured Reasoning Paradigm

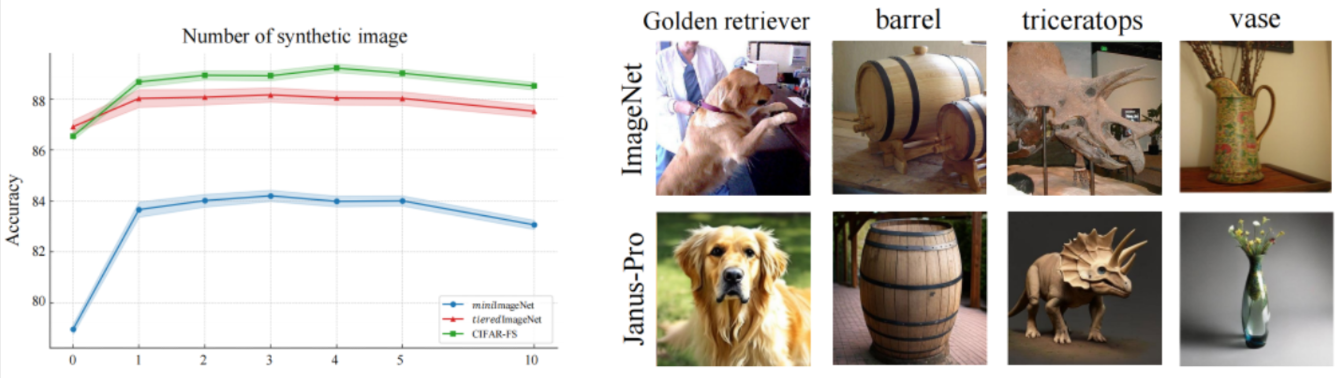
- **Strategy** Stage: Interpret the and propose initial semantic assumptions.
- **Perception** Stage: Extract visual patterns and shared attributes.
- **Refinement** Stage: Iteratively revise semantics via step-by-step reasoning aligned with visual evidence.
- **Conclusion** Stage: Output a unified, visually grounded class definition with formalized conceptual representation.



Cross-modal Iterative Prompting (CIP)

The generated description is then fed into a text-to-image generative model to produce synthetic imagesin a zero-shot manner. An LLM-based pairwise comparison strategy is designed to select top-K images per class by ranking them against the descriptions without compromising low-data regime





(a) The effect of the generated number (b) Visualization of generated images

Figure 7: The effect of the generated number and visualization of generated images.

Cross-modal Geometric Alignment (CGA)

NEURAL INFORMATION PROCESSING SYSTEMS

➢ **Geometric-aware structured and consistent multimodal learning**

Given multiple vectors $\{v_1, \dots, v_k\}$ to construct matrix $A = [v_1, \dots, v_k]$, the Gram is:

$$G(v_1, \dots, v_k) = A^{\mathrm{T}}A,\ G_{ij} = \langle \mathbf{v_i}, \mathbf{v_j} \rangle.$$

Gram matrix G reflects the square of the corresponding volume Vol :

$$\mathrm{Vol}(\mathbf{v_1}, \dots, \mathbf{v_k}) = \sqrt{\det(\mathbf{G})}\,.$$

To capture nonlinear relations, we compute volume in RKHS via an RBF kernel:

$$\mathrm{Vol}_{\mathcal{H}}(v_1, \dots, v_k) = \sqrt{\det(\mathbf{K})}\,, \quad K_{ij} = \kappa\big(v_i, v_j\big).$$

Given normalized triplets $\{v_t, v_v, v_s\}$ from textual, vision, and synthetic embeddings, minimizing the kernelized volume through a contrastive objective:

$$\mathcal{L}_{\mathcal{D}2\mathcal{A}} = \frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\ \left(-\mathrm{Vol}_{\mathcal{H}}\left(z_t^i, z_s^i, z_v^i\right)/\tau\right)}{\sum_{j=1}^{K}\ \exp\ \left(-\mathrm{Vol}_{\mathcal{H}}\left(z_t^i, z_s^i, z_v^j\right)/\tau\right)} \qquad \mathcal{L}_{\mathcal{A}2\mathcal{D}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\ \left(-\mathrm{Vol}_{\mathcal{H}}\left(z_t^i, z_s^i, z_v^i\right)/\tau\right)}{\sum_{j=1}^{K}\ \exp\ \left(-\mathrm{Vol}_{\mathcal{H}}\left(z_t^j, z_s^j, z_v^i\right)/\tau\right).}.$$

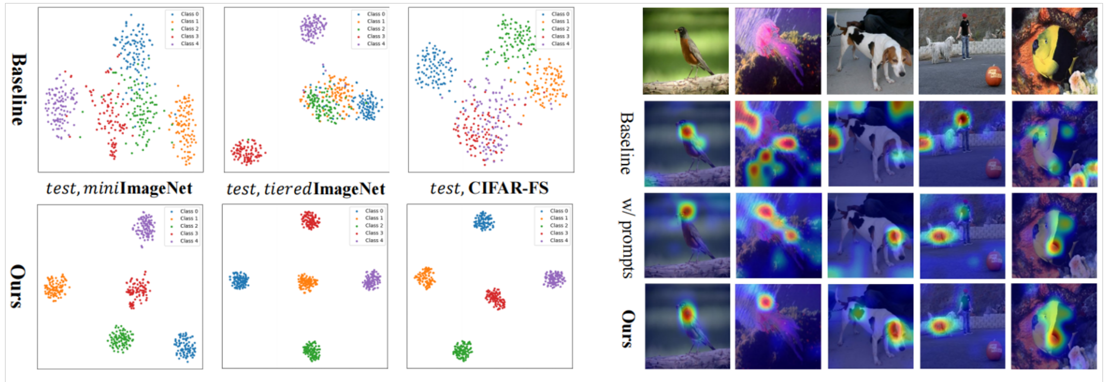
Experimental Results

On ten standard, cross-domain, and fine-grained datasets, improving accuracy by 4.2% on average. Bold and blue indicates the best and suboptimal results.

Model	Venue	Backbone	≈ # Params	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	*tiered*ImageNet 1-shot	*tiered*ImageNet 5-shot
CPEA [58]	ICCV'23	ViT-S/16	22.0M	71.97±0.65	87.06±0.38	76.93±0.70	90.12±0.45
FeatWalk [13]	AAAI'24	ResNet-12	12.4M	70.21±0.44	87.38±0.27	75.25±0.48	89.92±0.29
SemFew [28]	CVPR'24	Swin-T	29.0M	78.94±0.66	86.49±0.50	82.37±0.77	89.89±0.52
UAP [59]	NeurIPS'24	ResNet-12	12.4M	81.63±0.28	79.05±0.19	79.68±0.30	76.78±0.21
VT-FSL	ours	Visformer-T	10.0M	**83.66±0.31**	**88.38±0.25**	**88.02±0.34**	**91.71±0.27**

Method	Venue	CUB-200-2011 1-shot	CUB-200-2011 5-shot	Stanford-Dogs 1-shot	Stanford-Dogs 5-shot	Stanford-Cars 1-shot	Stanford-Cars 5-shot
SUITED [57]	AAAI'25	86.02±0.47	94.13±0.24	76.55±0.47	88.86±0.27	89.97±0.36	96.53±0.16
VT-FSL	ours	**91.08±0.28**	**94.63±0.19**	**86.58±0.30**	**90.69±0.25**	**92.95±0.24**	**96.62±0.15**

Method	Venue	CUB 1-shot	CUB 5-shot	Places 1-shot	Places 5-shot	Plantae 1-shot	Plantae 5-shot
StyleAdv [58]	CVPR'23	48.49±0.72	68.72±0.67	58.58±0.83	77.73±0.62	41.13±0.67	61.52±0.68
MEFP [59]	NeurIPS'24	51.55±0.70	73.61±0.66	52.06±0.69	73.78±0.61	41.55±0.65	61.39±0.67
SVasP [70]	AAAI'25	49.49±0.72	68.95±0.66	59.07±0.81	77.78±0.62	41.22±0.62	60.63±0.64
VT-FSL	ours	**66.86±0.47**	**81.02±0.36**	**73.68±0.41**	**81.52±0.33**	**45.90±0.40**	**61.54±0.38**



(a) T-SNE visualization on novel classes (b) Visualization of attention maps