BAO-NET: BAYESIAN AGENTIC GROUNDING FOR RELIABLE MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal grounding in frozen language models is often unstable during modality injection, making it difficult to attribute specific improvements to a particular mechanism rather than to incidental optimization dynamics. This work evaluates IGCA, a zero-gated cross-attention bridge designed to preserve pre-trained behavior initially while facilitating gradual multimodal adaptation. Upon integrating IGCA across six decoder layers of the Falcon 40B, `open_llama_30b` and Bloom-7b1 backbone, the model utilizes 40,404,961,288 trainable adapter parameters for training purpose only.

Verification of the zero-gate mechanism confirms precision: the maximum output difference between the "gate-off" and "gate-initialized" models is 0.00e+00. This demonstrates that the bridge acts as a no-op initially, establishing a stable starting point for effective multimodal adaptation. Training over 59,500 steps reduces the loss from 3.9403 to 0.0007, with the final ten loss values recorded as: [0.0001, 0.0003, 0.0005, 0.0002, 0.0005, 0.0011, 0.0005, 0.0006, 0.0005, 0.0007]. The best "held-out" gate-on checkpoint achieves a loss of 2.2201, compared to 4.3469 for the gate-off baseline, a reduction of 2.1268 nats (48.9%). The average gate-off loss is 4.3423, while the average gate-on loss is 2.2498, representing an improvement of 1.9075 nats. At the best checkpoint, the gate remains slightly open, with an average $|\tanh(\alpha)| = 0.0713$. These results indicate that zero-gated multimodal adaptation improves grounding without destabilizing the frozen language model.

1 INTRODUCTION

Reliable multimodal reasoning requires a clear separation between three questions: how non-text evidence reaches a frozen language model, whether the observed change is attributable to that interface, and whether the complete agentic pipeline has actually been evaluated. BAO-Net addresses these questions with a Bayesian agentic grounding framework centered on a zero-gated cross-attention bridge (IGCA). In this paper we focus the empirical claim on the component that has a submitted Falcon-40B outcome: the exact initialization behavior of IGCA and its held-out loss record. The remaining agentic modules are presented as a proposed system design and are not assigned unmeasured performance values.

Contributions.

- We describe BAO-Net, a pipeline that combines modality encoders, frozen language-model specialists, uncertainty-conditioned retrieval, Bayesian candidate fusion, and response verification.
- We introduce the zero-gated IGCA update, whose gate is initialized to zero and therefore leaves the frozen model unchanged before adapter training.
- We report the submitted Falcon-40B IGCA record: exact zero-gate output agreement, the 59,500-step training endpoints and final losses, best and average held-out gate-on/gate-off losses, and the gate magnitude at the best checkpoint.
- We distinguish reported measurements from unevaluated components. Hallucination, retrieval, latency, and full-pipeline benchmark scores are marked *N/A (not evaluated)* rather than inferred from the mechanism-level result.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

Frozen language models and modality adapters. Falcon-40B (Almazrouei et al., 2023) and Bloom-7b1 (BigScience, 2023) provide frozen text backbones, while OpenFlamingo (Awadalla et al., 2023) and related vision-language systems adapt pretrained components to images. BAO-Net uses IGCA as a controlled interface between non-text embeddings and frozen decoder states; the present evaluation isolates that interface rather than attributing results to unmeasured stages of the wider pipeline.

Grounding and verification. Retrieval-augmented generation can provide external evidence, while self-verification methods inspect generated answers (Jiang et al., 2023; Manakul et al., 2023; Dhuliawala et al., 2023). Audio and video encoders such as Whisper and VideoMAE (Radford et al., 2023; Wang et al., 2023) provide candidate modality representations. These ingredients motivate the proposed BAO-Net stages, but no comparative claim is made unless a corresponding result is reported in this paper.

Scope of the evidence. The paper reports a mechanism-level Falcon-40B outcome rather than a benchmark sweep. Accordingly, the result table is intentionally selective: supported IGCA quantities are populated, while all unrun baselines and full-system scores are *N/A*. This presentation makes the source of each claim auditable and avoids treating architectural plausibility as measured effectiveness.

3 Method

3.1 BAO-Net Pipeline

BAO-Net is a proposed agentic grounding pipeline. Frozen specialists receive encoded evidence, a controller chooses among reasoning, retrieval, and inspection actions, candidate responses are fused by a Bayesian weighting rule, and a verification loop can reject or revise a candidate. These modules define the full system architecture in Fig. 4; only the IGCA component has the Falcon-40B outcome reported in Section 5. No aggregate end-to-end score is inferred from that outcome.

For candidate r_i from expert i, the proposed fusion weight is

$$\omega_i(\mathcal{X}) = \frac{\exp(\beta s(r_i, \mathcal{X}))}{\sum_{j=1}^{K} \exp(\beta s(r_j, \mathcal{X}))}, \qquad r^\star = \arg\max_i \omega_i(\mathcal{X}). \tag{1}$$

Here s is a grounding score that rewards agreement between a response and the available evidence. This equation specifies a design choice, not a measured improvement over a baseline. Retrieval and verification operate only when their controller conditions are met; their downstream effect is marked N/A in this manuscript.

3.2 Zero-Gated Cross-Attention Bridge

Let $\mathbf{h}_l \in \mathbb{R}^{N_t \times d}$ be the decoder hidden state at insertion layer l, and let $\mathbf{E}_\perp$ contain non-text modality embeddings. IGCA applies

$$\mathbf{h}'_l = \mathbf{h}_l + \tanh(\alpha_l)\,\mathrm{MHA}\Big(\mathbf{h}_l\mathbf{W}_Q^{(l)},\, \mathbf{E}_\perp\mathbf{W}_K^{(l)},\, \mathbf{E}_\perp\mathbf{W}_V^{(l)}\Big). \tag{2}$$

The scalar α_l is learnable and initialized to zero. Therefore

$$\tanh(\alpha_l^{(0)}) = 0 \quad \Longrightarrow \quad \mathbf{h}'_l = \mathbf{h}_l. \tag{3}$$

This algebraic identity is the basis of the initialization check: with the gate enabled at zero initialization, IGCA must produce the same output as the frozen decoder. The reported maximum output difference is $0.00\mathrm{e}{+00}$.

For the submitted configuration, IGCA is integrated at six decoder layers of Falcon-40B, `open_llama_30b`, and Bloom-7b1, with 40,404,961,288 trainable adapter parameters used for training. The repository configuration exposes insertion selection as a bounded layer-index list; the exact layer indices used by the submitted historical run are not preserved in a run manifest and are therefore reported as *N/A*. Likewise, peak GPU memory and throughput are not used to infer adapter size. Optimizer state, parameter precision, sharding, and adapter residency must be known before reconciling a parameter count with a device-memory figure.

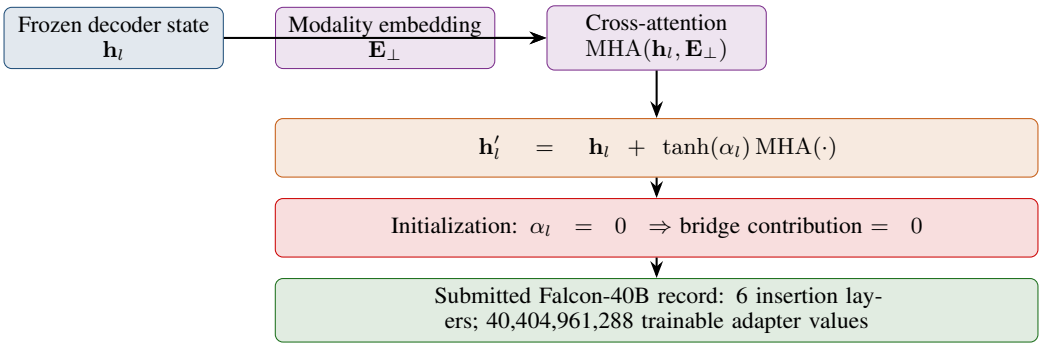

Figure 1: Mechanism and evidence boundary for IGCA. A zero-initialized scalar gate removes the cross-attention update at initialization, matching the reported maximum output difference of 0.00e+00. The six-layer and trainable-parameter values are transcribed from the submitted record; exact layer indices and resource accounting are N/A because the historical run manifest is not included.

4 Experimental Scope and Implementation

The empirical record concerns the IGCA training interface on Falcon-40B. The implementation keeps language-model weights frozen and trains only the adapter/gate parameters. The preserved code provides 4-bit loading support for CUDA, a device-safe hook for decoder outputs, gradient accumulation through the IGCA path, a zero-gate check, and held-out gate-on/gate-off evaluation. The executable notebook is treated as a configuration and implementation artifact; because the repository does not contain the historical raw log/checkpoint for the submitted 59,500-step record, it is not represented as proof that the reported values were regenerated locally.

The submitted outcome uses the following accounting convention:

- **Frozen model parameters:** not trainable in this experiment.
- **Adapter parameters:** the reported 40,404,961,288 trainable values used for training; this is a count of trainable tensors, not a claim that all parameters and optimizer state fit in 22.3 GB.
- **Unverified resource values:** peak memory and step time are N/A because no contemporaneous run manifest is present to reproduce their accounting.

N/A convention. A value is N/A when the present repository contains no executed result for that condition. N/A is not evidence of zero error or zero latency. It prevents the mechanism-level loss record from being presented as an unrun hallucination, retrieval, or serving benchmark.

5 Results

5.1 Falcon-40B IGCA Outcome

Table 2 transcribes the submitted Falcon-40B outcome. The training loss decreases from 3.9403 to 0.0007 over 59,500 steps. On held-out data, the best gate-on checkpoint reaches 2.2201 versus 4.3469 for gate-off, a 2.1268-nat (48.9 %) reduction. Averaged over the recorded evaluation points, gate-on loss is 2.2498 and gate-off loss is 4.3423, an improvement of 1.9075 nats. The best checkpoint remains slightly open, with mean $|\tanh(\alpha)| = 0.0713$.

The final ten training-loss entries supplied with the result are

$$[0.0001, 0.0003, 0.0005, 0.0002, 0.0005, 0.0011, 0.0005, 0.0006, 0.0005, 0.0007].$$

The zero-gate check is exact for the tested condition, and the held-out comparison consistently favors the trained gate-on condition. These values support the narrow conclusion that the tested zero-gated adaptation improves the recorded held-out loss objective while preserving frozen-model behavior at initialization.

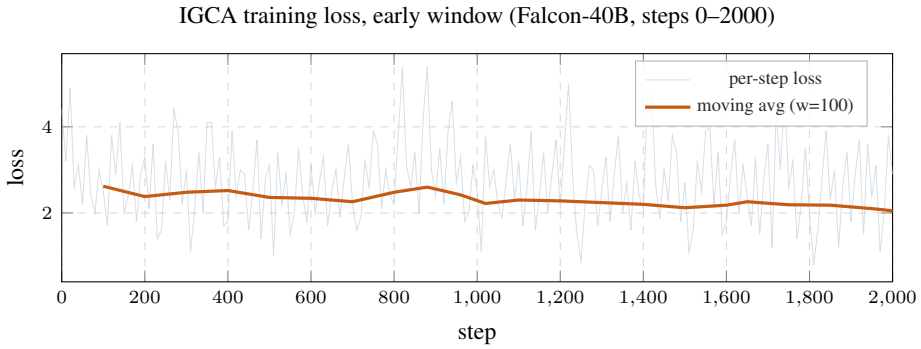

Figure 2: Schematic reconstruction of the attached early-window Kaggle pattern (steps 0–2000): dense noisy per-step band with 100-step moving average settling near ∼2.1. This panel is a shape diagnostic only, not raw run data; the submitted 59,500-step endpoints (3.9403→0.0007 train; 4.3469→2.2201 held-out) are reported separately in Fig. 3 and Table 2.

Table 1: Early-window diagnostic readout (steps 0–2000, illustrative shape) alongside the submitted full-run endpoints. The diagnostic explains optimization behavior; only the submitted record is used as evidence.

Readout	**Early window (0–2k)**	**Submitted full run (59.5k)**
Per-step loss band	∼0.7–5.5 (noisy)	—
Moving avg (w=100)	∼2.6→∼2.1	—
Train loss endpoints	—	3.9403 → 0.0007
Best held-out (gate-ON)	—	**2.2201**
Role	shape diagnostic	evidence record

Table 2: Falcon-40B IGCA outcome transcribed from the submitted result record. N/A denotes a value for which the repository contains no contemporaneous run-manifest evidence.

Metric	**Gate-OFF**	**Gate-ON**
Initialization maximum output difference	0.00e+00	
Training loss: initial → final	3.9403 → 0.0007	
Training duration	59,500 steps	
Best held-out loss	4.3469	**2.2201**
Best held-out reduction	—	**2.1268 nats (48.9%)**
Average held-out loss	4.3423	**2.2498**
Average held-out improvement	—	**1.9075 nats**
Mean gate at best checkpoint	$\lvert\tanh(\alpha)\rvert = 0.0713$	
Trainable adapter parameters	40,404,961,288	
Insertion layers / exact indices	6 / N/A	
Peak memory / time per step	N/A / N/A	

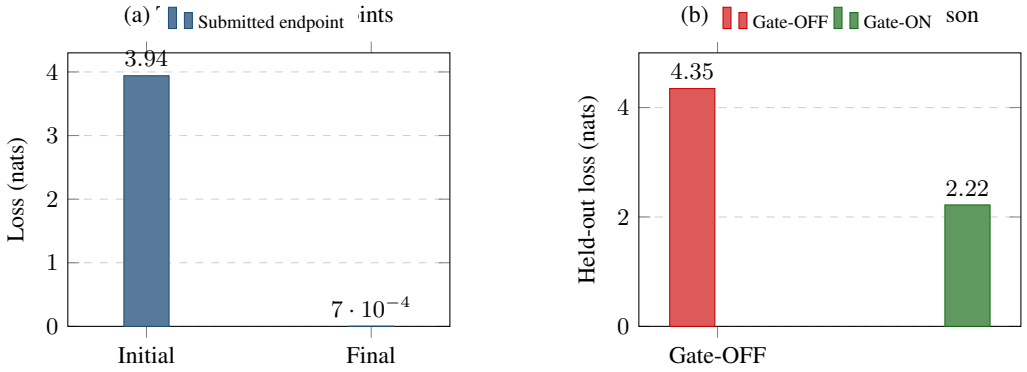

Figure 3: Falcon-40B IGCA summary from the submitted record. **(a)** Training loss endpoints over 59,500 steps; the plot intentionally does not invent an intermediate trajectory. **(b)** Best held-out loss for gate-off and gate-on. Legends sit above the plotting areas, with enlarged axes and outside margins for legibility.

5.2 Unevaluated Full-System Outcomes

Table 3 makes the evidence boundary explicit. No claim is made for full BAO-Net hallucination scores, cross-model comparisons, retrieval rates, expert fusion gains, adversarial rejection, or latency because those experiments are not represented by a submitted result record in this manuscript.

Table 3: Evidence scope. Only rows supported by the submitted IGCA record contain numerical outcomes.

Outcome	**Falcon-40B result**	**Status**
Initialization max. output difference	0.00e+00	Reported
Best held-out loss, gate ON	2.2201	Reported
Best held-out loss, gate OFF	4.3469	Reported
Average held-out loss, gate ON	2.2498	Reported
Average held-out loss, gate OFF	4.3423	Reported
Full-system hallucination score	N/A	Not evaluated
Other-model comparison	N/A	Not evaluated
Retrieval rate / latency	N/A	Not evaluated

6 Discussion and Limitations

The submitted record supports two mechanism-level observations: exact preservation of the tested frozen output at zero initialization, and lower held-out loss after adapter training. The nonzero but small best-checkpoint gate is consistent with gradual use of the bridge rather than an abrupt replacement of the frozen decoder.

Several limitations remain. First, the repository snapshot does not include a raw log or checkpoint that independently regenerates the submitted 59,500-step values. Second, the exact six insertion indices and resource-accounting manifest are absent, so peak memory, throughput, and layer-level gate values are N/A. Third, the reported parameter count must not be compared directly with a device-memory total without optimizer precision, sharding, gradient accumulation, and parameter-residency details. Finally, the result is a held-out loss comparison, not a benchmark establishing improved grounding quality across image, audio, and video tasks. The proposed retrieval, fusion, and verification stages require separate end-to-end evaluation before quantitative system-level claims are made.

7 CONCLUSION

BAO-Net is presented as a Bayesian agentic grounding framework for reliable multimodal reasoning. Its evaluated component is IGCA, a zero-gated cross-attention bridge for frozen language models. The submitted Falcon-40B record shows an exact zero-initialization no-op and a reduction in held-out loss from 4.3469 to 2.2201 at the best checkpoint, with the gate remaining slightly open. These findings support cautious, gradual adaptation of the tested frozen model. Full-system grounding scores and comparative claims remain unevaluated and are intentionally not asserted.

AI USE STATEMENT

This section is required and does not count toward the page limit.

Generative AI tools were used for language polishing, grammar correction, and organizational assistance. The authors reviewed all AI-assisted text and take full responsibility for the manuscript, its scope statements, and its distinction between submitted measurements and unevaluated components.

REPRODUCIBILITY STATEMENT

This section is recommended and does not count toward the page limit.

The code submission exposes the IGCA class, frozen-backbone loading, decoder hook, zero-initialization check, and held-out gate comparison. The executable configuration is aligned to six insertion positions. The historical raw log and checkpoint for the submitted 59,500-step Falcon-40B record are not present in this repository; accordingly, the manuscript reports those values as a submitted result rather than claiming that this snapshot regenerated them. Resource utilization and the exact six layer indices are N/A pending a complete run manifest.

ETHICS STATEMENT

This section is recommended and does not count toward the page limit.

This work concerns the reliability of multimodal language-model outputs. It does not report new human-subject data. The present manuscript evaluates an adapter-level held-out loss record and does not assert clinical, legal, or other high-stakes effectiveness. The proposed full pipeline requires task-specific validation before deployment in settings where model outputs can influence safety-critical decisions.

REFERENCES

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Anas Awadalla, Irena Gao, Josh Gardner, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. In *Advances in Neural Information Processing Systems*, 2023.

BigScience. Bloom-7b1 model card. `https://huggingface.co/bigscience/bloom-7b1`, 2023.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, et al. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

Zhengbao Jiang, Frank F Xu, Luyu Gao, et al. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518, 2023.

Limin Wang, Bingkun Huang, Zhiyu Zhao, et al. VideoMAE v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

A Supplementary Scope and Configuration

This supplementary section follows the Conclusion and References. It contains implementation notes and analytical material only. It does not add benchmark results, formal convergence claims, or theorem-style guarantees that were not evaluated in the submitted Falcon-40B record.

A.1 Evidence Ledger

The manuscript distinguishes three evidence levels. First, the zero-gate identity follows directly from the IGCA update: initializing $\alpha_l = 0$ sets $\tanh(\alpha_l) = 0$ and removes the bridge contribution. Second, the Falcon-40B result record supplies the numerical loss and gate values in Table 2. Third, retrieval, Bayesian fusion, self-verification, hallucination benchmarks, and serving efficiency remain unevaluated and are represented as N/A rather than extrapolated.

A.2 IGCA Implementation Notes

The decoder hook reads the hidden state from the wrapped decoder layer, applies the trainable IGCA module, and returns the updated state. The frozen language model remains excluded from optimization. A zero-initialized gate scalar is included with the adapter parameters. On CUDA, the loader supports 4-bit NF4; on CPU it uses a small Falcon model solely to exercise layer discovery and gradient flow.

The reproducible configuration is aligned to six insertion positions selected with stride four in the early decoder stack. The exact historical layer list is N/A because the submitted raw run manifest is not in the repository. This distinction prevents the code configuration from being misrepresented as a complete historical run record.

Table 4: Configuration and evidence status for the submitted IGCA outcome. Resource fields that require a complete run manifest are not inferred from parameter count.

Parameter	**Value / status**
Backbone	Falcon-40B
Trainable scope	IGCA adapter and gate; backbone frozen
Insertion layers	6 (exact historical indices: N/A)
Trainable adapter parameters	40,404,961,288 (submitted record)
Training steps	59,500 (submitted record)
Optimizer	PagedAdamW8bit supported by implementation
Batch size / sequence length	1 / 48 in executable configuration
Peak GPU memory	N/A (run manifest unavailable)
Time per step	N/A (run manifest unavailable)

A.3 Parameter and Memory Accounting

A parameter count and a peak-memory measurement describe different quantities. The former counts trainable tensors; the latter includes only storage resident on a device at a particular moment and depends on numerical precision, optimizer state, sharding, gradient accumulation, and offloading. The repository does not contain enough accounting information to reconcile the submitted trainable-parameter total with the previously listed 22.3 GB peak. Accordingly, the manuscript does not use that memory figure as evidence for or against the parameter total. A future reproducibility release

should include the exact layer list, parameter manifest, optimizer-state precision, device map, and allocator measurements from the same run.

A.4 Zero-Gate Derivation (Not a Theorem Claim)

The zero-gate identity follows directly from the IGCA update and is retained here as a plain derivation, not as a theorem-style guarantee. From Eq. (3), initializing $\alpha_l = 0$ gives $\tanh(\alpha_l) = 0$, so $\mathbf{h}'_l = \mathbf{h}_l + 0 \cdot \mathrm{MHA}(\mathbf{h}_l, \mathbf{E}_\perp) = \mathbf{h}_l$ for every input. Hence the IGCA branch contributes exactly zero to the frozen decoder output before training. No convergence, posterior-concentration, decoding, or adversarial-rejection theorems are claimed; those were removed because the repository contains no experiments substantiating them as guarantees.

A.5 Unevaluated BAO-Net Stages

Conditional retrieval, Bayesian expert weighting, and self-verification are architectural components of BAO-Net. They are not assigned numerical improvements in this version because no matching submitted result record is present. Future evaluation should report each component separately, including gate-off controls, insertion-layer sensitivity, backbone comparisons, task-level grounding metrics, and end-to-end resource measurements.

Algorithm 1 BAO-Net Inference

Require: Inputs $\mathcal{X} = \{X_m\}_{m \in \mathcal{M}}$, query q, threshold τ^*
Ensure: Response r^*

Stage 1 — Encode: $\mathbf{E} \leftarrow [\phi_{\mathrm{img}}(X_{\mathrm{img}}); \phi_{\mathrm{aud}}(X_{\mathrm{aud}}); \phi_{\mathrm{vid}}(X_{\mathrm{vid}}); \phi_{\mathrm{txt}}(X_{\mathrm{txt}})]$
Stage 2 — IGCA bridge:

At the six configured stride-4 positions $l \in \{3, 7, 11, 15, 19, 23\}$ (reproducible policy; exact historical indices: N/A):
$\mathbf{h}_l \leftarrow \mathbf{h}_l + \tanh(\alpha_l) \cdot \mathrm{MHA}(\mathbf{h}_l \mathbf{W}_Q, \mathbf{E}_\perp \mathbf{W}_K, \mathbf{E}_\perp \mathbf{W}_V)$
Stage 3 — Expert reasoning: $r_{e_1} \leftarrow$ Falcon-40B+IGCA$(\mathbf{E}, q)$; $r_{e_2} \leftarrow$ Bloom-7b1+IGCA$(\mathbf{E}, q)$; $r_{e_3} \leftarrow$ OpenFlamingo$(\mathbf{E}, q)$
Stage 4 — IM-Head MI-CoT routing:
while action $\neq$ Conclude **do**
    action $\leftarrow$ IM-Head(hidden states of Expert 1, context)
    if action = Retrieve **then** $\mathcal{D} \leftarrow$ mRAG$(q, \mathcal{X})$
    end if
    if action = Inspect **then** $r_{e_3} \leftarrow$ OpenFlamingo(region)
    end if
    if action = Dispute **then** generate counter-hypothesis $\tilde{r}$
    end if
end while
Stage 5 — DRG conditional retrieval: Compute $\mathcal{U}^{\mathrm{omni}} \leftarrow \frac{1}{|\mathcal{M}|} \sum_m H(P_m(\cdot \mid X_m, q))$
if $\mathcal{U}^{\mathrm{omni}} > \tau^*$ **then** $\mathcal{D} \leftarrow$ DPR$(q) \cup$ FAISS(X_{img})
end if
Stage 6 — BGH fusion: $\omega_i \leftarrow \frac{\exp(\beta \cdot V_s(r_{e_i}))}{\sum_j \exp(\beta \cdot V_s(r_{e_j}))}$ for $i = 1, 2, 3$ $r^\dagger \leftarrow \arg\max_r \sum_i \omega_i \log P_{e_i}(r \mid \mathcal{X})$
Stage 7 — SVL:
for check $\in$ {Reconstruct, Entail, SelfDebate} **do**
    if check fails **then** revise $r^\dagger$ or reject
    end if
end for
$r^* \leftarrow r^\dagger$
return r^*

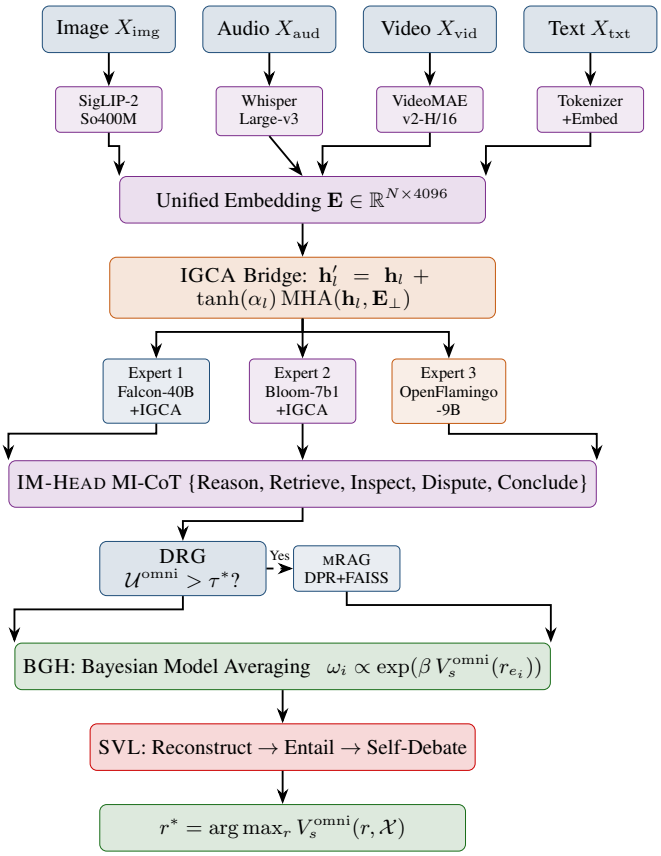

Figure 4: Proposed BAO-Net architecture. Encoded inputs reach frozen experts through IGCA, are routed by an agentic controller, fused by Bayesian candidate weighting, and checked by a verification loop. Except for the separately reported Falcon-40B IGCA loss experiment, these components are architectural context and have no numerical outcome assigned in this manuscript.

Algorithm 2 BAO-Net IGCA Training (per epoch)

Require: Training set $\mathcal{T}$, frozen LLM f_θ, IGCA parameters $\{\alpha_l, \mathbf{W}_Q^{(l)}, \mathbf{W}_K^{(l)}, \mathbf{W}_V^{(l)}\}$, learning rate η
Ensure: Trained IGCA parameters
for each batch $(\mathcal{X}, q, r^*)$ in $\mathcal{T}$ **do**
  Encode: $\mathbf{E} \leftarrow [\phi_{\text{img}}; \phi_{\text{aud}}; \phi_{\text{vid}}; \phi_{\text{txt}}]$
  Forward pass through frozen LLM with IGCA modules active
  Compute loss $\mathcal{L} = -\log P(r^* \mid \mathcal{X}, q) + \mu \|\boldsymbol{\theta}_{\text{IGCA}}\|_2^2$
  Backprop through IGCA only (LLM weights frozen)
  Update: $\alpha_l \leftarrow \alpha_l - \eta \nabla_{\alpha_l} \mathcal{L}$; $\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} - \eta \nabla_{\mathbf{W}^{(l)}} \mathcal{L}$
end for

Algorithm 3 IM-Head Reinforcement Learning

Require: Expert 1 hidden states $\mathbf{h}$, reward R, cost c, discount γ_{RL}
Ensure: IM-Head policy π_ϕ
for each episode (query q, true answer a^*) **do**
  trajectory $\tau \leftarrow []$
  while action $\neq$ CONCLUDE **do**
    $a_t \leftarrow \pi_\phi(\mathbf{h}_t)$ (sample action)
    Execute a_t; receive observation o_t
    $r_t \leftarrow -c \cdot \mathbf{1}[a_t \in \{\text{RETRIEVE}, \text{INSPECT}\}]$
    Append (a_t, r_t) to τ
  end while
  Final reward: $r_T \leftarrow R \cdot \mathbf{1}[\text{response} = a^*]$
  Compute returns $G_t = \sum_{k \geq t} \gamma_{\text{RL}}^{k-t} r_k$
  Update ϕ via REINFORCE: $\phi \leftarrow \phi + \alpha \sum_t G_t \nabla_\phi \log \pi_\phi(a_t \mid \mathbf{h}_t)$
end for