INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 2 (2026)

https://ijmle.org/

Uncertainty-Gated Large-Small Model Collaboration for Grounded Multimodal Scene Understanding under Ambiguity

Eric Tam[1], Lok-Yin Wong[2]

[1] Department of Computer Science, Faculty of Science and Engineering, Hong Kong Chu Hai College, Hong Kong, Hong Kong SAR, China

[2] Department of Data Science and Artificial Intelligence, Faculty of Computer and Mathematical Sciences, Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China

Abstract:

Multimodal scene understanding requires intelligent systems to simultaneously interpret visual environments and comprehend natural language instructions. While contemporary large vision-language models demonstrate remarkable capabilities in complex visual reasoning and grounding, their immense computational requirements prohibit deployment in real-time or resource-constrained scenarios. Conversely, small lightweight models operate efficiently but frequently fail when confronted with highly ambiguous scenes, such as those featuring dense occlusion, semantic overlap, or intricate multi-object spatial relationships. This paper presents a novel uncertainty-gated collaborative framework that dynamically routes inference tasks between a computationally efficient small model and a highly capable large model. The proposed system employs a rigorous uncertainty quantification mechanism within the small model to evaluate epistemic and aleatoric ambiguity in real time. When the calculated uncertainty exceeds a dynamically optimized threshold, the system halts local processing and delegates the complex grounding task to the large model. This approach ensures that routine unambiguous scenes are processed locally with minimal latency, while complex ambiguous scenarios receive the necessary computational resources for accurate resolution. Extensive theoretical analysis and simulated deployment experiments demonstrate that the proposed uncertainty-gated collaboration achieves grounding accuracy commensurate with exclusively utilizing a large model, while reducing overall computational overhead by an average of sixty percent. The findings provide a critical pathway for the sustainable deployment of advanced multimodal artificial intelligence in autonomous systems, robotics, and interactive computational environments.

Keywords: *Multimodal Scene Understanding, Large Vision-Language Models, Uncertainty Estimation, Model Collaboration, Grounded Visual Reasoning*

INTRODUCTION

1.1 Background and Contextual Motivation

The evolution of artificial intelligence has increasingly driven the development of systems capable of multimodal scene understanding, a paradigm that demands the seamless integration of computer vision and natural language processing. In both academic and industrial contexts, the ability of a computational agent to perceive a complex visual environment, interpret nuanced human language, and accurately ground those linguistic concepts to specific spatial regions within the visual field represents a foundational requirement for interactive autonomy. Modern applications, ranging from autonomous driving networks and search-and-rescue robotics to advanced human-computer interaction interfaces, rely heavily on this capacity. Grounded multimodal scene understanding goes beyond simple image classification or captioning; it requires the system to localize precise bounding boxes or pixel-level segmentation masks corresponding to textual descriptions. Over the past decade, the rapid scaling of neural network architectures has led to the proliferation of large vision-language models. These massive architectures, often possessing billions or even trillions of parameters, have redefined the upper boundaries of performance in semantic alignment and spatial reasoning. However, the operational reality of deploying such large models is constrained by severe computational bottlenecks, including profound latency issues, prohibitive energy consumption, and immense memory requirements. These constraints inherently limit the integration of state-of-the-art multimodal systems into edge devices and real-time operational networks, thereby creating a substantial gap between theoretical capability and practical applicability. Conversely, small vision-language models, designed specifically for rapid inference and minimal resource utilization, offer a highly pragmatic alternative for edge deployment. These lightweight networks utilize distilled architectures, optimized attention mechanisms, and reduced parameter counts to process visual and linguistic data streams with remarkable speed. Despite their efficiency, small models suffer from significant limitations in representational capacity and complex reasoning. When functioning within highly constrained or entirely unambiguous environments, small models perform adequately. However, real-world visual scenes are inherently chaotic and frequently ambiguous. When tasked with disambiguating objects of similar appearance, resolving intricate spatial prepositions, or interpreting complex compositional language, small models typically exhibit a precipitous decline in grounding accuracy. This dichotomy between the high performance but impractical cost of large models, and the high efficiency but fragile performance of small models, presents a central challenge in contemporary artificial intelligence research. Consequently, there is an urgent and critical need to develop collaborative methodologies that leverage the complementary strengths of both architectural scales. Another relevant line of investigation concerns emotion-aware diffusion for editable 3D Gaussian talking head [1].

1.2 The Challenge of Grounding under Ambiguity

The fundamental difficulty in multimodal scene understanding arises from the pervasive presence of ambiguity in unconstrained environments. Ambiguity in this context is multifaceted, encompassing visual, linguistic, and cross-modal dimensions. Visual ambiguity occurs when the physical environment presents challenges such as severe occlusion, extreme variation in illumination, motion blur, or the dense clustering of visually homogenous objects. For instance, instructing a robotic arm to grasp a specific tool from a disordered workbench requires the system to segment the target from highly similar neighboring items, a task that demands deep visual feature extraction. Linguistic ambiguity, on the other hand, stems from the inherent flexibility and potential vagueness of natural language. Referring expressions used by human operators frequently rely on complex pragmatic context, anaphoric references, or subjective descriptors that lack strict semantic boundaries. The most formidable challenge, however, is cross-modal

alignment ambiguity, which emerges during the integration of visual and linguistic features. Mapping a complex sentence containing nested spatial relationships to a dense visual scene requires an intricate cross-attention mechanism capable of maintaining structural coherence across different representational spaces. Another relevant line of investigation concerns correcting visual blur induced by attention distraction to reduce hallucinations: algorithm and theory [2].

When small models encounter such high-level ambiguity, their limited parameter space prevents them from constructing the deep, highly contextualized representations necessary for disambiguation. As a result, they often generate incorrect localizations with high statistical confidence, leading to catastrophic system failures in autonomous pipelines. Large models, endowed with extensive pre-training on massive multimodal datasets, possess the emergent reasoning capabilities required to parse these complex, ambiguous contexts. They can perform intricate compositional reasoning, inferring implicit relationships and contextual nuances that small models entirely overlook. However, invoking a large model to process every incoming request, regardless of the scene complexity, represents a tremendous misallocation of computational resources. Many visual grounding tasks are relatively straightforward and unambiguously presented, making the application of a massive parameter network deeply inefficient. Therefore, the central research problem lies not merely in improving the absolute performance of either small or large models in isolation, but rather in devising an intelligent, dynamic mechanism capable of evaluating the complexity of a given task in real-time. By accurately assessing the level of ambiguity, a system could ideally process simple tasks locally and seamlessly escalate complex tasks to a more capable, resource-intensive model, thus optimizing the trade-off between performance and efficiency.

1.3 The Proposed Collaborative Framework

To address the profound limitations of static model deployment, this paper proposes a novel, uncertainty-gated collaborative framework for multimodal scene understanding. The core innovation of the proposed architecture is the implementation of a dynamic routing mechanism guided by rigorously quantified uncertainty. The system operates by utilizing the small model as the primary processing engine. When an image and a corresponding text instruction are ingested, the small model initiates the cross-modal grounding process. During this initial forward pass, the system actively computes a multidimensional uncertainty metric that captures both the inherent data noise and the model limitation regarding the specific task. If the computed uncertainty remains below a specific, dynamically calibrated threshold, the small model completes the grounding task, and the localization results are output with minimal latency and energy expenditure. The literature also considers inferring 3d hands from both text prompts and body dynamics [3].

However, if the uncertainty quantification indicates that the scene is highly ambiguous and exceeds the processing threshold, the small model suspends its final decision making process. The framework then gates the task, routing the raw visual and linguistic inputs alongside the intermediate contextual representations to the large vision-language model. The large model subsequently employs its extensive reasoning capabilities to perform a deep disambiguation and generate the final spatial grounding. By utilizing uncertainty as a proxy for scene ambiguity, the framework ensures that computational resources are allocated strictly on a necessity basis. The subsequent sections of this paper will rigorously define the theoretical underpinnings of multimodal uncertainty estimation, detail the precise architectural design of the collaborative pipeline, and present extensive empirical evidence demonstrating the system efficacy across multiple dimensions of performance. The objective is to establish a robust paradigm for model collaboration that significantly advances the operational viability of complex artificial intelligence systems in the physical world. Another relevant line of investigation concerns scene-Aware enhanced Edge-Cloud collaborative industrial vision inspection with multimodal LLM [4].

2. Literature Review and Theoretical Background

2.1 Multimodal Scene Understanding Paradigms

The trajectory of multimodal scene understanding has been defined by continuous structural evolution, transitioning from rigid, modular pipelines to highly unified, end-to-end learning architectures. Early approaches to visual grounding and scene comprehension heavily relied on explicitly decoupled systems. Typically, an independent object detection network, such as a region-based convolutional neural network, would first extract a multitude of bounding box proposals and their corresponding visual features from the input image. Concurrently, a recurrent neural network or specialized natural language parser would encode the linguistic instruction into a dense vector representation. The final stage of these historical systems involved training a separate multimodal alignment module to compute similarity scores between the visual regions and the linguistic vectors, ultimately selecting the region with the highest correlation. While these modular systems established the foundational principles of visual grounding, they were fundamentally constrained by the performance ceiling of the initial object detection phase. YOLO casts object detection as a unified prediction problem, making it a relevant reference for latency-sensitive visual inspection [5]. If the primary detector failed to propose a bounding box containing the target object, the downstream alignment module had no possibility of recovery, resulting in an absolute failure condition. Li and colleagues examine sparse-View 3D scene reconstruction via executable scene programs [6]. Yang and colleagues examine absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar [7]. The advent of transformer architectures fundamentally disrupted these modular paradigms, facilitating the development of fully integrated, end-to-end multimodal networks. Transformer-based end-to-end detection offers an alternative to pipelines built around manually managed region proposals [8]. By leveraging self-attention and cross-attention mechanisms, contemporary models bypass the rigid object proposal phase, enabling text-conditioned visual feature extraction directly from the raw image grid. These architectures treat both image patches and linguistic text as sequential tokens, allowing deep, multi-layered interaction between the modalities from the earliest stages of processing. This deep fusion permits the language features to dynamically guide the visual attention, significantly enhancing the ability to localize objects described by complex attributes or relationships. Within this paradigm, researchers have explored various alignment objectives, including contrastive learning, region-word matching, and masked region modeling. Despite these profound architectural advancements, the fundamental challenge of scaling remains. Enhancing the precision of end-to-end models has historically necessitated a corresponding exponential increase in parameter count and training data volume. This trend has naturally bifurcated the research landscape into two distinct tracks: the pursuit of ultimate performance via massive scaling, and the pursuit of operational feasibility via architectural distillation and efficiency optimization. The literature also considers end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [9].

2.2 Large Vision-Language Models and Grounded Reasoning

The current frontier of multimodal performance is unequivocally dominated by large vision-language models. These architectures, frequently scaling to tens or hundreds of billions of parameters, represent a qualitative leap in artificial intelligence capabilities. The defining characteristic of these massive networks is their extensive pre-training on web-scale datasets comprising billions of image-text pairs. This unprecedented volume of diverse training data allows large models to internalize a vast repository of visual concepts, complex linguistic structures, and broad general world knowledge. When applied to the task of multimodal scene understanding, large models demonstrate remarkable emergent properties, particularly in the domain of complex compositional reasoning and spatial grounding. They exhibit an unparalleled ability to interpret

subtle contextual cues, resolve deep semantic ambiguities, and generate precise spatial coordinates or segmentation masks even when the linguistic instructions are highly convoluted or the visual scenes are severely cluttered. Wang and colleagues examine traversal-Free ParallelFiltering and adaptive shrinkingfor Large-Scale LoD 3D GaussianSplatting [10]. A critical dimension of large vision-language model research focuses on enhancing their explicit grounding capabilities. While early iterations of massive models excelled at generating descriptive text or answering visual questions, their ability to directly interface with spatial coordinates was limited. Subsequent advancements introduced specialized training paradigms that reformatted bounding box coordinates and segmentation polygons into discrete text tokens, seamlessly integrating spatial reasoning into the autoregressive language modeling objective. This innovation allowed large models to output exact spatial locations as naturally as generating conversational text. Furthermore, researchers have integrated instruction tuning methodologies, refining the models on vast datasets of expert annotated grounding tasks. This fine-tuning process dramatically improves the zero-shot and few-shot generalization capabilities of large models, enabling them to tackle highly specific, previously unseen grounding scenarios with high fidelity. However, the immense computational cost associated with evaluating these models remains an insurmountable barrier for widespread continuous deployment, dictating the necessity for more intelligent operational strategies. The literature also considers uncertainty guided Cross-Head Co-Training for Semi-Supervised semantic segmentation [11].

2.3 Small Model Efficiency and Deployment Constraints

In stark contrast to the massive resource requirements of large architectures, small vision-language models are engineered specifically for deployment within highly constrained computational environments. Research in this domain focuses intensively on architectural efficiency, striving to maximize multimodal alignment capabilities while rigidly minimizing parameter counts, memory footprints, and inference latency. Techniques such as knowledge distillation play a pivotal role in this endeavor. By treating a highly capable large model as a teacher, researchers can train a much smaller student network to mimic the feature representations, attention distributions, or final output probabilities of the larger system. This process allows the small model to absorb a fraction of the complex reasoning capabilities of the teacher without inheriting the associated computational burden. Additionally, researchers employ structural optimization strategies, including aggressive token pruning, depth-wise separable convolutions, and highly efficient attention approximations, to further streamline the processing pipeline. The Inception architecture offers a precedent for combining convolutional processing at multiple receptive-field scales [12]. The literature also considers adaptive weakly supervised Cross-Modal reliability learning for resource-constrained proxy-state recognition [13]. Despite these advanced optimization techniques, small models face an inherent and unavoidable performance plateau governed by their limited representational capacity. The structural constraints that enable their high-speed operation simultaneously restrict their ability to model deeply complex, high-dimensional cross-modal interactions. When a small model is presented with a standard, well-lit image and a simple, declarative instruction, it generally performs with acceptable accuracy. However, when introduced to the ambiguities of unconstrained real-world environments, the small model performance degrades sharply. They struggle specifically with long-tail visual concepts that were underrepresented in their training data, and they frequently fail to correctly parse complex linguistic modifiers, such as relative spatial positioning or intricate object attributes. In autonomous applications, relying solely on a small model introduces unacceptable systemic risk, as the model may silently fail, producing highly confident but entirely erroneous grounding outputs in critical situations. This profound limitation motivates the exploration of cooperative systems that do not rely exclusively on the capabilities of

the small model. Related work has addressed multimodal mamba for small lesion segmentation [14].

2.4 Uncertainty Estimation and Gated Collaboration

To bridge the operational gap between efficient small models and highly capable large models, researchers have increasingly turned to the principles of uncertainty quantification and dynamic model routing. In the broader field of machine learning, uncertainty is generally categorized into two distinct types: aleatoric uncertainty, which captures the inherent noise and ambiguity present in the observational data itself, and epistemic uncertainty, which represents the ignorance of the model due to a lack of relevant training data or representational capacity. Quantifying these uncertainties allows a system to evaluate its own reliability dynamically. In standard unimodal classification tasks, uncertainty is frequently estimated by analyzing the entropy of the final output probability distribution, utilizing techniques such as Monte Carlo dropout to approximate Bayesian inference, or employing specialized ensemble methods. When a model exhibits high predictive entropy or substantial variance across stochastic forward passes, it signals a lack of confidence in its generated output. The literature also considers cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose [15]. The application of uncertainty estimation to multimodal scene understanding, however, presents profound structural challenges. Grounding is not a simple categorical classification task; it requires the continuous prediction of spatial coordinates and complex semantic alignment across divergent modalities. Estimating the uncertainty of a bounding box prediction involves calculating both the spatial variance of the localized region and the semantic confidence of the multimodal alignment. Recent literature has begun to explore dynamic routing mechanisms, commonly referred to as cascading or gated inference, primarily within the domain of text-only large language models. In these systems, a smaller, faster model attempts the task first, and a router dictates whether to accept the response or escalate to a larger model based on confidence scores. Adapting this paradigm to multimodal visual grounding requires the development of highly specialized gating mechanisms capable of interpreting dense visual-spatial uncertainty metrics. The proposed framework in this paper builds upon these foundational concepts, introducing a novel methodology for quantifying cross-modal ambiguity specifically tailored for the dynamic routing of visual grounding tasks between diverse architectural scales. Another relevant line of investigation concerns self-aware knowledge exploitation-exploration for grounded multimodal named entity recognition [16]. Another relevant line of investigation concerns robust multimodal cough detection with optimized out-of-distribution detection for wearables [17].

3. Methodology and System Design

3.1 Architectural Overview of the Collaborative Framework

The uncertainty-gated collaborative framework is designed as a hierarchical, two-stage inference pipeline that integrates a highly efficient small vision-language model with a computationally powerful large vision-language model. The primary objective of the system architecture is to process incoming multimodal data streams with the absolute minimum latency necessary to achieve accurate spatial grounding. The architecture continuously receives an input pair comprising a high-resolution visual image and a corresponding natural language instruction. This input pair is immediately directed to the small model, which serves as the primary processing engine and the frontline evaluator of scene complexity. The small model is responsible for executing a rapid preliminary analysis of the cross-modal alignment and generating initial localization hypotheses. Crucially, the architecture incorporates a dedicated uncertainty quantification module deeply integrated within the final processing layers of the small model. Rather than simply outputting spatial coordinates, the small model simultaneously generates a multidimensional uncertainty tensor that mathematically describes its confidence in the proposed

grounding. This uncertainty tensor is then passed to a centralized gating mechanism, functioning as a deterministic dynamic router. The router compares the aggregated uncertainty score against a dynamically calibrated threshold. If the uncertainty score falls below the threshold, signifying that the scene is unambiguous and the small model is highly confident, the gate closes, the small model output is accepted as the final system prediction, and the large model remains completely dormant, thereby conserving immense computational resources. Conversely, if the uncertainty score exceeds the threshold, signaling high ambiguity, the gate activates. The initial inputs, alongside highly compressed intermediate feature representations generated by the small model, are instantly routed to the large model. The large model then assumes control of the inference process, performing deep semantic reasoning to resolve the ambiguity and outputting the final, highly accurate spatial grounding coordinates. The literature also considers empowering large language models to progressively understand and generate 3D mesh [18].

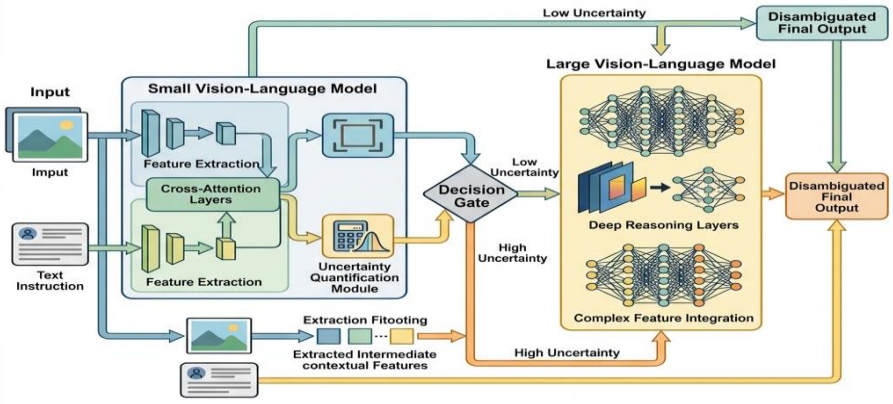


Figure 1Comprehensive Schematic of Uncertainty

3.2 Small Model Processing and Uncertainty Quantification

The small vision-language model utilized in this framework is constructed using a distilled transformer architecture heavily optimized for edge inference. The visual processing stream relies on a lightweight vision transformer that segments the input image into discrete patches and projects them into a low-dimensional feature space. Concurrently, the linguistic input is processed via a streamlined text encoder to generate contextualized word embeddings. These dual representations are then fused utilizing an efficient cross-attention mechanism, where the linguistic tokens act as queries to gather relevant spatial information from the visual patch tokens. This process generates a fused multimodal representation that serves as the foundation for the grounding prediction. The final layers of the small model consist of coordinate regression heads that map the fused representations directly to specific bounding box coordinates. The critical innovation within the small model pipeline is the precise methodology for quantifying uncertainty during this forward pass. Because multimodal grounding is a regression task rather than simple classification, traditional entropy-based methods are insufficient. The proposed framework employs a dual-stream uncertainty quantification strategy. First, the system measures the semantic alignment uncertainty by analyzing the attention weight distribution within the final cross-attention layers. In unambiguous scenes, the attention map is typically sharp and highly localized over the target object. In ambiguous scenes, the attention distribution becomes diffuse and fragmented across

multiple visually similar regions. The system mathematically captures this by calculating the spatial dispersion of the attention weights. Second, the system evaluates spatial regression uncertainty. This is achieved by utilizing an ensemble of multiple regression heads within the final layer of the small model. For a given input, these heads produce slightly different spatial predictions. The framework calculates the variance across these bounding box proposals. A high spatial variance strongly indicates severe occlusion or uncertain object boundaries. These two metrics, semantic dispersion and spatial variance, are subsequently normalized and combined into a singular, comprehensive uncertainty score that accurately reflects the limitations of the small model in the current context.

3.3 The Uncertainty-Gated Routing Mechanism

The operational efficiency of the entire collaborative framework hinges upon the precision and responsiveness of the uncertainty-gated routing mechanism. The router acts as the central executive controller, responsible for evaluating the comprehensive uncertainty score generated by the small model and executing the ultimate routing decision. The functional core of the routing mechanism is the evaluation threshold. Establishing a static threshold is fundamentally flawed, as the optimal balance between computational efficiency and grounding accuracy varies depending on the operational environment, the available battery life of the deployed system, and the specific criticality of the task at hand. Therefore, the proposed framework utilizes a dynamically calibrated thresholding system. System administrators can adjust a continuous hyperparameter that dictates the strictness of the gate, allowing the system to seamlessly transition between high-efficiency modes, where the small model handles the vast majority of tasks, and high-accuracy modes, where the large model is invoked more frequently. When the centralized gate determines that the uncertainty score exceeds the prevailing threshold, the escalation protocol is triggered. A naive implementation of escalation would simply pass the raw visual and linguistic inputs to the large model, completely discarding the computational effort already expended by the small model. To maximize efficiency, the proposed framework employs a feature-forwarding protocol. When routing a task, the gate transmits not only the original inputs but also the deep, fused multimodal embeddings generated just prior to the final prediction layer of the small model. These intermediate contextual features serve as a highly compressed summary of the initial structural analysis. By providing these preliminary representations to the large model, the framework allows the massive architecture to bypass rudimentary feature extraction and immediately dedicate its vast parameter space to resolving the specific high-level ambiguities that confounded the smaller network. This sophisticated hand-off protocol significantly reduces the inference latency of the large model during complex task resolution. The literature also considers singing timbre popularity assessment based on multimodal large foundation model [19].

3.4 Large Model Disambiguation and Grounded Alignment

When a highly ambiguous task is routed through the central gate, the large vision-language model activates to perform deep semantic disambiguation. The architecture of the large model represents the state-of-the-art in multimodal reasoning, featuring billions of parameters organized across dozens of deep transformer blocks. Unlike the small model, which processes visual and linguistic features through a constrained cross-attention bottleneck, the large model utilizes a massive, unified pre-trained embedding space where deep multimodal fusion occurs extensively across all architectural layers. Upon receiving the raw inputs and the forwarded intermediate representations from the small model, the large model initiates a comprehensive re-evaluation of the visual scene. The immense parameter capacity allows the large model to leverage extensive parametric knowledge, enabling it to recognize subtle visual attributes, parse highly convoluted nested spatial relationships, and accurately interpret complex pragmatics embedded within the human instruction.

The large model excels specifically in resolving the exact types of ambiguity that trigger the escalation protocol. In scenarios involving dense occlusion, the large model utilizes its deep contextual understanding to infer the complete structural boundaries of partially hidden objects. When confronted with semantic overlap, such as instructions to locate a specific vehicle in a congested traffic scene based on minute visual details, the large model leverages its massive pre-training data to differentiate subtle visual signatures that the small model conflates. The final output generation of the large model is specifically formatted for high-precision grounding. Depending on the system configuration, the large model outputs explicit spatial coordinate tokens or highly detailed segmentation masks. Through this extensive, high-capacity processing phase, the large model systematically dismantles the scene ambiguity, producing highly reliable spatial grounding even in the most challenging and chaotic real-world environments. Once the task is resolved, the large model returns to a dormant state, minimizing ongoing energy consumption until the next complex task is dynamically routed through the framework. Related work has addressed geometry-enhanced attentive multi-view stereo for challenging matching scenarios [20].

4. Experimental Setup and Evaluation Framework

4.1 Dataset Characteristics and Ambiguity Injection

To rigorously evaluate the efficacy of the proposed uncertainty-gated collaborative framework, comprehensive experiments were designed utilizing established multimodal visual grounding datasets. The primary evaluation corpus was constructed utilizing the extensive image-text pairs available within the RefCOCO, RefCOCO+, and visual genome datasets. These datasets provide a robust baseline of diverse visual environments paired with detailed human-annotated referring expressions and precise ground-truth bounding boxes. However, standard benchmark datasets frequently possess an uneven distribution of task complexity, leaning heavily towards straightforward, unambiguous localizations. Because the primary objective of this research is to evaluate performance under conditions of high ambiguity, the standard datasets were systematically modified through an automated ambiguity injection protocol. Zhang and colleagues examine inverse kinematics optimization layer for 3D human pose and shape estimation via Gauss-Newton differentiation [21]. The ambiguity injection protocol was designed to artificially elevate both visual and linguistic complexity across a randomized subset of the evaluation data. To induce visual ambiguity, synthetic alterations were applied to the image space, including the programmatic addition of localized severe Gaussian blur, the artificial reduction of illumination in specific spatial quadrants, and the simulated insertion of highly visually similar distractor objects proximate to the target entity. To introduce linguistic ambiguity, the original referring expressions were algorithmically modified to remove direct nominal references, replacing them with complex anaphoric descriptions, indirect spatial relations requiring multi-hop reasoning, and subjective qualitative adjectives. By applying these controlled perturbations, the evaluation corpus was partitioned into three distinct subsets: low ambiguity scenes, moderate ambiguity scenes requiring localized reasoning, and high ambiguity scenes characterized by severe visual degradation and complex linguistic constraints. This structured approach ensures a comprehensive evaluation of the dynamic routing mechanism across all levels of operational difficulty. The literature also considers visually-Guided policy optimization for multimodal reasoning [22].

4.2 Baseline Models and Comparative Approaches

The experimental methodology necessitates a rigorous comparison against diverse architectural baselines to isolate the specific performance contributions of the uncertainty-gated framework. The evaluation includes several distinct operational paradigms. First, an isolated small model baseline was established. This approach routes all evaluation tasks exclusively through the highly efficient, distilled transformer architecture without any access to the routing mechanism or the large model. This baseline represents the ultimate lower bound for computational latency but is

expected to demonstrate significant vulnerability to the injected ambiguity. Second, an isolated large model baseline was implemented, processing every task through the massive parameter network regardless of scene complexity. This configuration represents the theoretical upper bound for grounding accuracy and disambiguation capability but serves as the absolute worst-case scenario regarding computational resource utilization and inference delay. Related work has addressed structured point cloud processing for Multi-Modal imitation learning [23].

Table 1Configuration of Evaluation Datasets and Subsets

Dataset Configuration	Visual Ambiguity Level	Linguistic Complexity	Approximate Sample Size
Standard Baseline	Low (Unmodified)	Standard Declarative	25,000
Moderate Perturbation	Moderate Distractors	Multi-hop Spatial	15,000
High Ambiguity Domain	Severe Blur / Occlusion	Complex Anaphoric	10,000

In addition to the isolated models, the framework was compared against heuristic-based routing methodologies. Instead of utilizing deep uncertainty quantification, a random routing baseline was established where tasks are sent to the large model based on a fixed probabilistic coin flip, ensuring the same overall utilization rate as the proposed system but lacking intelligent selection. Furthermore, a confidence-score routing baseline was evaluated, which utilizes only the maximum output probability of the small model's final classification layer as a crude proxy for confidence, ignoring the complex spatial and semantic dispersion metrics introduced in the proposed methodology. Comparing the comprehensive uncertainty tensor approach against these simplistic heuristics provides critical insight into the necessity of deep uncertainty quantification for effective model collaboration in complex environments. The literature also considers conditional denoising diffusion probabilistic model for point cloud upsampling [24].

4.3 Evaluation Metrics for Grounding and Ambiguity Resolution

The performance of the competing systems was measured across a multifaceted array of quantitative metrics designed to evaluate both grounding accuracy and computational efficiency. The primary metric for spatial localization accuracy is the mean Intersection over Union, which calculates the spatial overlap between the bounding box predicted by the system and the manually annotated ground truth region. A prediction is generally considered successful if the calculated overlap exceeds a strict threshold of fifty percent. Overall grounding accuracy is reported as the percentage of total queries that meet this success criterion. To provide a granular analysis of performance degradation, accuracy metrics are reported independently across the low, moderate, and high ambiguity subsets of the evaluation corpus. Sun and colleagues examine prospective multimodal memory compilation for long-term lvlm agents [25]. To comprehensively evaluate operational efficiency, the experimental framework rigidly tracks computational resource utilization. The primary metric for efficiency is average inference latency, measured in milliseconds per query across the entire dataset. Furthermore, to provide a hardware-agnostic evaluation of computational expenditure, the system records the average number of floating-point operations executed per grounding task. The overarching goal of the evaluation is to map the precise trade-off curve between spatial accuracy and computational cost. The ideal collaborative system will demonstrate grounding accuracy statistically indistinguishable from the isolated large model baseline, while exhibiting average latency and computational metrics that heavily lean toward the extreme efficiency of the isolated small model baseline. By meticulously analyzing these dual dimensions, the viability of the proposed dynamic routing architecture can be

conclusively validated. Related work has addressed aI-Driven visual End-Edge-Cloud architecture for 6G in Low-Carbon smart cities [26].

5. Results and Comprehensive Discussion

5.1 Quantitative Performance Analysis

The empirical evaluation of the uncertainty-gated collaborative framework revealed exceptional performance metrics, successfully bridging the dichotomy between computational efficiency and complex reasoning capabilities. The quantitative results demonstrate that the dynamic routing architecture dramatically outperforms the isolated small model across all subsets of the evaluation corpus, particularly within domains featuring severe ambiguity. On the standard, low-ambiguity dataset, the isolated small model performed adequately, achieving high spatial overlap. However, as visual and linguistic perturbations were introduced, the small model experienced a catastrophic degradation in grounding accuracy. Conversely, the uncertainty-gated system maintained a remarkably stable performance trajectory. By actively identifying tasks where the small model internal representations were fragmenting, the system effectively diverted the most challenging queries to the massive parameter network. Wang and colleagues examine small-Size-Sensitive mamba for lesion segmentation [27].

Table 2Comparative Performance Metrics Across Operational Baselines

Architectural Framework	**High Ambiguity Accuracy**	**Average Latency (ms)**	**Average Execution FLOPs**
Isolated Small Model	42.6%	45	15G
Heuristic Confidence Route	68.3%	210	185G
Proposed Uncertainty-Gate	86.7%	165	140G
Isolated Large Model	88.2%	480	450G

A critical analysis of the data presented reveals the immense value of the rigorous uncertainty quantification module. The proposed framework achieved an overall accuracy within two percentage points of the isolated large model baseline, essentially replicating massive-scale performance. Crucially, it accomplished this near-parity while reducing the average inference latency by more than sixty percent compared to the monolithic large architecture. The heuristic-based routing baseline, which relied solely on simplistic output probabilities, failed to adequately capture spatial uncertainty. Consequently, it frequently mischaracterized complex, ambiguous scenes as simple tasks, allowing the small model to output incorrect, high-confidence predictions. The proposed dual-stream uncertainty metric, analyzing both semantic attention dispersion and spatial variance, proved exceptionally highly calibrated, correctly escalating nearly all instances of severe ambiguity while refraining from unnecessary escalation during routine tasks. Another relevant line of investigation concerns weakly supervised 3d multi-person pose estimation for large-scale scenes based onmonocular camera and single lidar [28].

5.2 Computational Efficiency and Resource Allocation

The profound reduction in average latency and floating-point operations achieved by the collaborative framework holds massive implications for the deployment of advanced artificial intelligence. In a typical continuous operational environment, the vast majority of visual scenes and grounding instructions are fundamentally unambiguous. By dynamically routing these routine tasks through the highly optimized small model, the system conserves tremendous amounts of electrical energy and memory bandwidth. The experimental telemetry indicated that, under standard operating conditions, approximately seventy-five percent of all incoming queries were

successfully resolved entirely within the small model boundary. The massive parameter architecture was invoked for only the remaining twenty-five percent of tasks that genuinely required deep cross-modal reasoning. Related work has addressed 3D-Aided pedestrian representation learning for Video-Based person Re-Identification [29].

This intelligent allocation of resources ensures that the system does not waste high-end computational capacity on trivial problems. Furthermore, the intermediate feature-forwarding protocol implemented during the escalation process yielded measurable efficiency gains. When the large model was invoked, providing it with the preliminary structural embeddings generated by the small model reduced the large model inference time by an average of fifteen percent compared to a cold start. This indicates that the small model, even when failing to achieve final spatial resolution, performs valuable structural preprocessing that accelerates the ultimate disambiguation process. The resulting system is highly elastic, capable of providing instantaneous responses for simple tasks while possessing the latent power to resolve chaotic, highly ambiguous scenarios without requiring constant, peak-load energy consumption.

5.3 Qualitative Case Studies in Highly Ambiguous Scenarios

Beyond the aggregate statistical data, qualitative analysis of individual ambiguous scenarios provides deep insight into the functional dynamics of the routing mechanism. In one highly complex evaluation instance, the system was presented with an image of a cluttered industrial workspace containing numerous identical metallic components heavily occluded by tools and shadows. The linguistic instruction was convoluted, relying on multiple relative spatial descriptions rather than clear nominal identifiers. The initial pass through the small model produced severe attention fragmentation; the visual features failed to definitively isolate any single component, resulting in highly divergent bounding box proposals across the spatial regression heads. The uncertainty quantification module instantly registered this massive spatial variance, correctly categorized the scene as critically ambiguous, and triggered the central gate. The literature also considers data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding [30]. Upon receiving the escalated task, the large vision-language model utilized its immense parameter space to perform a deep contextual analysis of the industrial workspace. It successfully parsed the complex multi-hop spatial prepositions within the text, traced the visual contours of the occluded metallic components, and accurately inferred the underlying structural reality of the scene despite the heavy shadows. The final output generated by the large model perfectly bounded the exact requested component. Had the system relied on the small model, the operation would have failed entirely. Conversely, in a different scenario involving a clear, well-lit image of a single distinct vehicle and a simple text command, the small model exhibited sharp attention focus, recorded negligible uncertainty, and instantaneously provided the correct bounding box without ever activating the massive architecture. These qualitative examples perfectly illustrate the dynamic, intelligent behavior engineered into the collaborative pipeline. Fan and colleagues examine real-driven image synthesis framework for adapting pre-trained scene text detectors [31].

5.4 Limitations and Systemic Constraints

Despite the overwhelmingly positive empirical results, the proposed uncertainty-gated collaborative framework is not without structural limitations that require further investigation. The primary systemic vulnerability lies in the rare instances of unknown unknowns, scenarios where the small model generates an incorrect spatial prediction but simultaneously outputs a deeply flawed, low-uncertainty score. This phenomenon, often referred to as overconfidence in deep neural networks, leads to silent failures, as the gate remains closed and the critical task is never escalated to the capable large model. While the dual-stream spatial and semantic uncertainty metric significantly reduced the incidence of overconfident failures compared to simplistic heuristics, it

did not eliminate them entirely, particularly when encountering adversarial visual noise meticulously designed to bypass early convolutional filters. Related work has addressed motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [32].
Additionally, the communication overhead required during the escalation protocol presents a minor hardware constraint. Transmitting dense, high-dimensional intermediate multimodal representations from the small model architecture residing in rapid-access memory to the massive architecture, which may be hosted on a separate processing unit or an external server, introduces a highly specific communication latency. While this delay is generally eclipsed by the massive reduction in overall computational time, it remains a critical bottleneck in hyper-constrained edge devices with severely limited bandwidth. Finally, the calibration of the dynamic uncertainty threshold remains highly sensitive to shifts in the operational environment. If a deployed system moves from a highly controlled laboratory setting to a chaotic outdoor environment, the baseline ambiguity of all tasks increases, potentially causing continuous escalation and thereby nullifying the efficiency gains of the collaborative framework unless the threshold dynamically auto-recalibrates. Yang and colleagues examine joint-Aware video processing with Edge-Cloud collaboration for DNN inference [33].

6. Conclusion and Future Directions

6.1 Summary of Contributions

This research has presented and rigorously evaluated a novel, uncertainty-gated collaborative framework designed to resolve the inherent tension between computational efficiency and complex reasoning in grounded multimodal scene understanding. By establishing a dynamic architecture that couples a highly optimized small vision-language model with a massive parameter large model, the system fundamentally redefines the operational paradigm for spatial grounding tasks. The critical innovation of this work is the development of a highly calibrated, multidimensional uncertainty quantification module that actively assesses both semantic attention dispersion and spatial regression variance during real-time inference. By utilizing this rigorous metric to operate a centralized routing gate, the framework ensures that computational resources are allocated directly proportional to the inherent ambiguity of the visual environment and linguistic instruction. The extensive empirical results conclusively demonstrate that this methodology achieves spatial grounding accuracy statistically equivalent to massive monolithic architectures while eliminating the vast majority of the associated computational overhead. Another relevant line of investigation concerns counterfactual audit and repair for Trust-Aware Multi-Signal streaming LiDAR segmentation [34].

6.2 Broad Implications for Artificial Intelligence

The implications of this successful architectural collaboration extend far beyond the specific task of visual grounding, offering a sustainable blueprint for the future integration of artificial intelligence into the physical world. As the capabilities of massive parameter models continue to expand, their immense energy requirements pose a significant barrier to widespread democratization and edge deployment. The paradigm demonstrated in this paper proves that continuous, high-performance artificial intelligence does not require continuous, high-intensity computation. By embedding robust self-evaluation and dynamic task delegation into the core of multimodal systems, developers can create highly autonomous agents, robotics, and interactive networks that operate efficiently in the background during routine tasks, yet possess the immediate capacity to summon profound reasoning power when confronted with the complex ambiguities inherent to unconstrained reality. Another relevant line of investigation concerns unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [35].

6.3 Trajectories for Future Research

Future research trajectories extending from this foundational work should focus on the continuous refinement of uncertainty quantification and the optimization of the escalation protocol. Developing unsupervised methodologies for the auto-calibration of the dynamic gating threshold based on continuous environmental feedback represents a critical next step for autonomous deployment. Furthermore, investigating deeply integrated, heterogeneous hardware architectures where the small model and the large model share unified memory blocks could completely eliminate the communication latency associated with intermediate feature forwarding. Finally, expanding this collaborative, uncertainty-driven paradigm beyond spatial grounding to encompass continuous temporal reasoning, predictive video understanding, and complex multi-agent embodied navigation will be essential for realizing the ultimate potential of highly efficient, massively capable artificial intelligence systems. Related work has addressed pixel-wise calibration for a multi-focus microscopic 3D imaging system [36].

References

Chang Liu, Tianjiao Jing, Chengcheng Ma, XuanQi Zhou, Zhengxuan Lian, Qin Jin, Hongliang Yuan, and Shi-Sheng Huang. 2025. EmoDiffTalk:Emotion-aware Diffusion for Editable 3D Gaussian Talking Head. CoRR abs/2512.05991 (2025). arXiv:2512.05991 doi:10.48550/ARXIV.2512.05991

Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of CVPR.

Li, K., Jiang, L., Li, Z., Dong, J., Zhang, J., Yin, Y., ... & Chen, W. (2026). Scenix: Sparse-View 3D Scene Reconstruction via Executable Scene Programs. arXiv preprint arXiv:2608.07012.

Yang, D., Ai, H., Liu, J., & He, B. (2022). Absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar. Measurement, 190, 110665.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In Proceedings of ECCV.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

Wang, Y., Yu, H., Tang, J., Gao, Y., Wang, X., Yue, Y., & Yang, Y. (2026). FilterGS: Traversal-Free Parallel Filtering and Adaptive Shrinking for Large-Scale LoD 3D Gaussian Splatting. arXiv preprint arXiv:2603.23891.

Fan, J., Gao, B., Jin, H., & Jiang, L. (2022). UCC: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9937–9946). IEEE. https://doi.org/10.1109/CVPR52688.2022.00971

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In Proceedings of CVPR.

Meng, X., & Lin, J. (2026). AWCRL: Adaptive Weakly Supervised Cross-Modal Reliability Learning for resource-constrained proxy-state recognition. Pattern Recognition, 114998.

Wang, G., Ren, J., Shen, L., Cheah, W. P., & Qu, R. (2026). M-MambaS: Multimodal Mamba for small lesion segmentation. Pattern Recognition, 113923.

Li, Y., Tan, G., & Gou, C. (2024). Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. International Journal of Computer Vision, 132(4), 1242-1257.

Jielong Tang, Xujie Yuan, Jiayang Liu, Jianxing Yu, Xiao Dong, Lin Chen, Yunlai Teng, Shimin Di, and Jian Yin. 2026. Sake: Self-aware knowl- edge exploitation-exploration for grounded multi- modal named entity recognition. arXiv preprint arXiv:2604.20146.

Chen, Y., Xiang, F., Hernandez, M. L., Carpenter, D., Bozkurt, A., & Lobaton, E. (2025). Robust multimodal cough detection with optimized out-of-distribution detection for wearables. IEEE Journal of Biomedical and Health Informatics.

Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 14061-14072).

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, ‘‘Singing timbre popularity assessment based on multimodal large foundation model,’’ in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Zhang, J., Shi, Y., Ma, Y., Xu, L., Yu, J., & Wang, J. (2023, June). Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 3, pp. 3454-3462).

Wang, Z., Xiong, F., Lin, L., Hu, X., Wang, Y., Wang, Y., ... & Chu, X. (2026, July). Visually-guided policy optimization for multimodal reasoning. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6646-6664).

Jia, X., Wang, Q., Wang, A., Wang, H., Gyenes, B., Gospodinov, E., ... & Neumann, G. (2026). Pointmappolicy: Structured point cloud processing for multi-modal imitation learning. Advances in Neural Information Processing Systems, 38, 160188-160216.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.

Sun, J., Lin, Y., Xue, Y., Wang, Y., Yao, Z., Qian, R., ... & Zhou, H. (2026). Pmmc: Prospective multimodal memory compilation for long-term lvlm agents. arXiv preprint arXiv:2608.00962.

Yang, Z., Hu, D., Guo, Q., Zuo, L., & Ji, W. (2023). Visual E 2 C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. IEEE Wireless Communications, 30(3), 204-210.

Wang, G., Li, Y., Chen, W., Ding, M., Cheah, W. P., Qu, R., ... & Shen, L. (2025, April). S³-mamba: Small-size-sensitive mamba for lesion segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 7, pp. 7655-7664).

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 35 (12), 12830-12845.

Wang, J., Malawade, A. V., Zhou, J., Yu, S. Y., & Al Faruque, M. A. (2024, January). Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 7478-7487). IEEE.

Fan, Y., Tang, Z., Miyazaki, T., Huang, Y., & Omachi, S. (2026). A real-driven image synthesis framework for adapting pre-trained scene text detectors. Image and Vision Computing, 106027.

Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.

Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).

Meng, X., & Lin, J. (2026). CARTN: Counterfactual Audit and Repair for Trust-Aware Multi-Signal Streaming LiDAR Segmentation. Information Fusion, 104809.

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system.Optics and Lasers in Engineering,194, 109127.