CMC UNIVERSITY
Aspire to Inspire the Digital World

ICAI
Impacts and Potentials

THE FIRST INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE: IMPACTS AND POTENTIALS IN 2025 (ICAI-IP 2025)

Training-Free Multi-Modal Alignment for Fine-Grained Couterfeit Fruit Detection

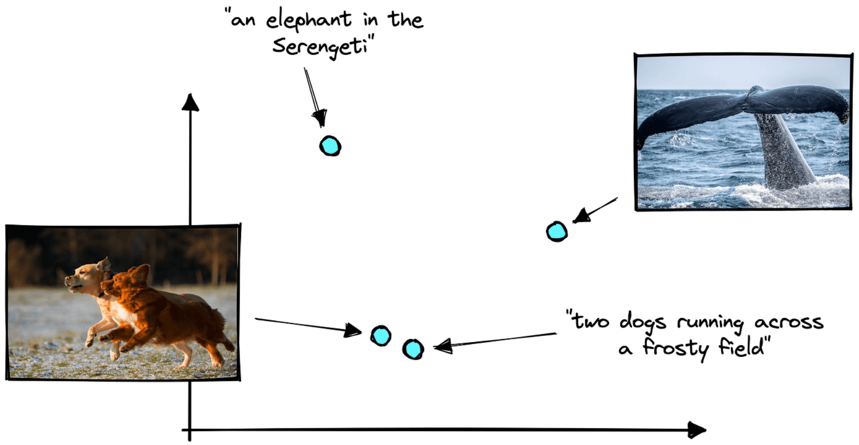
Quynh Nguyen Huu[1], Dat Tran-Anh[2], Thieu Huy Nguyen[2], Ngoc Anh Nguyen Thi[1], Nguyen Huu Gia Bach[3]

[1]CMC University, Hanoi, Vietnam;
[2]Thuyloi University, Hanoi, Vietnam;
[3]Yen Hoa High School, Hanoi, Vietnam

Few-Shot Class-Incremental Learning (FSCIL)

What is FSCIL?

- Learn classes sequentially across sessions (base → incremental).
- Each new session contains very few samples (5-shot).
- No access to old data, but must classify all seen classes so far ($Y_0 \dots Y_t$).



Key Challenges

✓ model forgets old classes as new classes arrive.

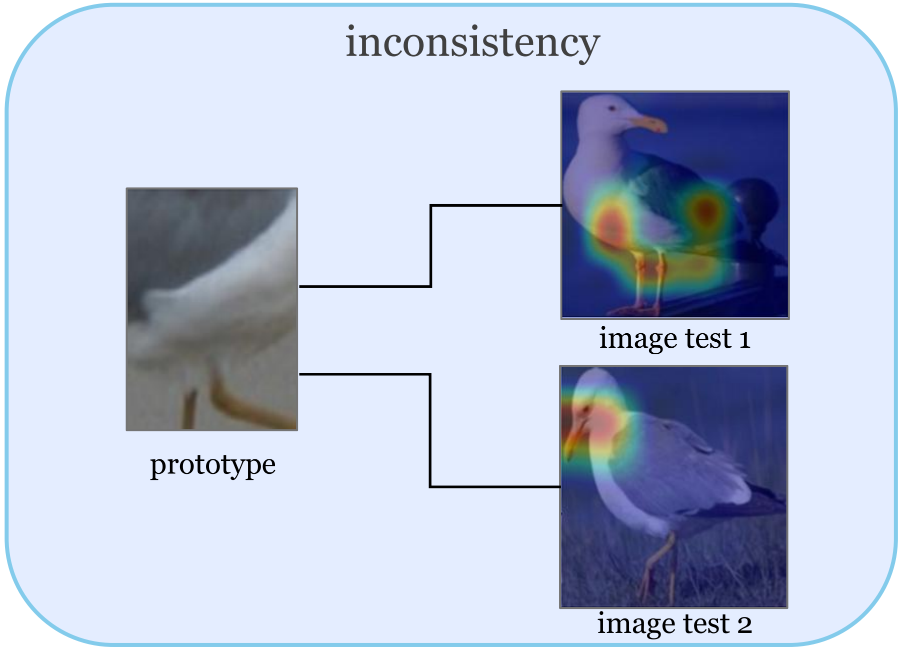
✓ few-shot prototypes are noisy and unstable.

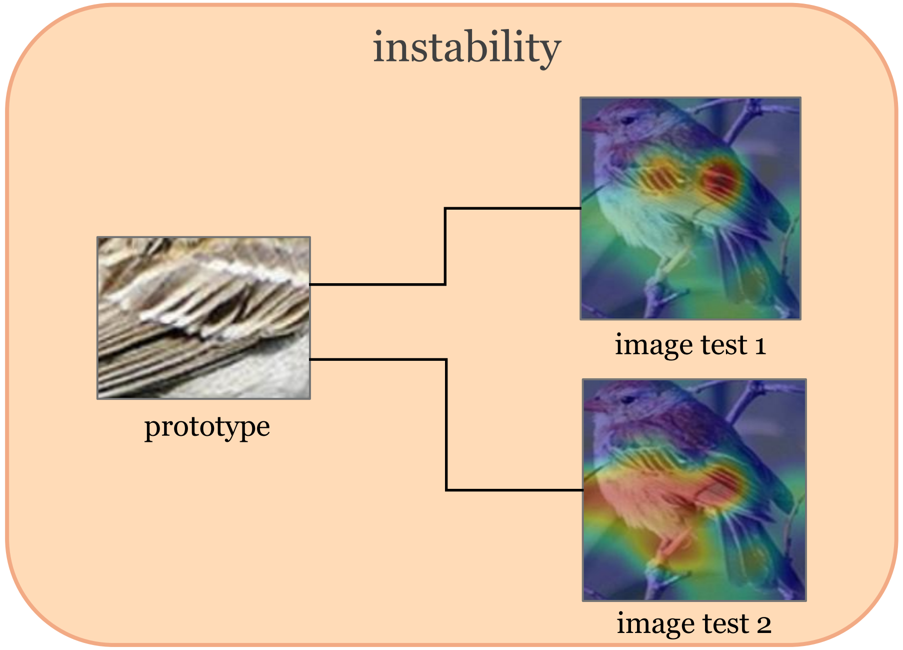
✓ Severe fine-grained similarity: new classes often visually overlap with old ones.

Motivation:

Existing Vision-Language Models (VLMs) in FSCIL paradigm suffers from modality mismatch and unstable prototypes.

- ➢ visual & textual embeddings are misaligned → inaccurate similarity



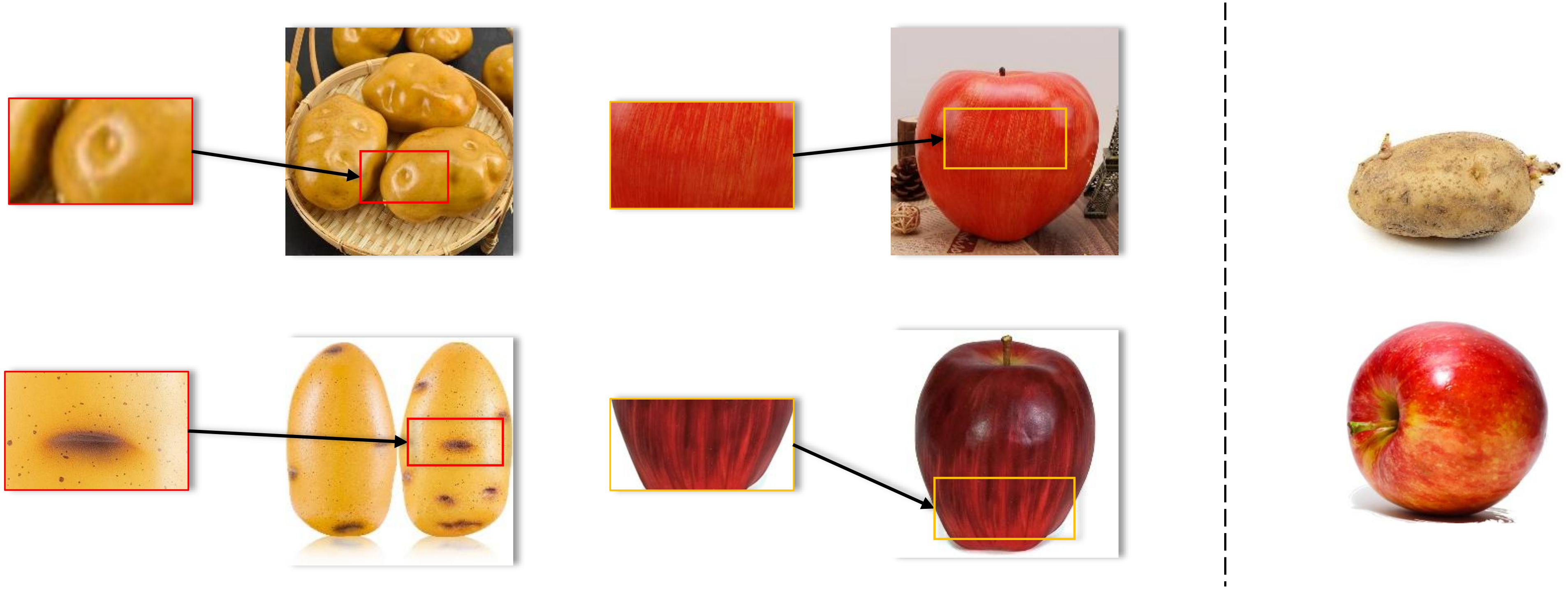


2. Introduction (2/4)

Motivation:

Existing Vision-Language Models (VLMs) in FSCIL paradigm suffers from modality mismatch and unstable prototypes.

- ➢ visual & textual embeddings are misaligned → inaccurate similarity
- ➢ counterfeit vs genuine fruit shares nearly identical appearance

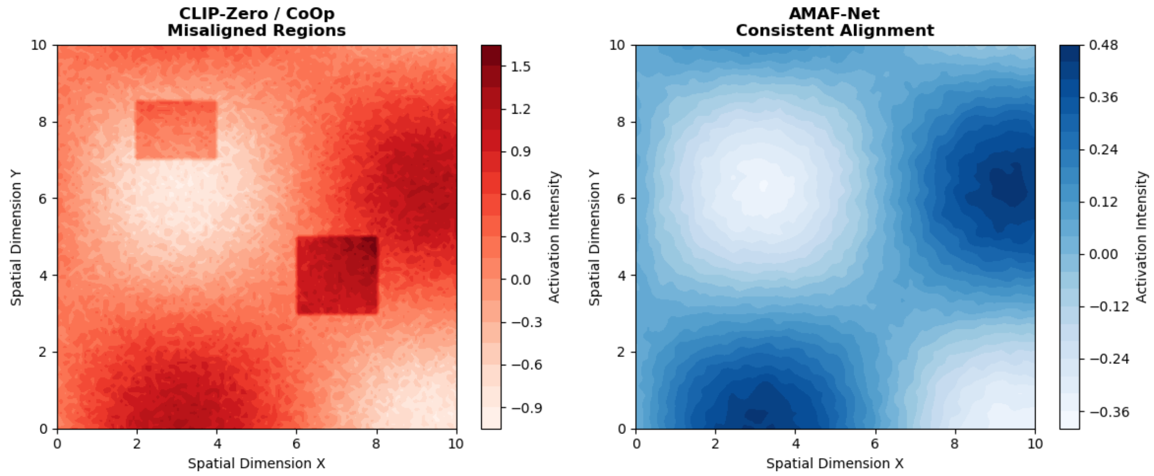


✗ Counterfeit ✓ Genuine

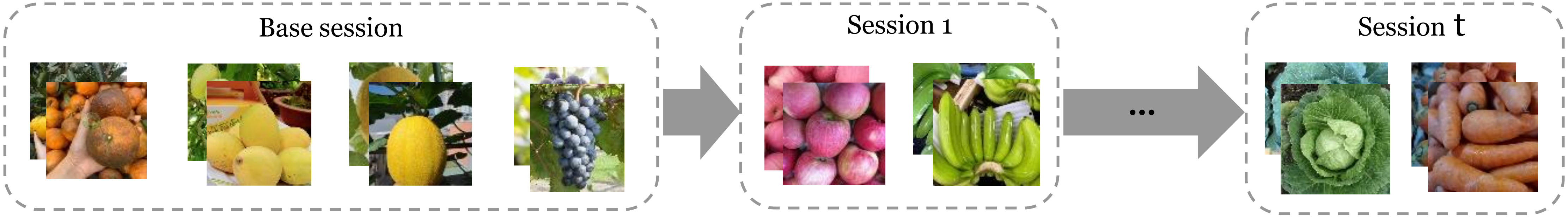
Contribution:

We propose **AMAF-Net** (Adaptive Multi-modal Alignment Framework) to overcome modality and prototype mismatch

✓ more discriminative and consistent representations for fine-grained counterfeit



✓ robust cross-session performance by reducing incremental drift



2. Introduction (4/4)

AMAF-Net (Adaptive Multi-modal Alignment Framework) Paradigm:

✓ widely needed for agricultural inspection and low-data incremental environments

Stage 1

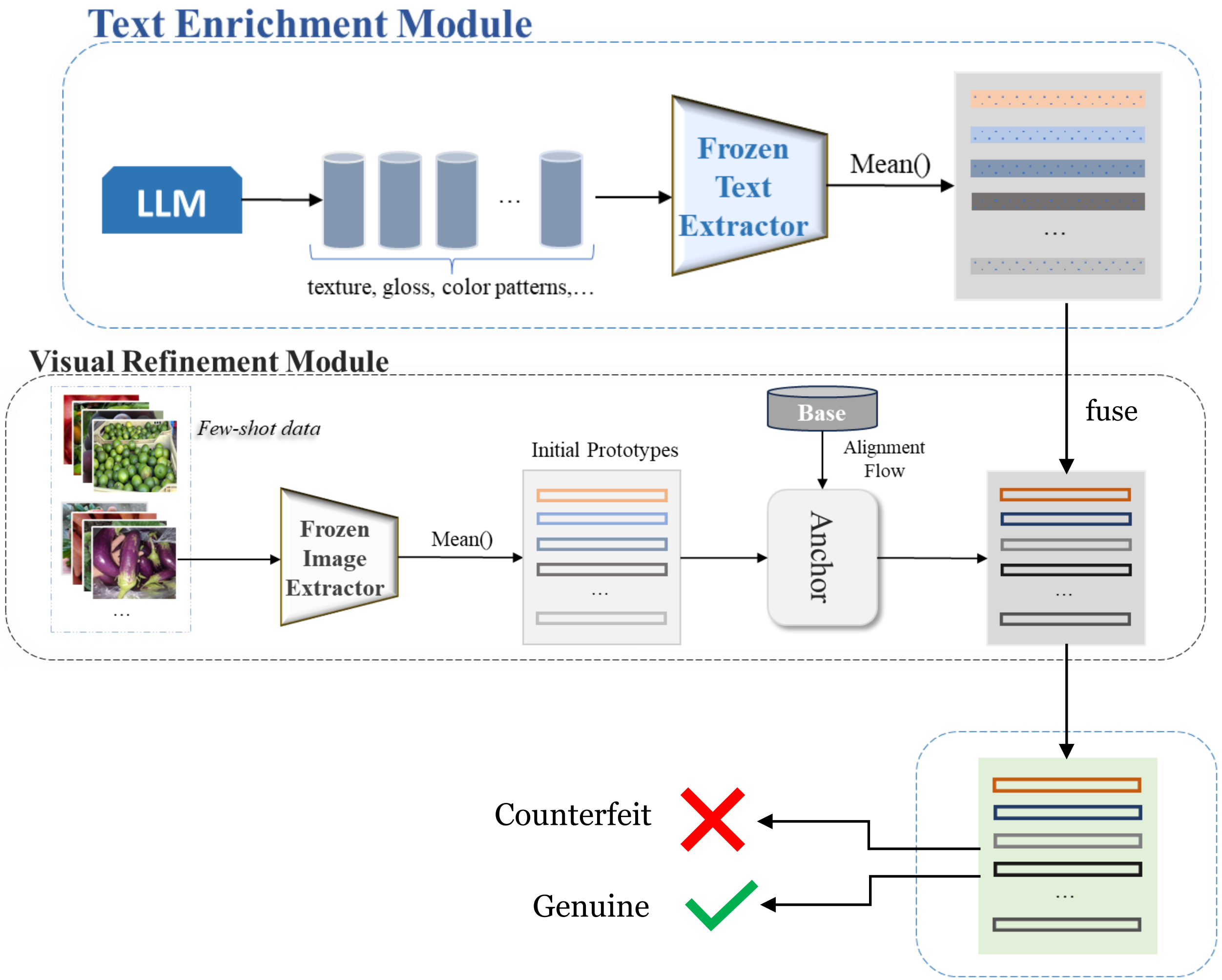
Text Enrichment: generate attribute-guided prompts via LLMs (texture, gloss, color gradient)

Stage 2

Visual Refinement: anchor few-shot prototypes to base-session knowledge

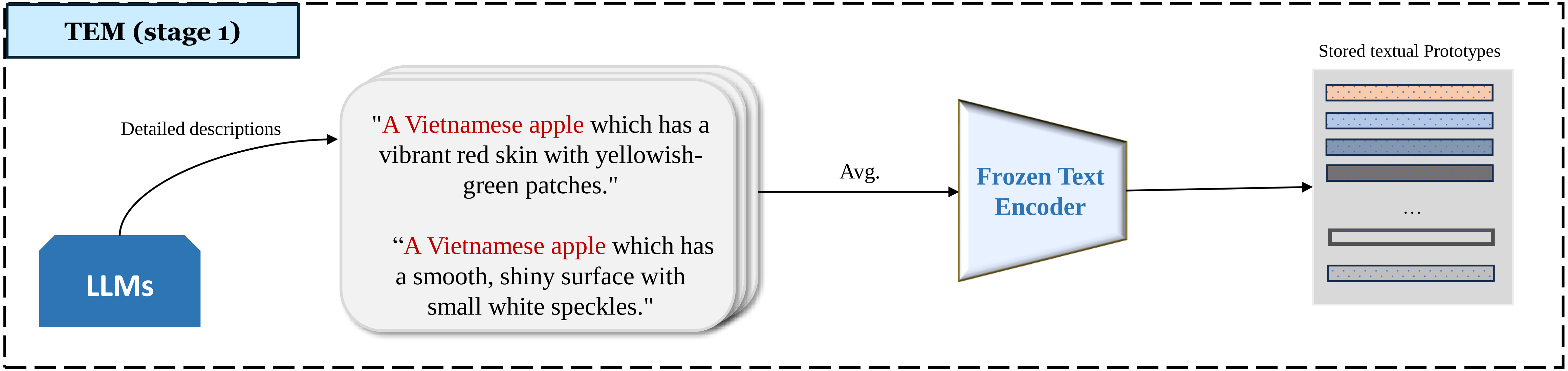
Stage 3

Adaptive Alignment Fusion: Align modalities and make predictions



CMC UNIVERSITY
Aspire to Inspire the Digital World
ICN
Impacts and Potentials

Text Enrichment Module (TEM)



❑ Object function

- ✓ Semantic Expansion: LLM generates K attribute-rich descriptions → richer textual embedding space.
- ✓ Prompt Diversity: Remove near-duplicate prompts (cosine similarity > 0.92) to ensure semantic variety.

Text Prototype Aggregation:

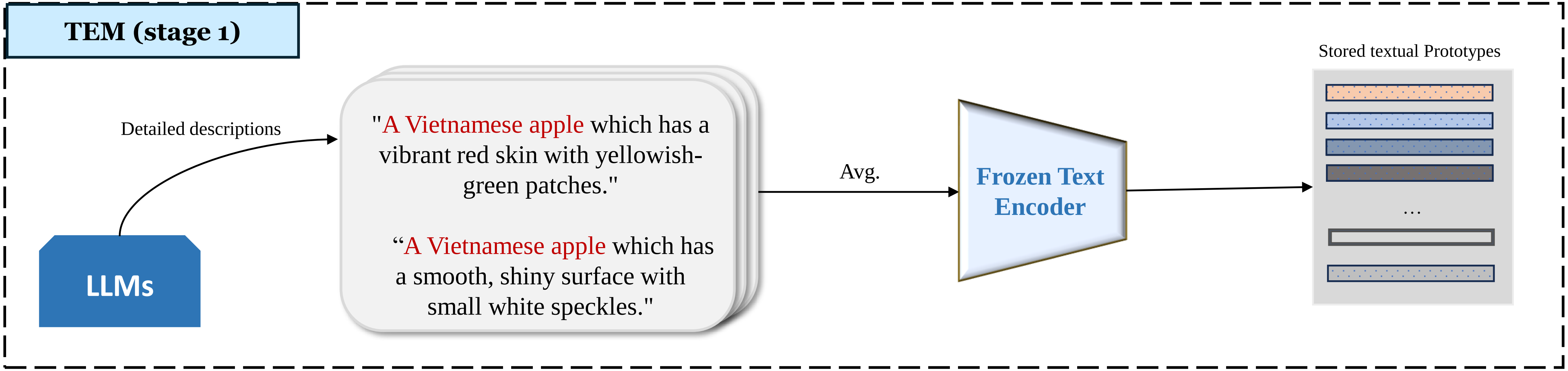
$$\boxed{t_c = E_{d \sim D_c}[f_{txt}(d)]}$$

$f_{txt}(d_c^K)$: a frozen text encoder

→ smooths noise and improves fine-grained discriminability.

CMC UNIVERSITY
Aspire to Inspire the Digital World
ICN
Impacts and Potentials

Text Enrichment Module (TEM)



❑ Object function

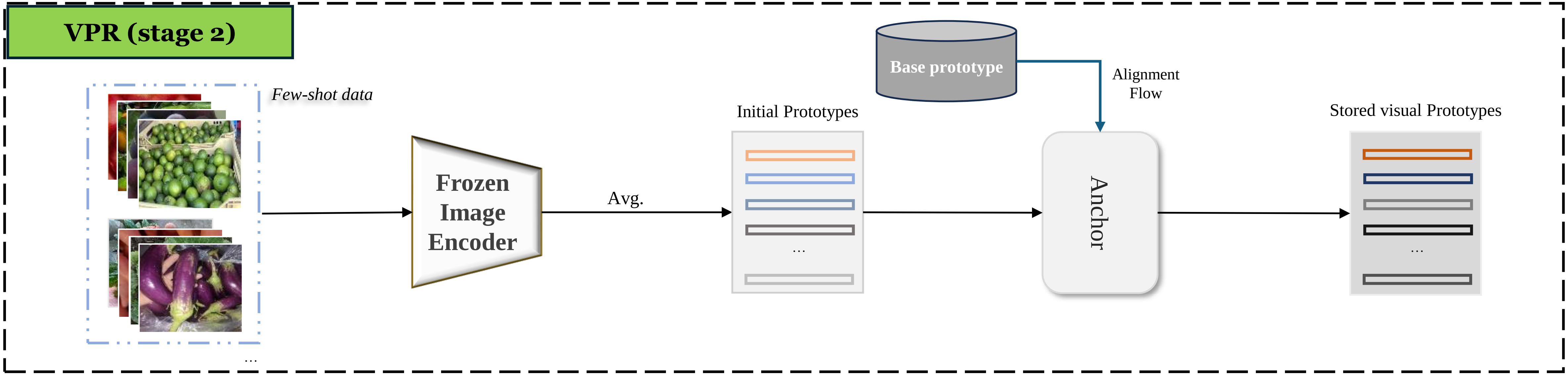
✓ Semantic Expansion: LLM generates K attribute-rich descriptions → richer textual embedding space.

✓ Prompt Diversity: Remove near-duplicate prompts (cosine similarity > 0.92) to ensure semantic variety.

- Without it, prompts too generic → poor separation of subtle counterfeit cues.
- With it, detailed semantic anchors → accurate, consistent boundaries.

CMC UNIVERSITY Aspire to Inspire the Digital World
ICN Impacts and Potentials

Visual Prototype Refinement (VPR)



❑ Object function

✓ Few-shot prototype instability: Raw prototypes from 1–5 images are noisy → drift across sessions.

✓ Adaptive reliability: fuses this prototype with its closest base anchor b_c (obtained from base-session classes):

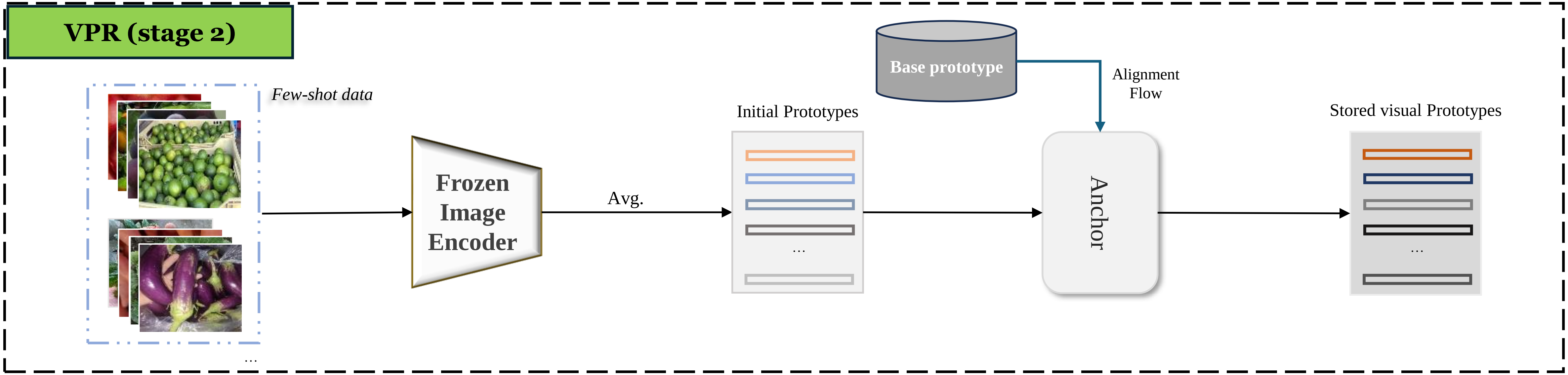
$$v_c = \eta v_c^{raw} + (1 - \eta) b_c$$

where b_c: nearest base-session anchor (stable visual reference).

CMC UNIVERSITY
Aspire to Inspire the Digital World

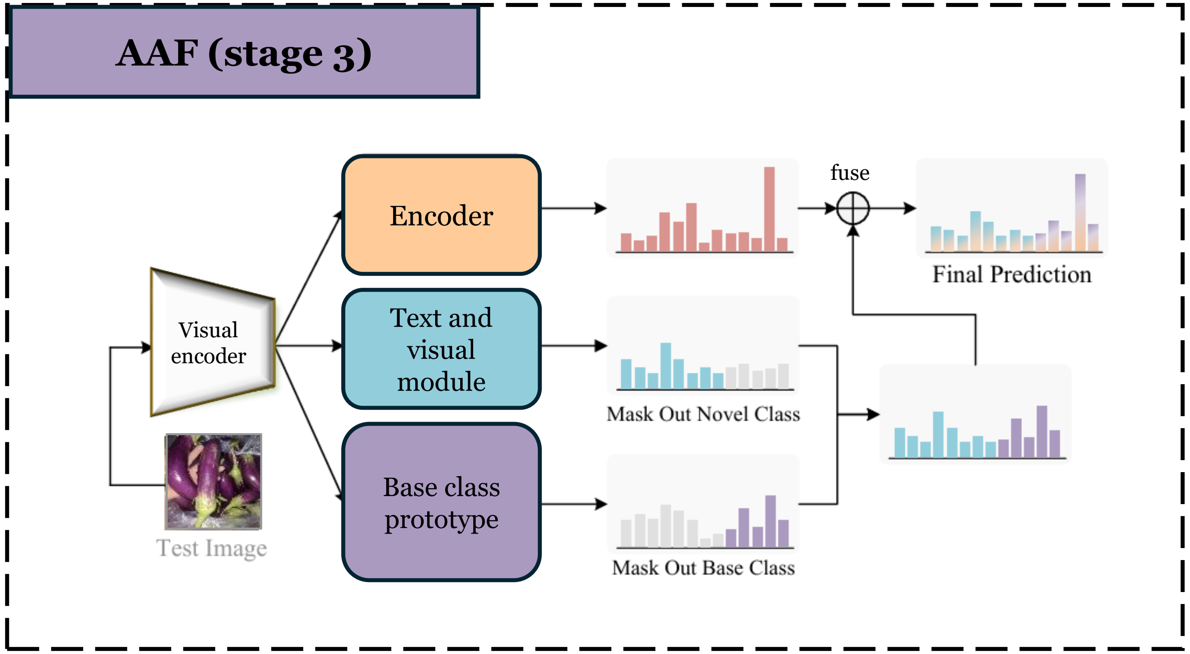
ICN
Impacts and Potentials

Visual Prototype Refinement (VPR)



- ❑ Object function
  - ✓ Effect:
    - more stable feature distribution
    - improved consistency across sessions
    - reduced catastrophic forgetting

Adaptive Alignment Fusion (AAF)



- **Stacked modality-specific scoring and mask out classes**
  - ✓ Visual-similarity block: provide image-based confidence and similarity, highlight discriminative visual cues.
  - ✓ Textual-similarity block: provide semantic confidence and similarity, capturing LLM-enriched
  - ✓ Fusion block: integrates both signals using adaptive confidence weighting ,ensure the more reliable modality contributes.

- Object function
  - ✓ confidence-guided fusion of visual & textual similarities

$$S_{img}(q,c) = \frac{z_q \cdot v_c}{\|z_q\| \, \|v_c\|},$$

$$S_{txt}(q,c) = \frac{z_q \cdot t_c}{\|z_q\| \, \|t_c\|}$$

✓ the final multimodal confidence is

$$S(q,c) = \beta_c S_{img}(q,c) + (1 - \beta_c) S_{txt}(q,c),$$

then predict:

$$\hat{y} = \arg\max S(q,c)$$

CMC UNIVERSITY Aspire to Inspire the Digital World ICN Impacts and Potentials

Quantitative results

❑Few-shot class-incremental setting:

✓ outperform all VLM-based baselines across MiniImageNet, CUB-200, EuroSAT, and FG-Fruit

✓ achieve top-tier performance on fine-grained datasets (CUB-200 & FG-Fruit), especially in counterfeit detection tasks with subtle attribute differences.

Method	MiniImageNet		CUB-200		EuroSAT		FG-Fruit	
	Acc_{avg} (%)	Drop (%)	Acc_{avg} (%)	Drop (%)	Acc_{avg} (%)	Drop (%)	Acc_{avg} (%)	Drop (%)
CLIP-Zero	84.4	**-0.24**	87.0	-0.30	86.7	**-0.41**	87.0	**-0.29**
CoOp	86.8	-0.11	89.3	-0.13	88.6	0.14	89.3	0.22
CoCoOP	87.4	0.021	89.5	**-0.41**	88.8	0.21	88.9	-0.16
FSIL-VL	86.1	-0.15	89.0	0.12	89.1	-0.18	88.4	0.26
Ours	**87.9**	-0.11	**90.6**	0.20	**89.6**	-0.12	**90.2**	0.13

Table 1 Performance comparison across four FSCIL datasets in the 5-shot setting. The most optimal results are emphasized in bold.

4. Experiments (2/3)

CMC UNIVERSITY
Aspire to Inspire the Digital World

Impacts and Potentials

Quantitative results

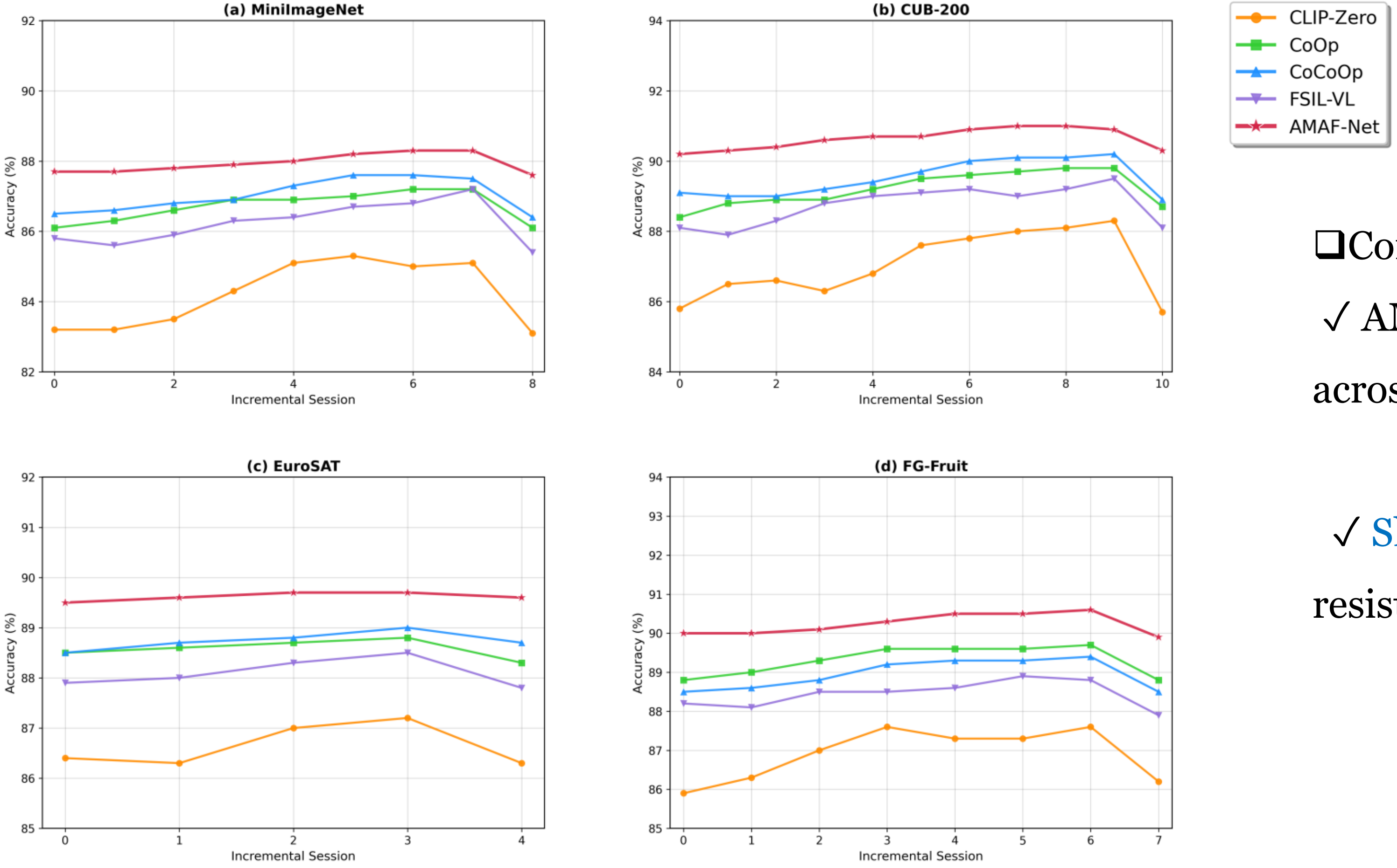


Figure 2 AMAF-Net demonstrates slower performance degradation and maintains higher accuracy across all sessions compared to other methods.

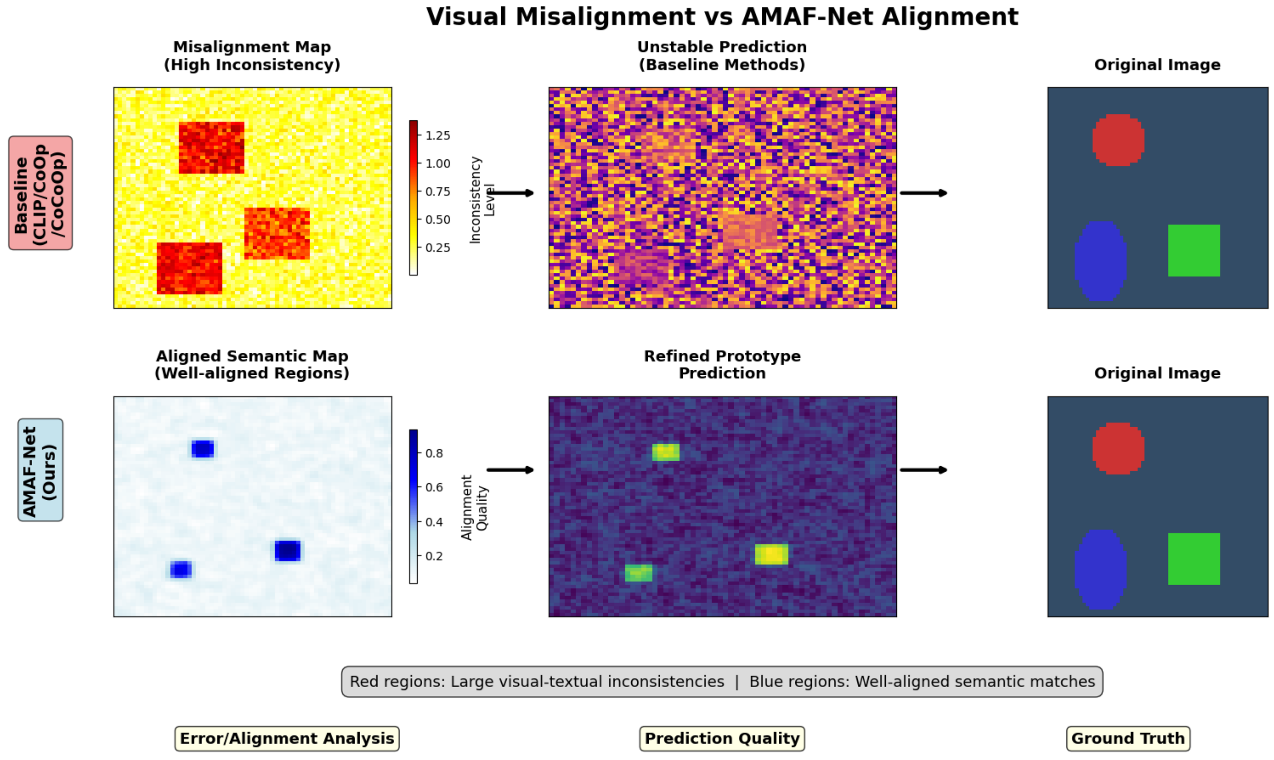
❑Comparison with VLM-based FSCIL methods:

✓ AMAF-Net consistently achieves the highest accuracy across all sessions.

✓ Slower degradation as sessions progress → strong resistance to incremental drift

4. Experiments (3/3)

Quantitative results



red regions: high modality inconsistency

blue regions: well-aligned semantics

❑Visualization of multi-modal alignment

our alignment maps demonstrate:

✓ the visually confusing regions with subtle differences are corrected through LLM-enriched attribute prompts

✓ the unstable prototype regions with high variance are stabilized by anchoring few-shot prototypes to base knowledge.

CMC UNIVERSITY
Aspire to Inspire the Digital World
ICN
Impacts and Potentials

Effect of Individual Modules on FG-Fruit dataset

✓ **TEM** — Text Enrichment Module

✓ **VPR** — Visual Prototype Refinement

✓ **AAF** — Adaptive Alignment Fusion

Method variant	TEM	VPR	AAF	$Acc_{avg}\uparrow$
Baseline CLIP-Zero	✗	✗	✗	87.0
+ TEM only	✓	✗	✗	88.4
+ VPR only	✗	✓	✗	88.1
+ TEM + VPR	✓	✓	✗	89.0
+ TEM + AAF	✓	✗	✓	89.3
+ VPR + AAF	✗	✓	✓	89.1
Full AMAF-Net (ours)	✓	✓	✓	**90.2**

Table 2 Effect of ATE, PRB, and AAF Modules on FSCIL Performance

✓ **TEM-only**
→ semantic prompts enrich class descriptions and reduce visual ambiguity.

✓ **VPR-only**
→ reducing drift across incremental sessions.

✓ **AAF-only**
→ improving robustness when either modality is unreliable.

→ Full AMAF-Net gains +3.2% over CLIP-Zero and maintains the highest stability.

Effect of Individual Modules on FG-Fruit dataset

✓ **TEM** — Text Enrichment Module

✓ **VPR** — Visual Prototype Refinement

✓ **AAF** — Adaptive Alignment Fusion

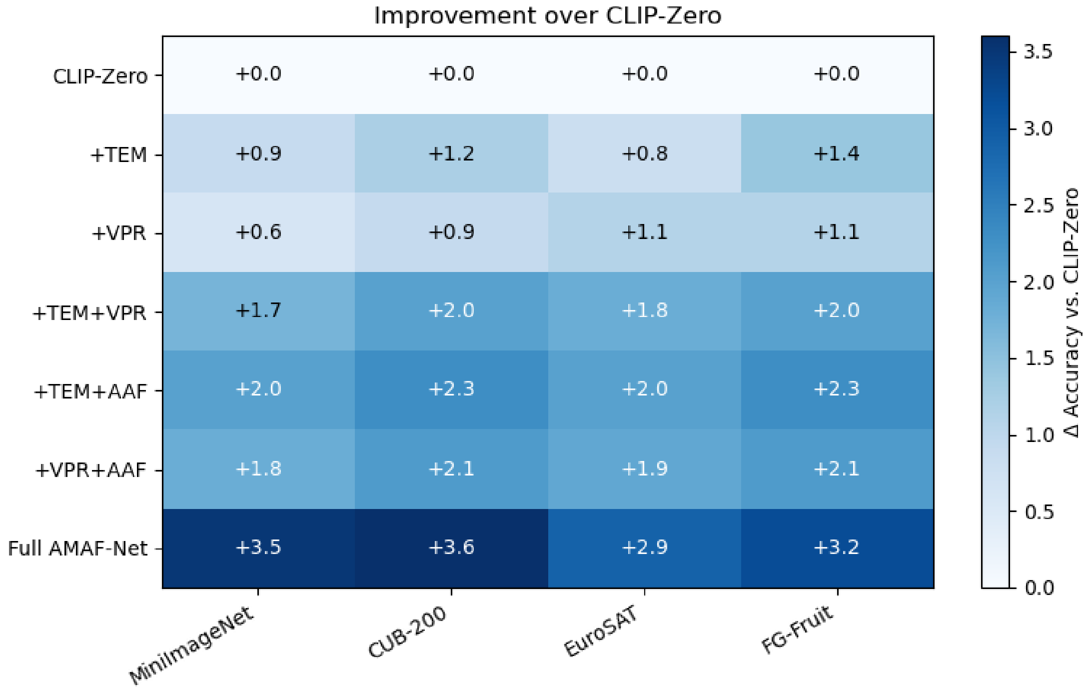


Figure 3 Improvement of ATE, PRB, and AAF Modules on all datasets

✓ **TEM-only**
→ semantic prompts enrich class descriptions and reduce visual ambiguity.

✓ **VPR-only**
→ reducing drift across incremental sessions.

✓ **AAF-only**
→ improving robustness when either modality is unreliable.

→ **Full AMAF-Net** gains +3.2% over CLIP-Zero and maintains the highest stability.

6. Conclusion and Future work

CMC UNIVERSITY
Aspire to Inspire the Digital World

ICN
Impacts and Potentials

Conclusion	Future work
✓ AMAF-Net introduces three modules: Text Enrichment, Visual Prototype Refinement, and Adaptive Alignment Fusion.	✓ Extend AMAF-Net to few-shot open-set and out-of-distribution scenarios.
✓ These modules jointly improve fine-grained recognition under FSCIL settings.	✓ Explore stronger task-specific prompt-generation strategies (LLM tuning).
✓ AMAF-Net achieves state-of-the-art results on MiniImageNet, CUB-200, EuroSAT, and FG-Fruit.	✓ Apply AMAF-Net to real-time agricultural monitoring and counterfeit inspection.
✓ Shows reduced incremental drift and more stable cross-modal alignment	✓ Investigate multi-view / multimodal extensions with RGB-T

CMC UNIVERSITY
Aspire to Inspire the Digital World

ICAI
Impacts and Potentials

THE FIRST INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE: IMPACTS AND POTENTIALS IN 2025 (ICAI-IP 2025)

THANKS FOR YOUR ATTENTION !