Resolving Task Objective Conflicts in Unified Model via Task-Aware Mixture-of-Experts

Extended Abstract

Jiaxing Zhang
Sichuan University
Chengdu, China
zjxing9972@gmail.com

Hao Tang
Peking University
Beijing, China
haotang@pku.edu.cn

ABSTRACT

Recently, multimodal understanding (MMU) and text-to-image generation (T2I) have been integrated into a single autoregressive (AR) architecture, achieving initial unification. However, existing works focus on representation-level studies and overlook potential conflicts in AR architectures' internal information flow during training different tasks. Motivated by this gap, we identify a deeper issue, **T**ask **O**bjective **C**onflict (TOC), arising from AR architectures' internal information flow, which causes negative transfer and catastrophic forgetting when training MMU and T2I jointly. To address this issue, we proposed **UniDecouple**, which decouples internal modules for different tasks to construct task-specific optimization subpaths. To implement UniDecouple, we employ a **T**ask-**A**ware **M**ixture **o**f **E**xperts (TA-MoE), comprising Hierarchical Expert Routing and Hybrid Expert Collaboration, trained in two stages: first to build task-specific experts, then jointly fine-tuned to balance specialization and overall coordination. Extensive experiments on both understanding and generation benchmarks demonstrate that UniDecouple preserves strong understanding ability while achieving generation quality comparable to state-of-the-art methods, offering a new perspective for unified modeling.

KEYWORDS

Unified Model; Autoregressive; Mixture-of-Experts

ACM Reference Format:
Jiaxing Zhang and Hao Tang. 2026. Resolving Task Objective Conflicts in Unified Model via Task-Aware Mixture-of-Experts: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/DLJD3800

1 INTRODUCTION

Human cognition, where understanding and generation intricately interleave, provides the foundation for reasoning and thought [5, 13, 16]. Inspired by this mechanism, researchers have sought unified models that integrate multimodal understanding and generation within a single framework. However, progress remains divided: autoregressive (AR) models [1, 7, 19, 20] dominate multimodal understanding (MMU), while diffusion models [8, 10, 15] excel at text-to-image generation (T2I). Recent works such as DALL·E [9] extend AR models to T2I, prompting further efforts [3, 4, 12, 14, 17] toward unified AR frameworks. Nevertheless, achieving a well-balanced trade-off between generation quality and understanding ability remains challenging. A widely recognized issue arises at the representation level: MMU requires high-level semantic abstraction, whereas T2I demands fine-grained detail preservation. Early approaches adopt shared or decoupled image encoders [2, 6, 17] to mitigate this conflict, but these methods primarily address surface-level representation mismatches.

Beyond representation, we identify a deeper challenge inherent to AR. By analyzing internal information flow, we observe that jointly optimizing MMU and T2I induces Task Objective Conflict (TOC) [11, 18], leading to negative transfer and catastrophic forgetting. Both theoretical analysis and empirical evidence confirm that AR-based joint training causes performance degradation across tasks, even when representation conflicts are alleviated.

To mitigate TOC, we propose UniDecouple, which integrates a Task-Aware Mixture of Experts (TA-MoE) into the AR paradigm. TA-MoE introduces task-specific routing to provide dedicated optimization paths for MMU and T2I, while enabling controlled collaboration through hierarchical routing and hybrid expert interaction. We further adopt a two-stage training strategy to balance expert specialization and overall coordination. Extensive experiments demonstrate that UniDecouple achieves competitive generation quality while effectively preserving multimodal understanding.

2 THE PROPOSED UNIDECOUPLE

Task Objective Conflict. In standard AR models, sequential dependency and causal attention create gradient interference: MMU mainly updates early token representations while T2I gradients flow from later tokens. The gradient overlap can produce negative inner products between the two tasks, leading to *negative transfer* and *catastrophic forgetting*, which limits multi-task learning under a single unified network.

Task-Aware Mixture of Experts (TA-MoE). To mitigate TOC, we introduce the TA-MoE layer, which partitions experts into task-specific groups for MMU and T2I, alongside a shared expert for cross-task knowledge exchange. Tokens are routed through a *Hierarchical Expert Routing* mechanism: a task-aware router first assigns tokens to the appropriate expert group, followed by a dynamic-assignment router that selects top-k experts within the group. The outputs of specific and shared experts are combined via weighted

Corresponding author: Hao Tang.

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/DLJD3800

aggregation, enabling task-specific processing while maintaining global coherence:

$$y = h_{\text{expert}} + \alpha \cdot \text{SharedExpert}(x), \tag{1}$$

where α is a learnable weight balancing shared and task-specific expert contributions.

Loss Function and Training Objective. The overall training objective combines task-specific losses and routing supervision:

$$\mathcal{L} = \sum_{t \in \{\text{MMU, T2I}\}} \lambda_t \, \mathcal{L}_t(y) + \gamma \, \mathcal{L}_{\text{group}}, \tag{2}$$

where λ_t and γ are hyperparameters, $\mathcal{L}_t$ denotes the cross-entropy loss for task t, and $\mathcal{L}_{\text{group}}$ supervises task-aware routing. This formulation ensures both expert specialization and accurate task assignment, mitigating conflicts between understanding and generation objectives.

Two-Stage Training Strategy. To fully exploit the complementary strengths of different experts, UniDecouple uses a two-stage training procedure. Stage 1 trains task-specific experts individually with frozen self-attention layers to obtain specialized FFNs:

$$\mathcal{W}^{t+1}_{\text{expert}} = \mathcal{W}^{t}_{\text{expert}} - \eta \nabla_{\mathcal{W}_{\text{expert}}} \mathcal{L}_{\text{gen}}. \tag{3}$$

Stage 2 replaces FFNs with TA-MoE and performs end-to-end LoRA-based fine-tuning (rank $r = 16$) jointly on both tasks. This parameter-efficient optimization allows UniDecouple to reconcile conflicting objectives, adapt to diverse tasks, and improve generalization.

3 EXPERIMENTS & RESULTS

Validation of TOC. To verify the existence of task objective conflicts (TOC), we compare single-task and multi-task variants on the POPE and GenEval benchmarks. The multi-task model exhibits slight performance drops, with antagonistic loss dynamics between understanding (cross-entropy) and generation (MSE), confirming inherent conflicts. UniDecouple addresses this through TA-MoE, effectively mitigating TOC while maintaining strong task-specific performance.

Table 1: Validation of TOC. "Und." and "Gen." denote understanding task and generation task.

	Task	POPE	GenEval	ΔPOPE	ΔGen.Eval
(a)	Und.	**86.9**	–	–	–
(b)	Gen.	–	**0.74**	–	–
(c)	Both	86.2	0.73	-0.7	-0.01

Quantitative Evaluation. UniDecouple preserves competitive MMU ability and achieves generation quality comparable to state-of-the-art methods. Table 2 highlights results on representative understanding and generation benchmarks. UniDecouple consistently outperforms prior multi-task models such as JanusPro, demonstrating the effectiveness of task-aware routing and shared expert design.

Table 2: UniDecouple performance vs. JanusPro on MMU (POPE, MMMU) and T2I (GenEval) benchmarks.

Model	POPE↑	MMMU↑	GenEval↑
JanusPro	86.2	36.3	0.73
UniDecouple (Ours)	**87.2**	**41.7**	**0.76**

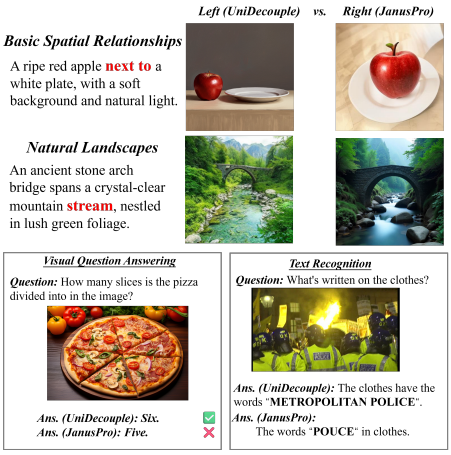

Figure 1: Qualitative comparison between UniDecouple and JanusPro on T2I (top) and MMU (bottom).

Qualitative Evaluation. In MMU tasks, UniDecouple demonstrates strong multimodal semantic comprehension, effectively handling VQA, visual description, and text recognition. In T2I tasks, UniDecouple produces coherent and detail-rich images across spatial reasoning, natural landscapes, human characters, and imaginative scenes. These results confirm that our TA-MoE design successfully balances high-level semantic modeling with fine-grained visual detail, mitigating task conflicts while preserving overall task quality.

Ablation and Expert Analysis. Ablation studies reveal that both the Task-Aware Router, which dynamically selects task-relevant experts, and the shared expert, which facilitates cross-task knowledge transfer, contribute substantially to performance gains. Incorporating a two-stage training strategy not only accelerates convergence but also enhances both understanding and generation results. Experiments on different expert ratios indicate that a 2:1 balance between group-specific and shared experts yields the best trade-off, allowing the model to effectively adapt to task-specific demands while maintaining robust generalization across multimodal tasks.

4 CONCLUSION

In this paper, we address the challenge of Task Objective Conflicts in unified multimodal autoregressive models. We propose UniDecouple, a novel framework that decouples internal modules and constructs task-specific optimization subpaths, effectively mitigating negative transfer between multimodal understanding (MMU) and text-to-image generation (T2I) tasks. Central to our approach is the Task-Aware Mixture of Experts (TA-MoE), which combines Hierarchical Expert Routing and Hybrid Expert Collaboration, trained with a two-stage strategy to balance task specialization and overall coordination. Extensive experiments demonstrate that UniDecouple preserves strong understanding capabilities while achieving generation quality on par with state-of-the-art methods. Our results highlight the importance of disentangling task-specific pathways in unified models and provide a promising direction for future research on efficient, high-performing multimodal architectures.

REFERENCES

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* 1, 2 (2023), 3.

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

[3] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135* (2024).

[4] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499* (2023).

[5] Raisa Islam and Owana Marzia Moushi. 2025. Gpt-4o: The cutting-edge advancement in multimodal llm. In *Intelligent Computing-Proceedings of the Computing Conference*. Springer, 47–60.

[6] Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. 2025. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422* (2025).

[7] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

[8] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[9] Mr D Murahari Reddy, Mr Sk Masthan Basha, Mr M Chinnaiahgari Hari, and Mr N Penchalaiah. 2021. Dall-e: Creating images from text. *UGC Care Group I Journal* 8, 14 (2021), 71–75.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[11] Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems* 31 (2018).

[12] Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818* (2024).

[13] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2024. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164* (2024).

[14] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. 2024. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673* (2024).

[15] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869* (2024).

[16] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848* (2024).

[17] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. 2024. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429* (2024).

[18] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems* 33 (2020), 5824–5836.

[19] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).

[20] Yichen Zhu, Minjie Zhu, Ning Liu, Zhiyuan Xu, and Yaxin Peng. 2024. Llava-phi: Efficient multi-modal assistant with small language model. In *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*. 18–22.