Beyond Structural Unification: A Perspective and Roadmap on Unified Multimodal Models

Anonymous ACL submission

Abstract

Unified Multimodal Models (UMMs) emerge as a promising paradigm for integrating understanding and generation within a single architecture. We argue that current approaches achieve primarily structural unification, creating a *Unification Paradox*: unified interfaces, token spaces, and parameters do not necessarily yield functional emergence. We identify critical layers where unification must be rethought: (1) *Asymmetric Input Unification*, where models must quantify modality utility rather than treating all modalities as equally reliable; (2) *Factorized Representation Unification*, where semantic abstraction and fine-grained reconstruction should be jointly supported through disentangled capacity allocation; (3) *Stateful Reasoning Unification*, where multimodal reasoning should maintain evolving belief states beyond text-centric generation; (4) *Synergistic Optimization Unification*, where conflicting understanding and generation objectives require coordinated optimization; and (5) *Multi-Centric Systemic Evaluation*, where benchmarks must shift from isolated outcomes to systemic competence, capability preservation, and cross-task robustness. We contend that the next leap in unified multimodal intelligence requires moving from structural homogenization toward full-stack functional unification.

1 Introduction

Unified Multimodal Models (UMMs) (Deng et al., 2025; Ma et al., 2025; Xie et al., 2024; Chen, 2024) have recently emerged as a transformative family of foundation models, aiming to seamlessly integrate perception, reasoning, and generation within a single network (Figure 1). By collapsing disparate modal processors into a singular backbone, UMMs promise to transcend the boundaries of modular pipelines, unlocking holistic multimodal intelligence. Consequently, the community has entered an era of *Structural Unification Paradigm*—a

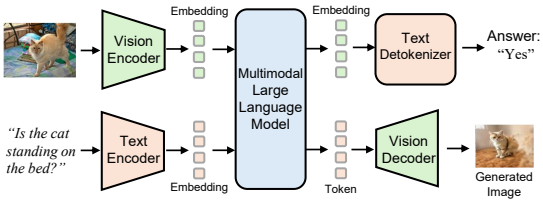

Figure 1: A general architecture of UMM.

pervasive belief that tokenizing pixels, unifying interfaces (e.g., next-token prediction across text and vision), and co-locating parameters automatically yields systemic synergy.

Existing UMMs pursue this paradigm through diffusion-based co-location (Swerdlow et al., 2025; Yang et al., 2025), autoregressive interface homogenization (Team, 2024; Wang et al., 2024b; Sun et al., 2024; Ge et al., 2024; Chen et al., 2025b), and fused architectures that stitch autoregressive decoding with diffusion or flow-based visual synthesis (Xie et al., 2024; Ma et al., 2025; Deng et al., 2025). These designs achieve remarkable engineering elegance, but they still prioritize structural and code-level interface unification over functional and behavioral coherence.

However, in this paper, we argue that this structural superficiality obscures a fundamental crisis. We unveil the "Unification Paradox": while modern architectures achieve seamless structural unification, they simultaneously aggravate a profound, systemic fragmentation across the entire functional stack. As illustrated in Figure 2, this paradox manifests as severe behavioral and cognitive dissonance. A model may correctly solve a visual puzzle in one trial but fail in an isomorphic setting, or generate a high-fidelity image that violates its own explicitly generated spatial-reasoning constraints (Zhou et al., 2025; Xu et al., 2026; Team, 2024). Crucially, this is not merely a surface-level alignment issue, but an architectural tax. As empirically demonstrated in TorchUMM (Luo et al., 2026), optimizing cur-

rent UMMs frequently triggers a zero-sum game: a significant performance leap on a specific isolated benchmark (e.g., visual understanding) may coincide with *catastrophic decoupling* in another core capability (e.g., generative fidelity). More counter-intuitively, TorchUMM's scaling analysis indicates that *a higher degree of structural unification does not linearly translate to superior systemic competence*. Without cross-layer coordination, forcing diametrically opposed objectives into a monolithic pipeline can exacerbate internal representation conflicts and gradient cannibalization.

To dismantle this structural myth, we redefine the ultimate goal of multimodal integration: True Unification must be a lossless, systemic gain, not a zero-sum architectural cost. If a capability degrades, vanishes, or becomes unmeasurable upon integration, it is structural co-location, not functional unification. The root cause of the Unification Paradox lies in the severe coordination cracks across five critical layers of the UMM functional stack: (1) Input Layer (Asymmetric Input Unification): Multimodal inputs possess intrinsic asymmetries in evidence reliability and information density. Current UMMs force them through identical sequence lengths or fixed structural pipelines, which paradoxically allows noisy or low-quality modalities to corrupt clean internal states, inducing overconfident hallucinated behaviors (Zollo and Zemel, 2025; Zitkovich et al., 2023). (2) Representation Layer (The Capacity Allocation Paradox): Understanding demands high-level semantic abstraction (lossy compression), whereas generation requires fine-grained reconstructive detail (lossless expansion). Forcing these orthogonal mandates into a unified shared latent space without structural factorization causes severe capacity encroachment—improving generation inherently erodes understanding, and vice versa. (3) Reasoning Layer (Stateful Reasoning Disregard): Current cross-modal reasoning is predominantly text-centric, treating visual tokens as passive static conditioning signals. The model lacks a unified, dynamic internal belief state to maintain multi-turn or cross-modal logical coherence, leading to severe attribute mixing and compositional failure during complex long-horizon decisions (Tang et al., 2025). (4) Post-Training Layer (Objective Cannibalization): The joint training of discriminative and generative paradigms relies on aggressive multi-task text-stitching. Without explicit gradient harmonization and Pareto-optimal reward shaping, the

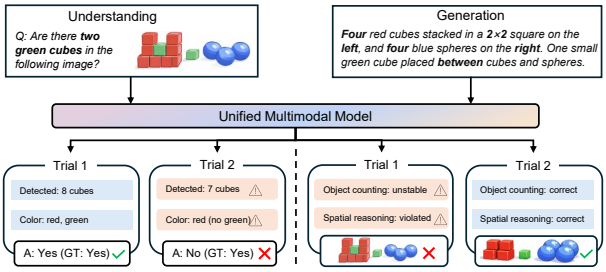

Figure 2: An illustration of Unification Paradox.

training objectives fiercely collide, tearing apart the model's internal consistency. (5) Evaluation Layer (The Single-Benchmark Illusion): Community assessments rely heavily on decoupled, outcome-oriented leaderboards (e.g., measuring understanding and generation through isolated metrics like VQA accuracy and FID). This single-benchmark optimization illusion can mask catastrophic decoupling inside structurally unified models, making local gains appear stronger than the unified system they are meant to evaluate.

Our position is that systemic competence and functional unification must supersede mere structural homogenization as the guiding principle for next-generation UMMs (Figure 3). True unification cannot be achieved by code-level and interface-level homogenization alone; it requires a concerted, full-stack orchestration of input calibration, factorized representation learning, stateful cross-modal reasoning, synchronized optimization, and holistic systemic evaluation. Through this lens, we chart a rigorous, evidence-driven path toward *truly coherent unified multimodal intelligence*, where capabilities are not merely structurally co-located, but functionally emergent.

2 Position Statement

We argue that structural co-location alone does not constitute functional emergence. As illustrated in Figure 3, next-generation UMMs should satisfy five full-stack *unification guarantees*:

(i). **Asymmetric Input Unification**: UMMs must explicitly quantify modality utility and reliability across modalities rather than blindly aggregate heterogeneous evidence. (§3)

(ii). **Factorized Representation Unification**: UMMs must allocate capacity across semantic abstraction and fine-grained reconstruction instead of forcing both into one monolithic latent space. (§4)

(iii). **Stateful Reasoning Unification**: UMMs must maintain evolving multimodal belief

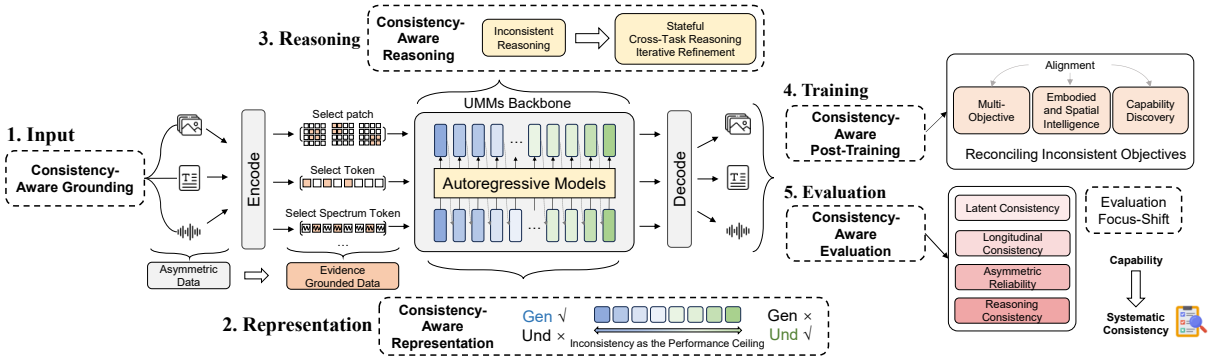

Figure 3: We argue that systemic competence should supersede structural unification as the priority for next-generation UMMs.

states beyond text-centric autoregressive outputs. (§5)

(iv). **Synergistic Optimization Unification**: Post-training must explicitly coordinate conflicting discriminative and generative objectives through robust gradient harmonization and Pareto-consistent reward shaping. (§6)

(v). **Multi-Centric Systemic Evaluation**: Evaluation must measure systemic competence, capability preservation, and cross-task robustness rather than isolated leaderboard wins or single-benchmark optimality (§7).

In short, true unification is not achieved by collapsing interfaces, but by jointly and actively coordinating input calibration, representation factorization, stateful reasoning, synergistic optimization, and systemic evaluation.

3 Asymmetric Input Unification

The first fracture of structural unification appears at the input level. Current UMMs often treat input unification as formatting: concatenate features, tokenize every modality into a shared sequence, or pass them through a fixed interface. Yet multimodal inputs exhibit inherent **evidence asymmetry** (Figure 4): modalities differ in information density, reliability, completeness, and task relevance (He et al., 2024; Zhang et al., 2024a). Thus, input unification should be dynamic evidence calibration rather than blind feature aggregation.

3.1 Challenges

The input-level unification paradox. Naive structural unification enforces homogeneous processing over heterogeneous evidence. Equal token allocation, uniform cross-attention, or fixed fusion pathways implicitly treat all modalities as similarly informative, allowing unreliable evidence to contaminate shared states. This yields two failures: adding modalities may *degrade* performance when noisy or irrelevant inputs inject alignment noise (Huang et al., 2022), and incomplete or contradictory evidence may trigger overconfident hallucination rather than calibrated uncertainty (Li et al., 2023; Zollo and Zemel, 2025; Zitkovich et al., 2023). Both arise because current UMMs unify interfaces without estimating evidence utility.

3.2 Research Directions

Dynamic input unification. True input unification requires task-aware utility modeling. A unified model should estimate which evidence is reliable, missing, or useful for the current objective; route unreliable signals away from shared states; and abstain when evidence is insufficient or mutually inconsistent. Because utility is contextual rather than intrinsic, this calibration must depend on the task, modality, and evidence state. The desired behavior is evidence-sensitive coordination: performance should improve when informative evidence is added, remain stable under irrelevant evidence, and degrade gracefully when critical evidence is missing. Missing evidence should not be treated as an ordinary token pattern to be imputed away: the model should propagate an explicit reliability state into later representation, reasoning, and generation modules. If the model is already confused about what evidence it has, later layers cannot reliably recover functional coherence.

Faithful generation under asymmetric evidence. Generation should condition on calibrated evidence states rather than fused vectors of unknown reliability. When visual input is noisy, text is underspecified, or modalities conflict, the model should avoid inventing unsupported details; calibrated abstention is therefore also a generative safety mechanism.

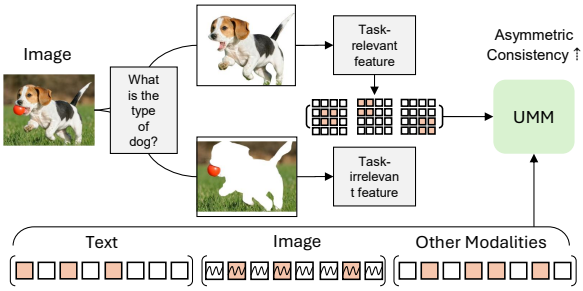

Figure 4: The input-level unification paradox: structurally homogeneous processing ignores asymmetric evidence quality, allowing unreliable modalities to corrupt unified internal states.

Existing ingredients point in this direction: token pruning and merging estimate patch utility (Rao et al., 2021; Ryoo et al., 2021; Bolya et al., 2023), mixture-of-experts enables conditional routing (Fedus et al., 2022), and selective prediction motivates risk-aware abstention (Geifman and El-Yaniv, 2019). Faithfulness benchmarks such as POPE and HallusionBench (Li et al., 2023; Guan et al., 2024), together with studies of language priors (Agrawal et al., 2018; Ramakrishnan et al., 2018), show why such utility estimates must target semantic reliability rather than efficiency alone.

4 Factorized Representation Unification

If input-level unification determines what evidence enters the model, representation-level unification determines what information can be faithfully preserved, transformed, and shared across capabilities. This is the core paradox of UMMs. Current models often assume that one shared latent space can simultaneously support understanding, reasoning, and generation, yet these capabilities impose conflicting representational requirements (Chen et al., 2025b; Team, 2024). The question is therefore what kind of representation can unify heterogeneous objectives without destructive interference.

4.1 Challenges

The representation unification paradox. Forcing all modalities and tasks into one latent space creates a clean interface but collapses two incompatible pressures. Understanding favors lossy semantic abstraction, while generation requires fine-grained reconstruction of layout, texture, color, and boundaries (Chow et al., 2024). If the representation is too semantic, details vanish; if it preserves too much low-level information, semantic reasoning becomes noisy. Scale cannot recover information erased at the encoding stage exposed in recent systems: decoupled pathways reduce capacity conflict but risk cross-branch inconsistency (Chen et al., 2025b; Chen, 2024; Deng et al., 2025; Fan et al., 2025), whereas highly unified representations are efficient but often imbalanced (Team, 2024; Xie et al., 2024). Chameleon discards fine visual detail during discretization (Team, 2024), and Tuna improves alignment while still leaning toward semantic encoders (Liu et al., 2025b). Thus, ideal representation unification is capacity allocation, not simple vector mixing.

4.2 Research Directions

From shared latent spaces to factorized unification. True representation unification should be factorized rather than monolithic. The goal is not to return to isolated task-specific systems, but to preserve semantic and reconstructive information in coordinated subspaces. A semantic subspace should support abstraction, reasoning, and grounding, while a reconstructive subspace should preserve spatial fidelity for generation. The key is the bridge: semantics must guide generation without letting reconstructive noise destabilize reasoning. This bridge should be selective: semantic states should constrain identity, relations, and intent during generation, while reconstructive states should contribute local geometry and appearance only when needed. Such exchange avoids both hard task splits and fully shared representations that let low-level perturbations enter reasoning space.

Theoretical foundations for capacity allocation. Classical results establish that downstream tasks can only extract information already preserved in the representation (Cover and Thomas, 2006); UMMs intensify this constraint because one representation is expected to jointly serve conflicting information requirements. Formalizing this constraint can clarify whether current gaps are architectural limitations or fundamental bounds (Yu et al., 2024). Since semantic and reconstructive information may occupy partially orthogonal subspaces (Huh et al., 2024; Tishby et al., 1999), UMMs need explicit policies for how much capacity to reserve for abstraction, reconstruction, and communication between them. These policies should be explicitly measurable: a factorized representation should preserve semantic invariance under texture or style changes while retaining enough reconstructive information for layout, boundaries, and fine visual attributes. If either property robustly collapses under joint training, the model has only

displaced the conflict to a hidden subspace.

Architectures for factorized representation unification. Promising directions include multi-scale features, task-adaptive subspaces, and factorized latent variables that separate semantic and reconstructive information while preserving cross-task interaction. This also changes how scaling should be interpreted: UMM scaling may depend as much on encoder capacity, subspace allocation, and bridge quality as on decoder size or data volume. Without such allocation, joint training may simply amplify the representation paradox. Unified representations should therefore be evaluated by *controlled interference*: generation-oriented changes should not degrade semantic probes, and semantic objectives should not erase reconstructive detail.

This view reframes existing design choices as partial responses to the same capacity problem: decoupled architectures reduce direct conflict but risk weak cross-branch reasoning (Chen, 2024; Chen et al., 2025b; Deng et al., 2025), while unified representations improve transfer yet may expose capability imbalance (Team, 2024; Xie et al., 2024). Recent alignment methods such as Tuna, RecA, UAE, UniGame, and T2I-R1 (Liu et al., 2025b; Xie et al., 2025a; Yan et al., 2025b; Su et al., 2025; Jiang et al., 2025) demonstrate the need for coordination, but factorized unification treats capacity allocation as an architectural principle rather than a post-hoc repair.

5 Stateful Reasoning Unification

While unified multimodal models (UMMs) have achieved promising capabilities in multimodal understanding and generation, their reasoning process remains fundamentally *text-centric*. Current UMMs reason through autoregressive token generation, where non-text modalities serve as conditional signals rather than active participants. Reasoning intelligence is therefore reflected in final text, while internal trajectories remain implicit and transient.

This exposes a fundamental **reasoning unification paradox**: although UMMs are trained under a unified architecture and latent space, their reasoning process is still fragmented into task-local autoregressive flows. Intermediate states are not explicitly preserved, revised, or shared across interactions: explanation, editing, and dialogue over the same scene are often treated as independent calls. Consistency is enforced at the output level rather than through persistent internal cognition.

True multimodal reasoning therefore requires **stateful reasoning unification**: persistent belief structures that can be accumulated, updated, verified, and geometrically anchored across modalities and tasks (Figure 5).

5.1 Challenges

The challenge is the mismatch between shared multimodal representations and unidirectional reasoning. Intermediate inferences are encoded only implicitly in generated tokens, so they are hard to explicitly preserve, revise, or reconcile with later evidence. In multi-step settings spanning understanding, explanation, editing, retrieval, and dialogue, this produces accumulated hallucinated revisions and cross-modal task-level contradictions.

5.2 Research Directions

Persistent Multimodal Belief States. Future UMMs should maintain structured belief states beyond autoregressive token generation. Memory systems, structured latent variables, or geometrically grounded feature spaces could make intermediate conclusions available for later reasoning.

Cross-Task Reasoning Consistency. Reasoning commitments formed in one task should propagate to downstream tasks. Object relations inferred during scene understanding, for instance, should constrain explanation, editing, and dialogue through shared, updateable belief states.

Dynamic Revision and Verification. Stateful reasoning should support revision rather than irreversible forward generation. Contradiction detection, selective rollback, and uncertainty-aware updates could transform multimodal reasoning into a self-correcting inference process.

Existing multimodal reasoning largely remains autoregressive and task-local (Zhang et al., 2025; Wang et al., 2025; Yan et al., 2025a; Wang et al., 2024a; Li et al., 2025a). Reflection and self-correction have improved language reasoning (Madaan et al., 2023; Ji and Lu, 2025), but analogous multimodal methods remain limited (Luo et al., 2025), especially when persistent states must survive across modalities and task boundaries.

6 Synergistic Optimization Unification

Post-training has become a critical stage for aligning foundation models with diverse downstream objectives, enabling substantial improvements in reasoning, instruction following, and generation quality (Christiano et al., 2017; Ouyang et al., 2022;

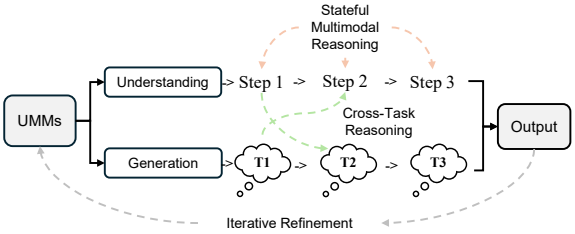

Figure 5: The reasoning unification paradox in UMMs. Although current UMMs operate under unified multimodal architectures, reasoning remains text-centric and autoregressive, lacking a persistent internal belief state across tasks. We envision stateful reasoning unification, where multimodal reasoning evolves through explicit, revisable, and cross-task consistent belief structures.

Rafailov et al., 2023). Parallel multimodal and generative work jointly shows similar benefits for image quality and controllability, but UMM post-training remains largely fragmented.

This reveals a core **optimization unification paradox**: while UMMs aim to jointly support multimodal understanding and generation, their alignment objectives are often stitched together through heterogeneous task-specific supervision. Discriminative objectives for perception and reasoning frequently conflict with generative or diffusion-based objectives for synthesis, producing inconsistent optimization signals and severe negative transfer. Post-training can therefore easily improve one capability at the expense of another.

Future UMMs require **synergistic optimization unification**: coordinated gradients, Pareto-aware rewards, and learning dynamics that make capabilities reinforce rather than cannibalize one another (Figure 6). This does not imply that every task should share every parameter update. Rather, post-training should decide when objectives should share gradients, when they should be isolated, and when one objective should regularize another. The central unit of alignment should therefore be a capability interaction, not a single task loss.

6.1 Challenges

Gradient Conflict Across Heterogeneous Objectives. Understanding tasks typically rely on discriminative supervision, emphasizing semantic accuracy and representation separability, whereas generation tasks optimize autoregressive or diffusion objectives that prioritize realism, diversity, and perceptual fidelity. These objectives can induce conflicting gradients in shared parameters: improving generation fidelity may degrade reasoning, while enhancing understanding can reduce generative diversity.

Lack of Unified Reward Signals. Existing post-training methods often rely on isolated reward proxies, including reconstruction objectives, image-text alignment scores, or task-specific preference models. Such signals evaluate local quality rather than global coherence across understanding, generation, editing, and reasoning; simplified proxies may therefore upper-bound unified intelligence.

Emergent Capability Underutilization. Current pipelines optimize capabilities independently, missing beneficial transfer: spatial perception can improve generation fidelity, while reconstruction can strengthen geometric understanding. Without synergistic objectives, latent multimodal capabilities may remain inaccessible.

6.2 Research Directions

Gradient Harmonization for Joint Optimization. Future UMMs should suppress destructive interference and amplify shared gradients through task-aware routing or modular optimization, so that specialized capabilities are preserved while the unified system remains globally consistent.

Pareto-Optimal Reward Shaping. Reward functions should model trade-offs and synergy regions among understanding, generation, reasoning, and editing instead of relying on manually tuned weights. Pareto-aware shaping can identify regimes where improvements reinforce rather than hinder one another. A practical reward model could score both local quality and cross-capability preservation: a generated image should be visually plausible, but also consistent with inferred spatial relations; an answer should be correct, but also preserve evidence needed for downstream editing or generation. This broader reward view turns alignment from a post-hoc preference filter into a coordination mechanism for the whole UMM stack.

Emergent Capability Discovery Through Unified Alignment. Coordinated post-training may reveal capabilities latent in pre-trained weights: jointly aligning spatial perception and reconstruction may induce implicit 3D reasoning, while coordinating understanding, generation, and editing may support self-consistent simulation.

Alignment for Spatial and Embodied Intelligence. Embodied agents require post-training objectives that jointly align perception, planning, and action generation under physical constraints. Likewise, unified 3D multimodal systems (Hu et al., 2025) demand optimization strategies ca-

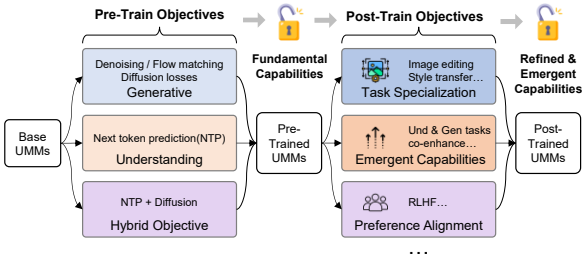

Figure 6: The optimization unification paradox in UMM post-training. Current approaches often stitch together heterogeneous objectives, causing gradient conflict and negative transfer between understanding and generation. We envision synergistic optimization unification, where gradients and rewards are coordinated to enable joint learning and robust multimodal capabilities.

pable of preserving geometric consistency across viewpoints and coordinate systems.

Prior post-training methods such as RLHF and DPO (Ouyang et al., 2022; Rafailov et al., 2023) and text-to-image preference optimization (Xu et al., 2023; Kirstain et al., 2023) remain task-specific. Recent UMM alignment efforts, including UniGame, RecA, UAE, and T2I-R1 (Su et al., 2025; Xie et al., 2025a; Yan et al., 2025b; Jiang et al., 2025), begin to bridge understanding and generation, but they still focus more on combining objectives than harmonizing gradients and rewards.

7 Multi-Centric Systemic Evaluation

Evaluation is the final layer where the Unification Paradox becomes visible. If UMMs are judged only by isolated task scores, structural unification can appear successful even when overall unified competence is weakening. Evaluation must therefore diagnose whether understanding, generation, reasoning, grounding, and alignment remain mutually compatible after unification, rather than inferring systemic competence from disconnected outcomes.

7.1 Challenges

Outcome-oriented fragmented metrics. Current UMMs are predominantly evaluated through isolated, task-specific leaderboards. Understanding is measured by VQA-style accuracy, language generation by BLEU-like scores (Papineni et al., 2002), image generation by FID-like perceptual metrics (Heusel et al., 2017), and reasoning by separate diagnostics. Such metrics track local skill, not mutual compatibility.

The single-benchmark optimality illusion. The fifth Unification Paradox is that a model can look better under one benchmark while becoming worse as a unified system. Empirical evidence from TorchUMM (Luo et al., 2026) suggests that UMM optimization often behaves like a zero-sum game: improving a model on one isolated benchmark, such as visual understanding, may coincide with a sharp decline in another capability, such as long-form generation or image fidelity. This pattern of *catastrophic decoupling* can remain invisible when evaluation records only local success, because the benchmark rewards the target task but not the collateral movement of the surrounding system.

The architectural myth. Evaluation must also expose the mistaken assumption that stronger structural unification automatically yields functional emergence. TorchUMM's analysis indicates that more architecturally unified designs do not necessarily produce linear performance gains (Luo et al., 2026). More importantly, it suggests that structural convergence and functional emergence should be treated as separate empirical questions. Without full-stack coordination, forcing understanding-oriented semantic compression and generation-oriented pixel reconstruction into the same pipeline can intensify representation conflict and gradient cannibalization (§4, §6). Output-only evaluation sees the final answer, but can miss the internal trade-offs introduced by hard architectural stitching.

7.2 Research Directions

Capability preservation matrices. UMM evaluation should measure collateral movement across capabilities after each architectural or training intervention. A cross-capability matrix can track understanding, generation, reasoning, grounding, editing, long-context language ability, and safety behavior, testing whether unification is a systemic dividend or an architectural tax. Reported across model variants and training stages, this matrix makes regressions visible: a model that improves VQA but loses image fidelity, editing stability, or long-context ability should not be described as more unified.

Pareto and zero-sum diagnostics. Instead of reporting single SOTA numbers, future evaluation should track Pareto frontiers across core UMM capabilities. If improving visual understanding consistently reduces generation fidelity, the model has not achieved functional unification; it has exposed an optimization conflict. Trade-off curves, negative-transfer coefficients, and capability interference scores convert this hidden zero-sum game into a measurable property (Luo et al., 2026). Such

diagnostics discourage misleading local optimization by asking not only "what is the best score?" but "what capability profile does that score require?"

Latent and longitudinal consistency. Systemic evaluation should probe mechanisms behind surface outcomes. Latent consistency tests whether representations balance semantic abstraction and reconstruction (§4), including intra-modal invariance (Aravindan et al., 2025) and cross-modal grounding geometry (Goel et al., 2022; Zhang et al., 2024b). Longitudinal consistency tests whether state remains coherent across interleaved understanding and generation, exposing semantic drift or contradictory reconstruction in round-trip generation (Kong et al., 2025; Xie et al., 2025a).

Stress tests for asymmetric and conflicting evidence. A systemic benchmark should include adversarial input perturbations that test whether the model remains calibrated under missing, imbalanced, or contradictory evidence (§3). Evidence monotonicity should check whether informative evidence improves performance while corrupted evidence induces uncertainty rather than hallucination. Conflict detection should test whether UMMs recognize modality conflicts without external supervision (Chen et al., 2025a; Ji et al., 2023); POPE and HallusionBench (Li et al., 2023; Guan et al., 2024) provide useful ingredients, but systemic evaluation must connect them to capability preservation. These stress tests can be paired with longitudinal protocols, such as question answering followed by editing and constraint checking, to reveal whether internal state, representation bridges, and generation pathways remain coordinated over time.

Recent benchmarks such as Uni-MMMU, MME-Unify, UmniBench, GGBench, GIR-Bench, Omni-Context, and UniSandBox (Zou et al., 2025; Xie et al., 2025b; Liu et al., 2025a; Wei et al., 2025; Li et al., 2025b; Wu et al., 2025; Niu et al., 2025) move toward mixed-modality and diagnostic settings. They mark important progress, but multi-centric systemic evaluation asks a stronger question: when one capability improves, what happens to the unified system? Answering this requires failure correlations, capability-preservation evidence, and cross-task robustness analysis, not leaderboard ranks alone.

8 Alternative Views

Argument: *Scaling/pre-training is all you need for unification.* Large-scale pre-training can improve cross-modal transfer (Team, 2024; Sun et al., 2024), but scale alone does not calibrate asymmetric evidence, factorize representation capacity, maintain stateful reasoning, harmonize conflicting objectives, or expose zero-sum trade-offs. Scaling may enlarge a structurally unified but functionally conflicted framework rather than correct it.

Argument: *Functional unification does not matter as long as final outputs are good.* Output-centric optimization is exactly what makes catastrophic decoupling difficult to detect. Functional unification matters because it preserves capabilities across tasks, supports reliable refinement, and prevents cascading failures in long-horizon multimodal reasoning.

9 Call to Action

We call on the community to treat systemic competence as a first-class design objective, not an emergent byproduct of structural homogenization. Model design should support asymmetric input calibration, factorized representations, and stateful reasoning; training should move from task-local refinement to gradient- and reward-level coordination; evaluation should report capability preservation, Pareto trade-offs, latent stability, longitudinal consistency, and cross-task robustness under asymmetric evidence. Papers should state which unification guarantees they target, and progress should be judged by whether capabilities remain coherent under transfer, perturbation, and joint optimization rather than by single-benchmark optimality alone.

10 Conclusion

This paper argued that current UMMs often achieve structural unification without functional emergence. Without full-stack coordination, input noise contaminates shared states, representation capacity is cannibalized, reasoning remains text-centric, post-training objectives collide, and isolated benchmarks can mask catastrophic decoupling. We proposed systemic competence as the organizing principle for next-generation UMMs, operationalized through five unification guarantees: asymmetric input unification, factorized representation unification, stateful reasoning unification, synergistic optimization unification, and multi-centric systemic evaluation. We hope this perspective reorients the field from assembling capabilities toward building genuinely unified systems.

11 Limitation

The proposed architectural roadmaps serve as high-level conceptual guidelines rather than fully deployed algorithmic implementations. Second, although the multi-centric systemic evaluation matrix offers a theoretical blueprint to diagnose performance trade-offs, building the standardized benchmark suite and normalizing heterogeneous task metrics remains an ongoing task that requires collective community engineering efforts.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of CVPR*.

Ashwath Vaithinathan Aravindan, Abha Jha, and Mihir Kulkarni. 2025. Do vlms have bad eyes? diagnosing compositional failures via mechanistic interpretability. In *ICCV*, pages 704–712.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token merging: Your vit but faster. In *Proceedings of ICLR*.

Tianle Chen, Chaitanya Chakka, Arjun Reddy Akula, Xavier Thomas, and Deepti Ghadiyaram. 2025a. Some modalities are more equal than others: Decoding and architecting multimodal integration in mllms. *arXiv:2511.22826*.

Xiaokang Chen. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *Preprint*, arXiv:2410.13848.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Wei Chow, Juncheng Li, Qifan Yu, Kaihang Pan, Hao Fei, Zhiqi Ge, Shuai Yang, Siliang Tang, Hanwang Zhang, and Qianru Sun. 2024. Unified generative and discriminative training for multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:23155–23190.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *NeurIPS*, 30.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. John Wiley & Sons.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, and 1 others. 2025. Emerging properties in unified multimodal pretraining. *arXiv:2505.14683*.

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, and 1 others. 2025. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of ICML*.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *NeurIPS*, 35:6704–6719.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385.

Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. 2024. Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25(47):1–39.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30.

JiaKui Hu, Shanshan Zhao, Qing-Guo Chen, Xuerui Qiu, Jialun Liu, Zhao Xu, Weihua Luo, Kaifu Zhang, and Yanye Lu. 2025. Omni-view: Unlocking how generation facilitates understanding in unified 3d model based on multiview images. *arXiv preprint arXiv:2511.07222*.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, pages 9226–9259. PMLR.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Jiazhou Ji and Xinru Lu. 2025. Reflair: Enhancing multimodal reasoning via structured reflection and reward-guided learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25401–25413.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *EMNLP*, pages 1827–1843.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. 2025. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663.

Lecheng Kong, Xiyuan Wang, Yixin Chen, and Muhan Zhang. 2025. Round-trip reinforcement learning: Self-consistent training for better chemical llms. *arXiv:2510.01527*.

Can Li, Ying Liu, Ting Zhang, Mei Wang, and Hua Huang. 2025a. Visiomath: Benchmarking figure-based mathematical reasoning in lmms. *arXiv preprint arXiv:2506.06727*.

Hongxiang Li, Yaowei Li, Bin Lin, Yuwei Niu, Yuhang Yang, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. 2025b. Gir-bench: Versatile benchmark for generating images with reasoning. *arXiv:2510.11026*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP*.

Kai Liu, Leyang Chen, Wenbo Li, Zhikai Chen, Zhixin Wang, Renjing Pei, Linghe Kong, and Yulun Zhang. 2025a. Umnibench: Unified understand and generation model oriented omni-dimensional benchmark. *arXiv:2512.17196*.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, and 1 others. 2025b. Tuna: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*.

Wenjie Luo, Ruocheng Li, Shanshan Zhu, and Julian Perry. 2025. Coherent multimodal reasoning with iterative self-evaluation for vision-language models. *arXiv preprint arXiv:2508.02886*.

Yinyi Luo, Wenwen Wang, Hayes Bai, Hongyu Zhu, Hao Chen, Pan He, Marios Savvides, Sharon Li, and Jindong Wang. 2026. Torchumm: A unified multimodal model codebase for evaluation, analysis, and post-training. *arXiv preprint arXiv:2604.10784*.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, and 1 others. 2025. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Yuwei Niu, Weiyang Jin, Jiaqi Liao, Chaoran Feng, Peng Jin, Bin Lin, Zongjian Li, Bin Zhu, Weihao Yu, and Li Yuan. 2025. Does understanding inform generation in unified multimodal models? from analysis to path forward. *arXiv:2511.20561*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741.

Sanjay Ramakrishnan, Aishwarya Agrawal, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. 2021. Tokenlearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhaolong Su, Wang Lu, Hao Chen, Sharon Li, and Jindong Wang. 2025. Unigame: Turning a unified multimodal model into its own adversary. *arXiv preprint arXiv:2511.19413*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. 2025. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*.

Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. 2025. Unilip: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv:2405.09818*.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Jingxuan Wei, Caijun Jia, Xi Bai, Xinglong Xu, Siyuan Li, Linzhuang Sun, Bihui Yu, Conghui He, Lijun Wu, and Cheng Tan. 2025. Ggbench: A geometric generative reasoning benchmark for unified multimodal models. *arXiv:2511.11134*.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, and 1 others. 2025. Omnigen2: Exploration to advanced multimodal generation. *arXiv:2506.18871*.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. 2025a. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. 2025b. Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models. *arXiv:2504.03641*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Pengcheng Xu, Peng Tang, Donghao Luo, Xiaobin Hu, Weichu Cui, Qingdong He, Zhennan Chen, Jiangning Zhang, Charles Ling, and Boyu Wang. 2026. Towards generalized multi-image editing for unified multimodal models. *arXiv preprint arXiv:2601.05572*.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2025a. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11798–11827.

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin, Hao Li, Xue Xu, and 1 others. 2025b. Unified multimodal model as auto-encoder. *arXiv preprint arXiv:2509.09666*.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2025. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2024. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and 1 others. 2024a. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024b. Cross-modal consistency in multimodal large language models. *arXiv:2411.09273*.

Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, and 1 others. 2025. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*.

Zijun Zhou, Yingying Deng, Xiangyu He, Weiming Dong, and Fan Tang. 2025. Multi-turn consistent image editing. *arXiv preprint arXiv:2505.04320*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

Thomas P Zollo and Richard Zemel. 2025. Confidence calibration in vision-language-action models. *arXiv preprint arXiv:2507.17383*.

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. 2025. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv:2510.13759*.