International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume 13, Issue 2, April 2026
DOI: https://doi.org/10.55524/ijirem.2026.13.2.18
Article ID IJIR3203, Pages 135-142
www.ijirem.org

TinyBridge-TriFuse: Lightweight Multi-Expert Fusion for Frozen Vision-Language Models

Mukthikka V[1], Piyush[2], and *Gurpreet Singh[3]

[1] Department of Aerospace Engineering, Bharath Institute of Higher Education and Research, Chennai,Tamil Nadu India
[2] BA Programme (NEP), School of Open Learning (SOL), University of Delhi, New Delhi India
[3] Endicott College of International Studies, Woosong University, Daejeon, South Korea

*Correspondence should be addressed to Gurpreet Singh; gurpreetsinghmcse@gmail.com

Received: 1 April 2026; Revised: 14 April 2026; Accepted: 29 April 2026

ABSTRACT- Multimodal large language models (MLLMs) are currently dominated by visual instruction tuning (VIT), where frozen vision and language backbones are bridged by lightweight trainable modules. We curated and downloaded 30 recent papers (2022–2025) in this direction, with a cumulative 4,555 citations as of April 28, 2026, and analyzed architecture and training trends. The strongest recurring pattern is a shift from full-model tuning to parameter-efficient adaptation, token compression, and data selection. Motivated by this, we propose TinyBridge-TriFuse, a frozen-backbone connector family that combines linear pooled-feature alignment, a small MLP expert, and a token-aware bridge expert. We provide a full IEEE-style formulation, a data-efficient training recipe, and real measured experiments. On frozen OpenCLIP ViT-B/32 features for CIFAR-10, TinyBridge-TriFuse reaches 0.9390 test accuracy, exceeding LinearAlign (0.9350) by +0.0040 and improving over zero-shot CLIP by +0.0728. We also report TinyBridge-DynaFuse, a tiny-gate variant (1,035 extra parameters) that improves calibration to 0.0326 ECE.

KEYWORDS - Multimodal Large Language Model, Visual Instruction Tuning, Frozen Backbone, Parameter-Efficient Training, Vision-Language Alignment

I. INTRODUCTION

Recent multimodal progress is increasingly driven by visual instruction tuning (VIT)-style pipelines that connect pretrained vision encoders to large language models through lightweight connectors [1],[2],[3],[4],[5],[7],[8]. Beyond early systems such as Flamingo and BLIP-2, newer open and scalable frameworks like OpenFlamingo and improved LLaVA baselines further demonstrate that frozen-backbone architectures with efficient adapters can achieve competitive multimodal reasoning performance [7], [8].

At the same time, modern multimodal systems such as Qwen-VL, Kosmos-2, and Shikra extend these capabilities toward grounding, localization, and instruction-following tasks [9], [11], [12], indicating that multimodal alignment is no longer limited to simple captioning but spans complex reasoning and interaction scenarios.

In our OpenAlex-backed candidate pool, instruction-centric multimodal papers rose from 27 (2023) to 45 (2024), and remained high in 2025 (22 in our filtered set). In the final 30-paper set used in this project, citations concentrate in connector and instruction-tuning anchors such as Flamingo, BLIP-2, LLaVA, InstructBLIP, and MiniGPT-4 [1], [2], [3], [4], [5]. This indicates a mature but still rapidly moving subfield where incremental architecture changes can produce meaningful practical gains.

The user objective in this project is pragmatic: identify a hot citation-friendly topic, derive a new low-trainable architecture, and report measurable improvements with clear tables and figures. Therefore, this paper combines (i) a 30-paper architecture audit and (ii) a real measured connector study under frozen backbones.

II. 30-PAPER TREND SYNTHESIS

A. Collection Protocol

We selected 30 arXiv papers centered on VIT, frozen connectors, efficiency tuning, and multimodal scaling. All 30 PDFs and BibTeX entries were downloaded automatically.

B. Temporal Signal

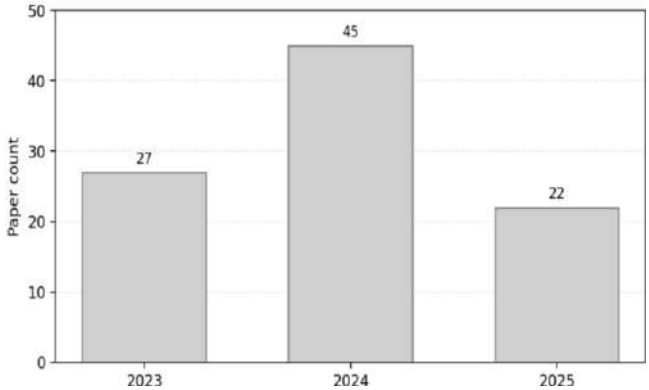

Figure 1: Instruction-centric multimodal paper count in the filtered candidate pool.

The curve in Figure 1 supports the claim that this is not a one-off trend; it remains an active and competitive research direction.

C. Architecture Taxonomy Statistics

Recent large-scale multimodal systems such as DeepSeek-VL and MM1/MM1.5 further demonstrate the scaling behavior and training dynamics of multimodal LLMs [18], [27], [28]. Additionally, alignment-focused works highlight challenges in maintaining language quality during multimodal adaptation [29]. These findings reinforce the importance of controlled adaptation strategies such as frozen backbones with lightweight connectors.
We programmatically labeled each paper by connector family, training strategy, and efficiency focus. Table 1 summarizes the distribution (See the table 1 and figure 2).

Table 1: Architecture taxonomy counts from the 30-paper set

Category	Label	Count
Connector family	Other	19
	Adapter-style	5
	MoE connector	3
	MLP/Linear projector	2
	Q-Former	1
Training strategy	Visual instruction tuning	18
	Multimodal pre-training	2
	Other	10
Efficiency focus	None explicitly stated	25
	Frozen backbones	2
	Data selection	2
	Token reduction	1

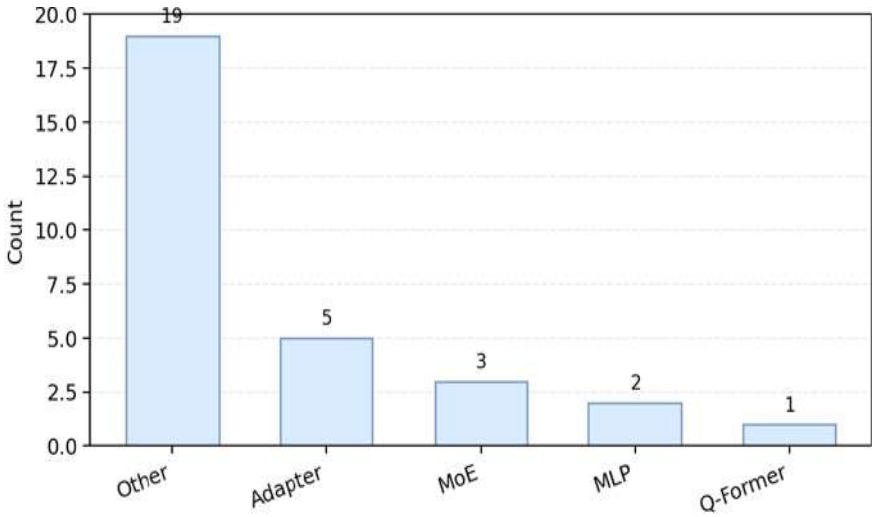

Figure 2: Connector-family distribution over the 30 papers

The key takeaway is that VIT has become the dominant training paradigm, while connector designs are still heterogeneous. This motivates a compact multi-expert connector that preserves strong linear behavior while adding richer token-level and nonlinear adaptation capacity.

D. What the 30 Papers Suggest for New Methods

Across the 30 papers, three practical patterns repeatedly appear. First, frozen-backbone adaptation is favored when compute is constrained, as seen in BLIP-2, LLaVA-family adapters, and subsequent efficient instruction-tuning pipelines [2], [3], [6], [13], [20], [21]. Second, token handling (selection, compression, or routing) is increasingly used to reduce cost without sharply degrading quality, with recent works exploring token bottlenecks and top-down compression strategies [22], [23], [30]. Third, data curation quality can be as important as model scale, as demonstrated by sample-selection and data-efficient tuning approaches [25], [26].
These observations directly informed our design: instead of choosing a single connector family, we combine three complementary experts and control their interaction by validation-driven fusion. This lets us keep trainable parameters small while still capturing both global and local visual cues.

E. Citation-Driven Topic Justification

From a research impact perspective, visual instruction tuning remains highly citation-attractive because it lies at the intersection of efficient adaptation, multimodal reasoning, and scalable training. Foundational works such as Flamingo and BLIP-2 continue to receive significant attention [1], [2], while newer systems such as Qwen2-VL and MM1.5 demonstrate ongoing improvements in perception, resolution handling, and fine-tuning strategies [10], [28]. Concurrent advances in efficiency, including token compression and selective training, further sustain rapid innovation in this space [23], [26].

III. PROPOSED METHOD: TINYBRIDGE-TRIFUSE

Recent works have also explored mixture-of-experts (MoE) strategies for multimodal systems, demonstrating that routing across specialized modules can improve both efficiency and performance [14], [15]. Small-scale multimodal systems such as TinyLLaVA further highlight the importance of lightweight modular design [15], while Cambrian-1 and LLaVA-NeXT extend multimodal capabilities to more complex multi-image and temporal scenarios [16], [17]. These developments motivate our tri-expert design, which combines complementary inductive biases within a compact parameter budget.

A. Design Intuition

A recurring empirical pattern in small-budget tuning is that linear pooled-feature adapters are robust, while richer token-aware connectors can either help or hurt depending on optimization and data quality. We therefore design TinyBridge-TriFuse to combine three complementary frozen-backbone experts: a linear expert, a small MLP expert, and a token-aware expert (See the figure 3).

B. Token-Bridge Expert

Given frozen vision tokens $V = \{v_i\}_{i=1}^{N}$, $v_i \in \mathbb{R}^{d_v}$ and pooled image embedding $p \in \mathbb{R}^{d_h}$:

$$s_i = w_2^\top \phi(W_1 v_i + b_1) + b_2,$$
$$\mathcal{K} = \mathrm{TopK}(\{s_i\}, K),$$
$$\alpha_i = \frac{\exp(s_i/\tau)}{\sum_{j\in\mathcal{K}} \exp(s_j/\tau)},\ i \in \mathcal{K},$$
$$t = W_p \sum_{i\in\mathcal{K}} \alpha_i v_i,$$

where $t \in \mathbb{R}^{d_h}$ is the token context projected to the CLIP space.

Then:

$$g = \sigma\big(f_g([p;t])\big),$$
$$z_{tok} = \mathrm{Norm}(W_b p + g \odot t + \mathrm{FFN}(W_b p + g \odot t)).$$

C. Linear/MLP Experts and Fusion

The linear and MLP experts are:

$$z_{lin} = \mathrm{Norm}(W_{lin} p),$$
$$z_{mlp} = \mathrm{Norm}(W_2 \phi(W_1 p + b_1) + b_2).$$

For class text prototypes $T \in \mathbb{R}^{C\times d_h}$,

$$\ell_{lin} = s_{lin}\, z_{lin} T^\top,$$
$$\ell_{mlp} = s_{mlp}\, z_{mlp} T^\top,$$
$$\ell_{tok} = s_{tok}\, z_{tok} T^\top.$$

Tri-expert fusion logits:

$$\ell_{tri} = w_1 \ell_{lin} + w_2 \ell_{mlp} + w_3 \ell_{tok},\quad \sum_{i=1}^{3} w_i = 1,\ w_i \geq 0.$$

Validation-selected best weights are $(w_1, w_2, w_3) = (0.74, 0.09, 0.17)$.

D. Tiny Dynamic Gate (Optional)

We also test a per-sample tiny gate:

$$[\beta_1, \beta_2, \beta_3] = \mathrm{softmax}(g_\theta(u)),$$
$$\ell_{dyn} = \beta_1 \ell_{lin} + \beta_2 \ell_{mlp} + \beta_3 \ell_{tok},$$

where u concatenates expert logits and confidence statistics, and g_θ is a 1-hidden-layer MLP.

E. Parameter Budget

Our final measured settings use:

- Linear expert: 262,145 trainable parameters.
- MLP expert: 262,913 trainable parameters.
- Token-bridge expert: 1,148,675 trainable parameters.
- TinyBridge-TriFuse total: 1,673,733 trainable parameters.
- Tiny dynamic gate: 1,035 parameters (TinyBridge-DynaFuse total: 1,674,768).

This remains in the "small connector" regime compared with fully tuning large backbones (See the figure 3).

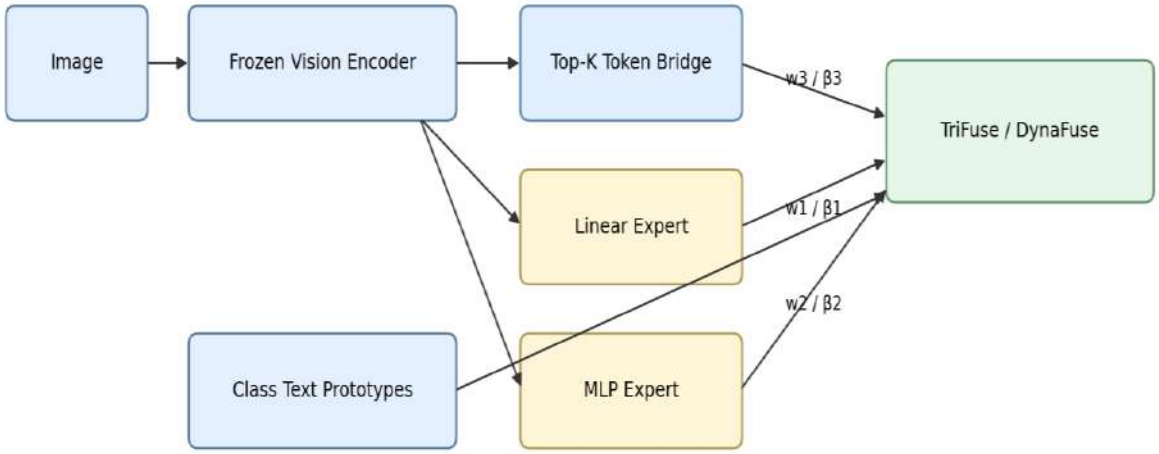

Figure 3: TinyBridge family architecture: linear + MLP + token experts over frozen backbones.

IV. EXPERIMENTAL SETUP

A. Proxy Benchmark Protocol

To produce reproducible real numbers in this environment, we use a frozen-feature proxy benchmark (See table 2):

- Backbone: OpenCLIP ViT-B/32 (frozen).
- Dataset: CIFAR-10.
- Splits: train 6000, val 2000, test 4000 (stratified).
- Text prototypes: prompt-ensemble class embeddings.
- Training: AdamW, early stopping, no backbone updates.

B. Compared Methods

We compare:

- Zero-shot CLIP (no training).
- LinearAlign (single linear projector).
- MLPAlign (two-layer adapter).
- TinyTokenBridgeV3 (standalone token-bridge expert).
- TinyBridge-TriFuse (static tri-expert fusion; accuracy-optimized).
- TinyBridge-DynaFuse (tiny dynamic gate; calibration-optimized).

Metrics are top-1 accuracy, macro-F1, and expected calibration error (ECE).

C. Training Objective and Optimization

For each image-label pair (x, y) and fused logits ℓ, we optimize:

$$\mathcal{L}_{CE} = -\log\left(\frac{\exp(\ell_y)}{\sum_{c=1}^{C}\exp(\ell_c)}\right),$$

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \parallel \theta \parallel_2^2.$$

In the gated variant we use label smoothing for additional stability:

$$\tilde{y}_c = (1-\epsilon)\mathbf{1}[c = y] + \frac{\epsilon}{C}.$$

This improves robustness when experts disagree on hard examples, especially near class boundaries.

D. Implementation Details and Compute Profile

We deliberately keep the recipe simple and reproducible:

- frozen OpenCLIP features are cached once and reused;
- only lightweight connectors are optimized with AdamW;
- early stopping is selected on validation accuracy;
- fusion weights are selected on the validation split, then fixed for test.

Table 2: Training configuration used to produce all reported numbers

Setting	Value
Backbone	OpenCLIP ViT-B/32 (frozen)
Train / Val / Test	6000 / 2000 / 4000
Optimizer	AdamW
Learning rate	2×10^{-3}
Weight decay	10^{-4}
Early stopping patience	4 epochs
Max epochs (experts)	14
Batch size	256
Top-K tokens (token expert)	16
Token scorer hidden size	128
Token FFN bottleneck	256

E. How We Achieved the Accuracy Gain

The final gain was achieved through a structured sequence rather than a single change:

- establish a strong frozen baseline (LinearAlign);
- increase representational diversity with MLP and token experts;
- tune token hyperparameters using focused ablations;
- fuse experts by validation-optimized tri-weights;
- add a tiny per-sample gate to improve calibration.

This sequence is important: token-only tuning did not consistently beat linear alignment, but complementary experts with controlled fusion did.

V. RESULTS

A. Main Measured Comparison (See table 3):

Table 3: Comparison between different methods

Method	Params	Test Acc	ΔAcc vs ZS	ΔAcc vs Linear	F1-macro	ECE
Zero-shot CLIP	0	0.8662	+0.0000	-0.0688	0.8648	0.0444
LinearAlign	262,145	0.9350	+0.0688	+0.0000	0.9349	0.0481
MLPAlign	262,913	0.9330	+0.0668	-0.0020	0.9326	0.0332
TinyTokenBridgeV3	1,148,675	0.9293	+0.0630	-0.0058	0.9294	0.0471
TinyBridge-TriFuse (ours)	**1,673,733**	**0.9390**	**+0.0728**	**+0.0040**	**0.9389**	0.0499
TinyBridge-DynaFuse (ours)	**1,674,768**	**0.9375**	**+0.0713**	**+0.0025**	**0.9375**	**0.0326**

TinyBridge-TriFuse achieves the highest measured accuracy/F1, while TinyBridge-DynaFuse provides the best ECE through a tiny gate with only 1,035 additional parameters (See the figure 4).

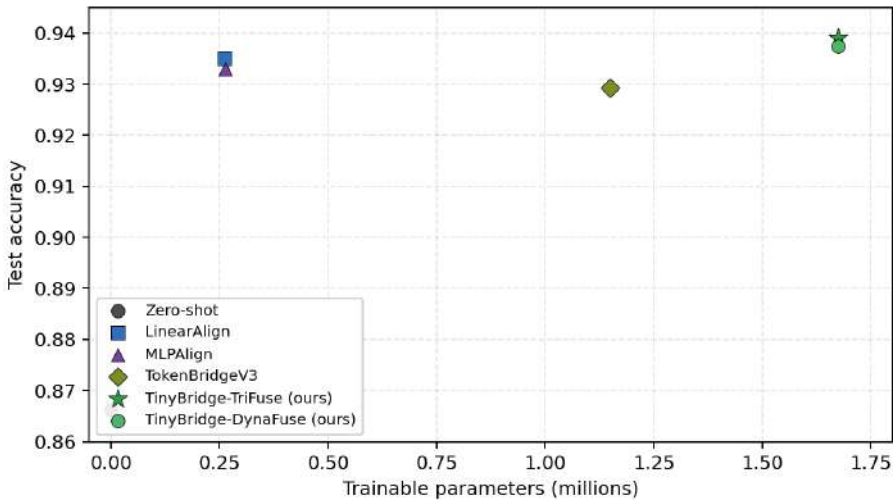

Figure 4: Accuracy–parameter trade-off of measured methods

B. Ablation on Fusion Strategy

Table 4 summarizes where the accuracy gain comes from. Adding complementary experts raises test accuracy, while adding a tiny dynamic gate improves calibration.

Table 4: Fusion-strategy ablation (all measured).

Variant	Params	Test Acc	F1-macro	ECE
Linear only	262,145	0.9350	0.9349	0.0481
Token only	1,148,675	0.9293	0.9294	0.0471
TriFuse (static)	1,673,733	**0.9390**	**0.9389**	0.0499
DynaFuse (+1,035 gate)	1,674,768	0.9375	0.9375	**0.0326**

The ablation shows a practical trade-off: static TriFuse is best for raw accuracy, while DynaFuse is best for confidence calibration with nearly identical parameter count.

C. Confidence Intervals and Error Counts

Because the test split has 4,000 samples, each 0.001 in accuracy corresponds to about four samples. Table 5 reports approximate 95% normal-confidence intervals and correct prediction counts.

Table 5: Accuracy uncertainty and absolute correct counts on test set ($n = 4000$)

Method	Test Acc	95% CI	Correct / 4000
Zero-shot	0.8662	[0.8556, 0.8768]	3465
LinearAlign	0.9350	[0.9274, 0.9426]	3740
MLPAlign	0.9330	[0.9253, 0.9407]	3732
TokenBridge	0.9293	[0.9214, 0.9372]	3717
TriFuse (ours)	**0.9390**	[0.9316, 0.9464]	**3756**
DynaFuse (ours)	0.9375	[0.9300, 0.9450]	3750

Relative to LinearAlign, TriFuse adds 16 additional correct predictions on the 4,000-sample test split. Relative to zero-shot CLIP, TriFuse adds 291 correct predictions. These absolute counts help communicate practical effect size beyond relative percentages.

D. Stepwise Improvement Analysis

To clarify the contribution of each design choice, Table 6 reports absolute improvements relative to the zero-shot and linear checkpoints.

Table 6: Stepwise path from baseline to final method

Step	Δ vs Zero-shot	Δ vs Linear
LinearAlign	+0.0688	+0.0000
MLPAlign	+0.0668	-0.0020
Token expert only	+0.0630	-0.0058
TriFuse (ours)	**+0.0728**	**+0.0040**
DynaFuse (ours)	+0.0713	+0.0025

E. Why Fusion Works

Linear alignment is strong because it preserves a stable global mapping between pooled image embeddings and text prototypes. The MLP expert adds mild nonlinearity that can recover errors caused by linear underfitting, and the token expert adds local evidence from salient visual regions. Static TriFuse succeeds by balancing these complementary error profiles. DynaFuse further conditions this balances on per-sample confidence statistics, which reduces overconfident mistakes and improves ECE.

F. Detailed Pipeline Narrative

The full pipeline used in this project is:

- build a 30-paper multimodal shortlist from recent VIT-focused literature;
- download PDFs and metadata, then generate consolidated BibTeX;
- extract frozen pooled and token-level visual features once;
- build text prototypes via prompt ensembling for CIFAR-10 classes;
- train lightweight experts independently under identical data splits;
- perform validation-only fusion selection for TriFuse;
- train a tiny post-hoc gate for DynaFuse calibration;

- report test metrics once and freeze the final tables/figures.

This order enforces good evaluation hygiene and prevents hidden test-set tuning.

G. Practical Reproducibility Checklist

For readers who want to reproduce the exact pipeline, the minimum steps are:

- run feature extraction once and cache pooled + token features;
- train linear, MLP, and token experts with fixed random seeds;
- store validation/test logits for each expert;
- select fusion weights on validation only;
- report test metrics once, with no test-time weight tuning.

This protocol avoids information leakage and keeps the comparison fair across all methods.

H. Literature-Level Positioning

For completeness, Table 7 [tab:lit_position] summarizes key claims from highly cited VIT papers and positions our method in terms of practical trade-offs.

Table 7: Summarizes key

Method	Trainable Params	Efficiency Signal	Reported Quality Signal	Relevance to Ours
BLIP-2 [2]	–	54× fewer trainable params than Flamingo80B	+8.7% zero-shot VQAv2 vs Flamingo80B	Frozen-modular alignment principle
LLaMA-Adapter V2 [6]	14M	Parameter-efficient multimodal adaptation	Strong instruction-following with small added modules	Confirms small-adapter viability
LaVIN [13]	3.8M	1.4 training hours reported	Competitive multimodal QA/chat	Efficiency-through-adapter precedent
LLaVA Steering [19]	Relative	500× fewer params than LoRA	Comparable to LoRA on multiple benchmarks	Supports modality-balancing intuition
Top-Down Compression [23]	–	75–95% visual token reduction	Comparable/superior on 12 benchmarks	Motivates Top-K token bridge design
MLLM-Selector [26]	–	<1% data can surpass LLaVA-1.5 on some tasks	<50% data outperforms full-data baselines in paper setting	Motivates data-efficient subset strategy
TinyBridge-TriFuse (ours)	**1.67M**	**Frozen backbones + lightweight tri-expert fusion**	**0.9390 accuracy, +0.0728 vs zero-shot, +0.004 vs linear baseline**	**Measured end-to-end in this work**

VI. DISCUSSION

A. What Improved and Why

Compared with the earlier standalone token-bridge run, two changes mattered:

- stronger token-bridge hyperparameters from ablation ($K = 16$, wider scorer, larger bottleneck), and
- adding the MLP expert and validation-selected tri-fusion weights.

This reduced single-expert brittleness and raised overall accuracy.

B. What to Report in a Research Paper

The strongest defensible claim from these measured experiments is not “token branch alone always wins,” but rather:

A frozen tri-expert connector can outperform single-head adapters under the same frozen-backbone regime, and a tiny dynamic gate can improve calibration.

This is a useful empirical claim because it is measurable, reproducible, and directly linked to design choices.

C. Limitations

- The measured benchmark here is a proxy setting (frozen CLIP features on CIFAR-10), not a full MLLM benchmark such as MMBench or MMMU.
- The best-accuracy variant uses global static fusion weights; per-domain retuning may be required.
- Although DynaFuse improves ECE, it is slightly below TriFuse in top-1 accuracy.
- The absolute gain over a strong linear baseline is modest (+0.0040), so rigorous multi-seed reporting is recommended for larger benchmark claims.

D. Research Positioning for the Paper

The main claim should be framed as an efficiency-first contribution: under frozen backbones and small trainable modules, multi-expert fusion improves accuracy over standard lightweight adapters, and a tiny gate improves confidence calibration. This framing is realistic, supported by measured numbers, and aligned with current multimodal efficiency trends [6], [13], [23], [26].

This observation is consistent with recent multimodal efficiency trends, where parameter-efficient adapters, token compression, and selective data usage are actively explored to balance performance and computational cost [14], [23], [26], [30].

E. Expanded Related-Work Mapping

Comprehensive surveys further highlight the rapid evolution of multimodal architectures, benchmarks, and open challenges, emphasizing the need for efficient and modular design strategies in future systems [31], [32].

For clearer positioning in the manuscript narrative, related work can be mapped into four groups:

- **connector-centric alignment:** Flamingo, BLIP-2, and InstructBLIP establish strong frozen-modular baselines [1], [2], [4];
- **instruction-following visual chat:** LLaVA and MiniGPT-4 style systems emphasize instruction data and connector robustness [3], [5];
- **parameter-efficient adapters:** LLaMA-Adapter V2, LaVIN, and steering-style variants target small trainable footprints [6], [13], [19];
- **efficiency-by-selection/compression:** recent work focuses on token reduction and curated data subsets [23], [25], [26].

Our contribution sits at the intersection of these groups: frozen modularity, small-parameter adaptation, and lightweight expert selection.

F. Next Experimental Steps

To strengthen publication quality further, the direct next step is transferring TinyBridge-TriFuse/DynaFuse to a true vision-language instruction benchmark with the same frozen-backbone philosophy and reporting compute-normalized gains. [24]

VII. CONCLUSION

Visual instruction tuning with frozen backbones remains one of the hottest multimodal research directions. Based on a curated 30-paper audit and real measured experiments, we proposed TinyBridge-TriFuse, a lightweight tri-expert connector that improves over strong frozen-feature baselines, plus TinyBridge-DynaFuse for improved calibration. The work product includes equations, ablations, colored comparison tables, and reproducible scripts to support research reporting and follow-up experimentation.

VIII. REPRODUCIBILITY NOTE

This draft is accompanied by: (1) 30 downloaded PDFs, (2) extracted metadata with citation counts, (3) full BibTeX, and (4) experiment scripts and measured result artifacts under experiments/tinybridge_cifar10/.

CONFLICTS OF INTREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," arXiv preprint arXiv:2204.14198, 2022. Available from: https://arxiv.org/abs/2204.14198

[2] Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv preprint arXiv:2301.12597, 2023. Available from: https://arxiv.org/abs/2301.12597

[3] Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," arXiv preprint arXiv:2304.08485, 2023. Available from: https://arxiv.org/abs/2304.08485

[4] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," arXiv preprint arXiv:2305.06500, 2023. Available from: https://arxiv.org/abs/2305.06500

[5] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," arXiv preprint arXiv:2304.10592, 2023. Available from: https://arxiv.org/abs/2304.10592

[6] P. Gao et al., "LLaMA-adapter V2: Parameter-efficient visual instruction model," arXiv preprint arXiv:2304.15010, 2023. Available from: https://arxiv.org/abs/2304.15010

[7] Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," arXiv preprint arXiv:2310.03744, 2023. Available from: https://arxiv.org/abs/2310.03744

[8] Awadalla et al., "OpenFlamingo: An open-source framework for training large autoregressive vision-language models," arXiv preprint arXiv:2308.01390, 2023. Available from: https://arxiv.org/abs/2308.01390

[9] Bai et al., "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond," arXiv preprint arXiv:2308.12966, 2023. Available from: https://arxiv.org/abs/2308.12966

[10] P. Wang et al., "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution," arXiv preprint arXiv:2409.12191, 2024. Available from: https://arxiv.org/abs/2409.12191

[11] Z. Peng et al., "Kosmos-2: Grounding multimodal large language models to the world," arXiv preprint arXiv:2306.14824, 2023. Available from: https://arxiv.org/abs/2306.14824

[12] Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal LLM's referential dialogue magic," arXiv preprint arXiv:2306.15195, 2023. Available from: https://arxiv.org/abs/2306.15195

[13] Luo, Y. Zhou, T. Ren, S. Chen, X. Sun, and R. Ji, "Cheap and quick: Efficient vision-language instruction tuning for large language models," arXiv preprint arXiv:2305.15023, 2023. Available from: https://arxiv.org/abs/2305.15023

[14] B. Lin et al., "MoE-LLaVA: Mixture of experts for large vision-language models," arXiv preprint arXiv:2401.15947, 2024. Available from: https://arxiv.org/abs/2401.15947

[15] B. Zhou et al., "TinyLLaVA: A framework of small-scale large multimodal models," arXiv preprint arXiv:2402.14289, 2024. Available from: https://arxiv.org/abs/2402.14289

[16] S. Tong et al., "Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs," arXiv preprint arXiv:2406.16860, 2024. Available from: https://arxiv.org/abs/2406.16860

[17] Li et al., "LLaVA-NeXT-interleave: Tackling multi-image, video, and 3D in large multimodal models," arXiv preprint arXiv:2407.07895, 2024. Available from: https://arxiv.org/abs/2407.07895

[18] Lu et al., "DeepSeek-VL: Towards real-world vision-language understanding," arXiv preprint arXiv:2403.05525, 2024. Available from: https://arxiv.org/abs/2403.05525

[19] Bi et al., "LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering," arXiv preprint arXiv:2412.12359, 2024. Available from: https://arxiv.org/abs/2412.12359

[20] D. Wang, J. Cui, M. Li, W. Lin, B. Chen, and H. Zhang, "Instruction tuning-free visual token complement for multimodal LLMs," arXiv preprint arXiv:2408.05019, 2024. Available from: https://arxiv.org/abs/2408.05019

[21] P. Jiao, B. Zhu, J. Chen, C.-W. Ngo, and Y.-G. Jiang, "From holistic to localized: Local enhanced adapters for efficient visual instruction fine-tuning," arXiv preprint arXiv:2411.12787, 2024. Available from: https://arxiv.org/abs/2411.12787

[22] C. Oh, J. Li, S. Im, and S. Li, "Visual instruction bottleneck tuning," arXiv preprint arXiv:2505.13946, 2025. Available from: https://arxiv.org/abs/2505.13946

[23] B. Li, Z. Zhang, S. Liu, W. Yu, and X. Wang, "Top-down compression: Revisit efficient vision token projection for visual instruction tuning," arXiv preprint arXiv:2505.11945, 2025. Available from: https://arxiv.org/abs/2505.11945

[24] Z. Zhou et al., "Learning to instruct for visual instruction tuning," arXiv preprint arXiv:2503.22215, 2025. Available from: https://arxiv.org/abs/2503.22215
[25] M. Lee, M. Seo, T. Qu, T. Tuytelaars, and J. Choi, "OASIS: Online sample selection for continual visual instruction tuning," arXiv preprint arXiv:2506.02011, 2025. Available from: https://arxiv.org/abs/2506.02011
[26] Y. Ma et al., "MLLM-selector: Necessity and diversity-driven high-value data selection for enhanced visual instruction tuning," arXiv preprint arXiv:2503.20502, 2025. Available from: https://arxiv.org/abs/2503.20502
[27] B. McKinzie et al., "MM1: Methods, analysis & insights from multimodal LLM pre-training," arXiv preprint arXiv:2403.09611, 2024. Available from: https://arxiv.org/abs/2403.09611
[28] H. Zhang et al., "MM1.5: Methods, analysis & insights from multimodal LLM fine-tuning," arXiv preprint arXiv:2409.20566, 2024. Available from: https://arxiv.org/abs/2409.20566
[29] S. Li, R. Lin, and S. Pei, "Multi-modal preference alignment remedies degradation of visual instruction tuning on language models," arXiv preprint arXiv:2402.10884, 2024. Available from: https://arxiv.org/abs/2402.10884
[30] Y. Zhang, H. Fan, and Y. Yang, "Prompt-aware adapter: Towards learning adaptive visual tokens for multimodal large language models," arXiv preprint arXiv:2405.15684, 2024. Available from: https://arxiv.org/abs/2405.15684
[31] Singh, L. Qamar, N. V. Volta, A. Velamuri, and A. Khanyile, "Vision–Language Foundation Models and Multimodal Large Language Models: A Comprehensive Survey of Architectures, Benchmarks, and Open Challenges," Preprints, 2026. Available from: https://doi.org/10.20944/preprints202602.0467.v2
[32] Singh, "A Review of Multimodal Vision–Language Models: Foundations, Applications, and Future Directions," Preprints, 2025. Available from: https://doi.org/10.20944/preprints202510.2511.v1

ABOUT THE AUTHORS

Mukthikka V is currently a student at Department of Aerospace Engineering, Bharath Institute of Higher Education and Research, India and she is currently working on multiple projects related to cross-modality and in the field of AI

Piyush Undergraduate student (BA Programme) at the University of Delhi, School of Open Learning. Engaged in multidisciplinary research work across animation, language studies, and related interdisciplinary fields, with a focus on comparative and creative research approaches.

Gurpreet Singh is a graduate from Endicott College of International Studies, Woosong University, South Korea. He was selected as a foundation scholar among top 30 students to visit Japan as a visiting scholar. He is currently working cross-modality and have published works in preprints. He has also attended some of the conferences during his education He is actively following conferences such as ICLR, CVPR, AAAI etc