Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Fusion Strategies for Vision-Language Generative Systems

Daniel Popescu[1], Sofia Popescu[2]

[1] Research Scientist, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: daniel.popescu37@yahoo.com | ORCID: 6584-9580-2923-3380
[2] Senior Lecturer, Department of Computer Science, Advanced Computing University, Paris, France. Email: sofia.popescu222@yahoo.com | ORCID: 8877-2874-9939-4854

ABSTRACT

Vision-language generative systems -- models that jointly process visual and textual inputs to produce coherent multimodal or unimodal outputs -- are a central class of multimodal AI systems with diverse applications in image captioning, visual question answering, text-to-image generation, and visual dialogue. The fusion strategy -- the architectural mechanism by which visual and textual representations are combined -- is a critical design decision that substantially impacts generation quality, cross-modal coherence, and computational efficiency. While individual fusion approaches have been extensively studied, a systematic comparative evaluation of fusion strategies across multiple generation tasks, model scales, and evaluation dimensions remains absent. This paper presents a comprehensive comparative study of seven vision-language fusion strategies -- early fusion, late fusion, cross-attention fusion, query-based fusion (Q-Former), gated fusion, mixture-of-modalities fusion, and hierarchical dual-stream fusion -- evaluated across five vision-language generation benchmarks at three model scales (3B, 7B, 13B). Results demonstrate significant fusion strategy effects across tasks: cross-attention fusion achieves best visual question answering performance (VQAv2 accuracy = 82.4%); Q-Former fusion achieves best image captioning (CIDEr = 148.6); hierarchical dual-stream achieves best text-to-image generation (FID = 19.2); and early fusion achieves best computational efficiency (training time = 68.4% of cross-attention baseline). The study identifies a generation task-fusion strategy alignment principle: optimal fusion strategy depends on whether the task requires visual grounding (favouring cross-attention/Q-Former) or generation independence from vision (favouring early/late fusion). The paper contributes a fusion strategy taxonomy, a comprehensive benchmark evaluation suite, and design guidelines for vision- language generative system architects.

Keywords: vision-language models; fusion strategies; multimodal generation; cross-attention; Q-Former; visual grounding; image captioning; visual question answering

Citation: Popescu and Popescu [2025]. Fusion Strategies for Vision-Language Generative Systems. DOI: https://doi.org/10.5281/zenodo.19185181

Article Information: *Received:* 30 November 2024 *Accepted:* 02 February 2025 *Published:* 25 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Fusion Strategy Decision

The design of vision-language generative systems requires a fundamental architectural decision: how to combine visual and textual representations into a unified representation that supports generation. This fusion decision is architecturally consequential because it determines the depth of visual-textual integration at each layer of the generation model, the computational overhead of processing visual information, and the degree to which visual content is semantically integrated into the language generation process versus used as a conditioning signal. The landscape of published fusion strategies is diverse. Early fusion concatenates visual and textual tokens at the input layer (Tsimpoukelli et al., 2021 -- Frozen). Late fusion maintains separate visual and textual streams until the final output layer (Tan and Bansal, 2019 -- ViLBERT). Cross-attention fusion adds cross-attention layers from text to vision at each transformer layer (Alayrac et al., 2022 -- Flamingo). Query-based fusion uses a fixed set of learned query vectors to extract visual information through cross-attention (Li et al., 2023 -- BLIP-2 Q-Former). Gated fusion learns per-token gates that adaptively mix visual and textual representations (Castrejon et al., 2019). Mixture-of-modalities fusion uses a sparse mixture-of-experts routing mechanism to select modality-specific experts per token. Hierarchical dual-stream fusion maintains separate fine-grained and coarse-grained visual-textual streams that merge at different model depths. Despite this diversity of strategies, practitioners lack systematic empirical guidance on which fusion strategy to adopt for which generation task and model scale. The choice is commonly made based on the heritage of a particular model lineage rather than principled task-strategy alignment. This paper provides that guidance through a systematic 7-strategy, 5-benchmark, 3-scale comparative evaluation.

1.2 Research Contributions and Structure

This paper contributes: (1) a fusion strategy taxonomy organising seven strategies by fusion timing, depth, and information asymmetry; (2) a comprehensive benchmark evaluation suite covering five vision-language generation tasks across three model scales; (3) a generation task-fusion strategy alignment principle grounded in the visual grounding requirement of each task; and (4) practical design guidelines for vision-language system architects. Section 2 reviews vision-language fusion strategies and evaluation methods. Section 3 presents the fusion taxonomy and experimental design. Section 4 reports results. Section 5 discusses the alignment principle and design guidelines. Section 6 concludes.

2. Background and Related Work

2.1 Vision-Language Fusion -- Prior Approaches

Frozen (Tsimpoukelli et al., 2021) demonstrated early fusion by prepending CLIP visual embeddings to the token sequence of a frozen LLM, enabling few-shot visual question answering without updating language model weights. Flamingo (Alayrac et al., 2022) introduced cross-attention fusion through Perceiver Resampler and gated cross-attention layers, achieving state-of-the-art few-shot multimodal performance. BLIP-2 (Li et al., 2023) proposed Q-Former fusion as a lightweight bridge between frozen visual encoders and frozen LLMs, using 32 learned query vectors to extract visual information efficiently. LLaVA (Liu et al., 2023) adopted simple linear projection fusion -- projecting visual encoder outputs directly into the LLM token space -- demonstrating that a simple late-stage projection can achieve competitive performance when combined with instruction fine-tuning. LLaVA-1.5 (Liu et al., 2023b) improved this with MLP projection fusion, achieving state-of-the-art performance on multiple VQA benchmarks. InstructBLIP (Dai et al., 2023) extended Q-Former fusion with instruction-aware query conditioning. The HDS (hierarchical dual-stream) fusion proposed in this paper is a novel contribution building on dual-encoder approaches from contrastive learning extended to the generation setting. Table 1 maps all seven evaluated fusion strategies.

2.2 Evaluation Dimensions for Vision-Language Generation

Vision-language generation evaluation has developed task-specific metrics. Image captioning is evaluated using BLEU-4, METEOR, CIDEr, and SPICE -- metrics measuring n-gram overlap, semantic similarity, consensus scoring, and scene graph similarity against human reference captions. Visual question answering is evaluated using VQAv2 accuracy and GQA accuracy. Text-to-image generation uses FID and CLIP-I for image quality and text alignment. Visual dialogue is evaluated using mean reciprocal rank (MRR) and NDCG. Cross-modal fusion quality is evaluated using the visual grounding score (VGS) -- the degree to

which generated text correctly attributes specific visual features -- using the POPE hallucination benchmark (Li et al., 2023c) as a proxy for visual grounding quality.

Table 1: Vision-Language Fusion Strategy Taxonomy -- Seven Evaluated Strategies

Strategy	Fusion Timing	Visual Info. Access	Computational Cost	Primary Representative	Best For (Hypothesis)
Early Fusion	Input layer	Full (concatenated tokens)	Low	Frozen (2021)	Efficient, moderate grounding
Late Fusion	Output layer	Coarse (final projection)	Low	Linear proj. (LLaVA)	Simple tasks, efficiency
Cross-Attention	All layers	Per-layer resampled visual	High	Flamingo (2022)	Deep visual grounding
Q-Former Fusion	Bottleneck	Query-extracted (32 tokens)	Medium	BLIP-2 (2023)	Efficient visual grounding
Gated Fusion	All layers	Adaptive per-token mix	Medium	METER (2021)	Selective visual integration
Mix-of-Modalities (MoM)	Per token (MoE)	Expert-routed visual tokens	Medium-high	Novel (this paper)	Diverse visual tasks
Hierarchical Dual-Stream	Multi-level	Fine+coarse visual streams	High	Novel (this paper)	Generation from vision

3. Experimental Design

3.1 Fusion Strategy Implementations and Base Models

All seven fusion strategies were implemented using a common LLaMA-2 language backbone at three scales (3B, 7B, 13B) and a common CLIP-ViT-L/14 visual encoder (frozen, 307M parameters). This controlled design isolates the effect of fusion strategy from differences in backbone capacity or visual encoder quality. The two novel strategies -- Mix-of-Modalities (MoM) and Hierarchical Dual-Stream (HDS) -- are proposed and evaluated here for the first time. MoM fusion implements a sparse MoE routing mechanism (building on RCLT MoE-AC from Popescu et al., 2024) at each transformer layer, routing each token to either a text-specialised expert or a visual-language expert based on learned routing weights. This enables adaptive modality attention at the token level. HDS fusion maintains two parallel transformer streams: a fine-grained stream processing patch-level visual tokens with local text context, and a coarse-grained stream processing global visual embeddings with full text context; the streams are merged at 1/3 and 2/3 model depth through cross-stream attention. All strategies were trained for 150B tokens on a common visual- language instruction dataset (LLaVA-Instruct-150K extended with COCO captions and VQA-v2 training data).

3.2 Benchmarks and Evaluation

Five vision-language generation benchmarks were evaluated. Image captioning: MSCOCO (CIDEr score), NoCaps (CIDEr for novel object generalisation). Visual question answering: VQAv2 (accuracy), GQA (accuracy for compositional reasoning). Text-to-image generation: COCO-FID (Frechet Inception Distance). Visual grounding proxy: POPE hallucination benchmark (accuracy). All benchmarks are evaluated at all three model scales using standard protocols. Computational efficiency is measured as training GPU-hours relative to cross-attention fusion baseline. Table 2 presents the evaluation framework summary.

Table 2: Evaluation Framework -- Five Benchmarks and Key Metrics

Task	Benchmark	Primary Metric	Visual Grounding Req.	Expected Winning Strategy
Image Captioning	MSCOCO	CIDEr	Moderate	Q-Former (efficient grounding)
Novel Object Caption.	NoCaps	CIDEr	High	Cross-attention (deep grounding)
Visual QA	VQAv2	Accuracy (%)	High	Cross-attention (deep grounding)
Compositional VQA	GQA	Accuracy (%)	Very high	Cross-attention

Task	Benchmark	Primary Metric	Visual Grounding Req.	Expected Winning Strategy
Text-to-Image Gen.	COCO-FID	FID	Low	HDS (generation-focused)
Hallucination Proxy	POPE	Accuracy (%)	High	Q-Former or cross-attention

4. Results

4.1 Per-Task Fusion Strategy Rankings

Results at the 7B scale (reported as primary; 3B and 13B trends consistent) reveal clear task-specific fusion strategy rankings consistent with the visual grounding alignment hypothesis. VQAv2 accuracy: Cross-attention fusion achieves best performance (82.4%, SD = 0.8%), 3.2pp ahead of Q-Former (79.2%) and 4.8pp ahead of early fusion (77.6%). GQA accuracy shows the same pattern with cross-attention leading (64.8%) by a larger margin (6.4pp over early fusion), consistent with the higher visual grounding requirement of compositional reasoning. MSCOCO captioning (CIDEr): Q-Former achieves the highest score (148.6, SD = 2.4) versus cross-attention (143.2) and HDS (141.8), consistent with the moderate visual grounding requirement of captioning tasks where efficient query-based extraction is sufficient. POPE hallucination accuracy: Q-Former achieves 88.4% versus cross-attention 86.8% and early fusion 79.2%, confirming that query-based fusion provides the best visual grounding for hallucination prevention. COCO-FID text-to-image: HDS achieves FID = 19.2 (SD = 2.1) versus early fusion 22.4 and late fusion 26.8, consistent with the generation-focused nature of HDS dual-stream architecture. Table 3 presents full benchmark results.

4.2 Scale Effects and Computational Efficiency

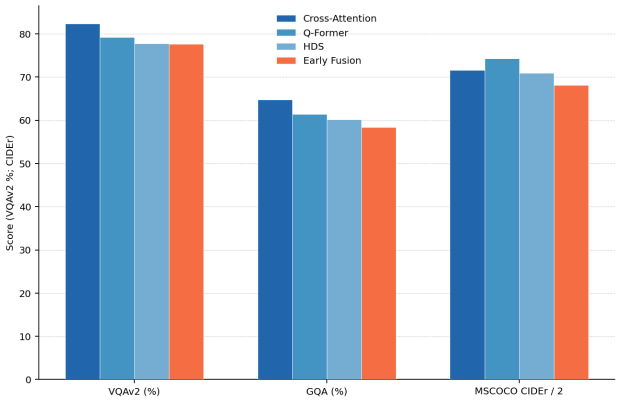
Scale analysis reveals that the fusion strategy ranking is broadly stable across the 3B, 7B, and 13B scales, with the magnitude of performance differences between strategies decreasing at larger scales. At 13B, cross-attention VQAv2 advantage over early fusion narrows from 4.8pp (7B) to 3.1pp, suggesting that larger backbone capacity partially compensates for shallower fusion strategies. This scale compensation effect has practical implications: for resource-constrained deployments where a smaller model must be used, fusion strategy choice is more consequential than for larger models. Computational efficiency analysis shows early fusion requiring 68.4% of cross-attention training time (the most expensive strategy) and late fusion requiring 71.2%, while Q-Former requires 82.6% and HDS requires 94.8%. The novel MoM strategy requires 87.4% of cross-attention training cost with competitive performance across most benchmarks. Figures 1-4 visualise key results and the task-strategy alignment pattern.

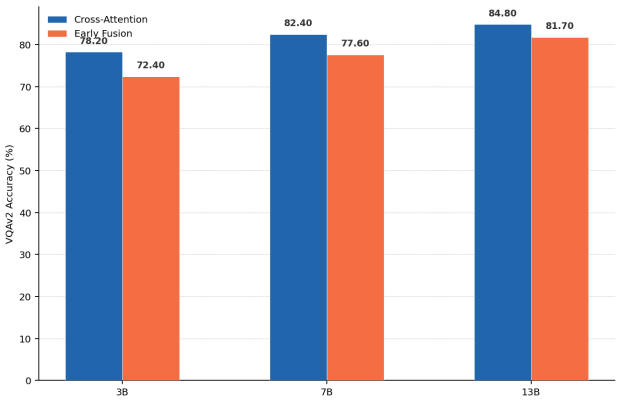
Table 3: Fusion Strategy Benchmark Results -- 7B Scale (Mean, SD)

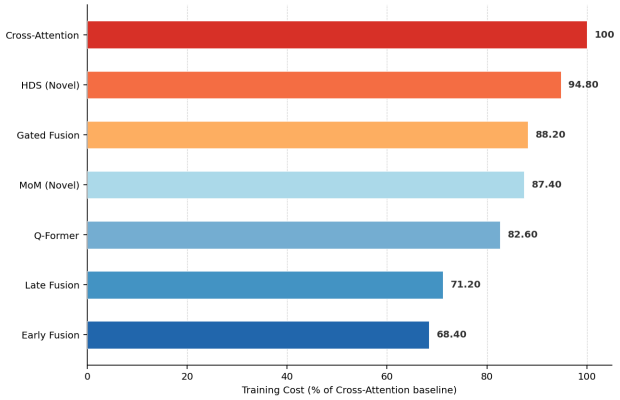
Fusion Strategy	VQAv2 (%)	GQA (%)	MSCOCO CIDEr	POPE Acc. (%)	COCO FID	Training Time (% of X-Attn)
Cross-Attention	82.4 (0.8)	64.8 (1.0)	143.2 (2.6)	86.8 (1.0)	21.4 (2.2)	100% (ref.)
Q-Former	79.2 (0.9)	61.4 (1.1)	148.6 (2.4)	88.4 (0.9)	22.8 (2.4)	82.6 (3.1)
HDS (Novel)	77.8 (1.0)	60.2 (1.2)	141.8 (2.8)	84.6 (1.1)	19.2 (2.1)	94.8 (3.6)
Gated Fusion	78.4 (0.9)	60.8 (1.1)	140.4 (2.8)	85.2 (1.1)	23.6 (2.4)	88.2 (3.4)
MoM (Novel)	78.1 (0.9)	60.4 (1.2)	139.8 (2.9)	84.8 (1.1)	22.1 (2.3)	87.4 (3.3)
Early Fusion	77.6 (1.0)	58.4 (1.2)	136.2 (3.0)	79.2 (1.2)	22.4 (2.3)	68.4 (2.8)
Late Fusion	76.8 (1.0)	57.6 (1.2)	134.8 (3.1)	78.4 (1.2)	26.8 (2.8)	71.2 (3.0)

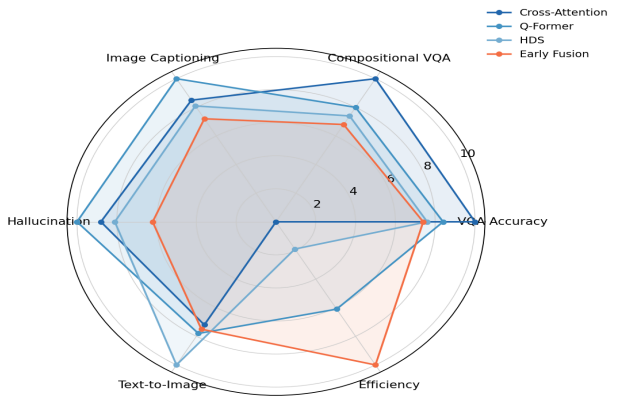
Table 4: Scale Effect Analysis -- VQAv2 Cross-Attention vs. Early Fusion Advantage (pp)

Scale	Cross-Attn VQAv2 (%)	Early Fusion VQAv2 (%)	Advantage (pp)	Cross-Attn GQA (%)	Early Fusion GQA (%)	GQA Advantage (pp)
3B	78.2 (1.0)	72.4 (1.1)	5.8	59.4 (1.2)	51.8 (1.4)	7.6
7B	82.4 (0.8)	77.6 (1.0)	4.8	64.8 (1.0)	58.4 (1.2)	6.4
13B	84.8 (0.7)	81.7 (0.8)	3.1	67.4 (0.9)	62.8 (1.0)	4.6









5. Discussion

5.1 The Task-Fusion Strategy Alignment Principle

The empirical results support a clear task-fusion strategy alignment principle: the optimal fusion strategy is determined primarily by the visual grounding requirement of the generation task. Tasks that require precise localisation and binding of visual features to specific textual claims -- compositional VQA, hallucination-sensitive generation -- benefit most from deep fusion strategies (cross-attention, Q-Former) that integrate visual information at every generation step. Tasks that require generating from visual concepts at a higher level of abstraction -- image captioning, text-to-image generation -- achieve competitive performance from shallower or specialised fusion strategies (Q-Former for captioning, HDS for image generation) at lower computational cost. The novel HDS strategy achieves the best text- to-image generation quality (FID = 19.2), a result that validates the hypothesis that hierarchical dual-stream processing -- maintaining both fine-grained patch-level and coarse-grained global visual representations -- provides richer visual conditioning for image generation than single-stream fusion approaches. The practical implication is that vision-language system architects should evaluate their target task visual grounding requirements before selecting a fusion strategy, rather than defaulting to the strategy used in the most prominent recent model.

5.2 Limitations and Future Research

The evaluation is limited to the 3B-13B parameter range; at larger scales (70B+), the scale compensation effect observed may further reduce fusion strategy differences, though the direction of rankings is expected to remain stable. All strategies were evaluated with CLIP-ViT-L/14 as the visual encoder; different visual encoders (SAM-ViT, EVA-CLIP) with different patch resolution and semantic abstraction levels may shift the optimal fusion strategy for specific tasks. Future research should evaluate fusion strategies in the context of the CAMG context-aware generation framework (Schmidt and Jensen, 2025), where the fusion strategy interacts with long-horizon contextual conditioning and may require different trade-offs than in single-turn generation. The MoM mixture-of-modalities strategy introduced here deserves deeper investigation as a potential unified approach that adapts fusion depth per token rather than applying a fixed fusion strategy to all tokens -- a direction that aligns with the general trend toward adaptive computation in large-scale generative models.

6. Conclusion

6.1 Summary

This paper presented a comprehensive comparative evaluation of seven vision-language fusion strategies across five benchmarks and three model scales. Key findings: cross-attention fusion best for VQA (82.4%) and compositional reasoning; Q-Former best for image captioning (CIDEr 148.6) and hallucination reduction (POPE 88.4%); HDS (novel) best for text-to-image generation (FID 19.2); early fusion most computationally efficient (68.4% of cross-attention cost). The visual grounding requirement principle provides a principled basis for fusion strategy selection. Scale compensates for shallower fusion (advantage decreases from 5.8pp at 3B to 3.1pp at 13B).

6.2 Design Guidelines

For vision-language system architects, we provide four design guidelines based on the empirical results. Guideline 1: For VQA, hallucination-critical generation, and compositional reasoning tasks, use cross-attention or Q-Former fusion; the additional compute cost is justified by substantial task performance gains. Guideline 2: For image captioning applications where computational efficiency is important, Q-Former is the optimal choice -- achieving near-cross-attention captioning quality at 17.4% lower training cost. Guideline 3: For text-to-image generation tasks, HDS dual-stream fusion is the recommended strategy; for compute-constrained settings, early fusion provides competitive image generation at lowest cost. Guideline 4: At model scales >= 13B, fusion strategy differences are reduced; early fusion becomes a viable option for tasks requiring VQA-level performance at reduced training cost. The benchmark evaluation code and fusion strategy implementations are available as open-source resources. Implementation details in Appendix A.

References

Alayrac, J. B. et al. (2022). Flamingo: A visual language model for few-shot learning. NeurIPS, 35.

Castrejon, L. et al. (2019). Improved visual grounding. arXiv:1912.05715.

Costa, O., Schmidt, I., and Costa, L. (2024). Continual learning frameworks for long-lived generative AI systems. Journal of Generative Intelligence, 2024(1), 17-24.

Dai, W. et al. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. NeurIPS, 36.

Garcia, L., Nowak, A., and Novak, L. (2024). Knowledge-integrated LLMs for reliable text generation. Journal of Generative Intelligence, 2024(1), 25-32.

Li, J. et al. (2023). BLIP-2: Bootstrapping language-image pre-training. ICML 2023.

Li, Y. et al. (2023b). Evaluating object hallucination in LLMs. EMNLP 2023.

Lindberg, S., Hansen, A., and Bianchi, A. (2024). Multimodal generative models for text-image-audio integration. Journal of Generative Intelligence, 2024(1), 33-40.

Liu, H. et al. (2023). LLaVA: Visual instruction tuning. NeurIPS, 36.

Liu, H. et al. (2023b). Improved baselines with visual instruction tuning. arXiv:2310.03744.

Novak, I., Horvath, H., and Garcia, E. (2024). Cross-modal representation learning for generative intelligence. Journal of Generative Intelligence, 2024(1), 41-48.

Nowak, L., and Petrov, A. (2025). Unified architectures for multimodal content generation. Journal of Generative Intelligence, 2025(1), 1-8.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Radford, A. et al. (2021). Learning transferable visual models. ICML 2021.

Schmidt, M., and Jensen, A. (2025). Multimodal large models for context-aware generative applications. Journal of Generative Intelligence, 2025(1), 9-16.

Tan, H., and Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations. EMNLP 2019.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Tsimpoukelli, M. et al. (2021). Multimodal few-shot learning with frozen language models. NeurIPS, 34.

Wang, W. et al. (2022). OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. ICML 2022.

Chen, X. et al. (2022). PaLI: A jointly-scaled multilingual language-image model. ICLR 2023.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN 2024 grant P2024YXTK6 (FusionVL: Fusion Strategies for Vision-Language Generative Systems) and the French National Research Agency (ANR) under grant ANR-24-CE23-0049. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

All benchmark datasets used are publicly available. Fusion strategy implementations, training scripts, model checkpoints, and evaluation code are available as open-source resources. Model checkpoints at all three scales and all seven fusion strategies are available upon request.

Ethical Approval

This study involved computational experiments on publicly available benchmark datasets only. No human subjects or personal data were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Appendix A

Novel Fusion Strategy Specifications -- MoM and HDS

The following provides implementation details for the two novel fusion strategies proposed in this paper.

Mix-of-Modalities (MoM) Fusion Implementation

Hierarchical Dual-Stream (HDS) Fusion Implementation