HOURGLASS: TOWARDS BETTER UNIFIED MULTIMODAL MODELING THROUGH HIERARCHICAL SHARING AND CONFLICT-AWARE GRADIENT SURGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) have attracted growing attention due to their unified task interfaces and potential for broader generalization. However, existing unified methods largely follow either partially shared designs with limited cross-task coordination or fully shared architectures that may induce capacity competition and gradient interference. To address these limitations, we propose Hourglass, a unified framework that facilitates beneficial collaboration at both the architectural and optimization levels. Motivated by the semantic complementarity and task heterogeneity of visual understanding and generation, Hourglass employs a hierarchical decoupled-shared-decoupled design, retaining task-specific processing in both shallow and deep stages while introducing a selective intermediate sharing mechanism. This structures a balanced trade-off that promotes beneficial cross-task collaboration while preserving task specialization. Nevertheless, jointly training the shared parameters may expose them to conflicting gradients stemming from the distinct natures of understanding and generation. To mitigate this, we further present CocoGrad, a gradient surgery method that dynamically partitions the shared-parameter coordinates into collaborative and conflicting subspaces, preserving common descent directions while coordinating conflicting updates. Extensive experiments across various visual understanding and generation benchmarks demonstrate that our method achieves state-of-the-art or highly competitive performance, providing a viable and scalable blueprint for UMMs.

1 INTRODUCTION

Recent advances in multimodal large language models Liu et al. (2023b); Li et al. (2024); Deitke et al. (2025); Team (2026b), alongside diffusion models and generative Transformers Podell et al. (2024); Team et al. (2026); Peebles & Xie (2023); Wu et al. (2025a), have substantially enhanced the capabilities of visual understanding and generation, respectively. Compared to these isolated paradigms, building a unified multimodal model (UMM) Xie et al. (2025); Deng et al. (2025); Tian et al. (2026) holds greater promise in terms of shared semantic representations, unified task interfaces, and overall system generality Yang et al. (2025). Recent unified models have explored objective-level unification, shared visual representations, and different degrees of architectural coupling. However, architectures that enable visual understanding and generation to learn complementary semantic relationships from each other while preserving task-specific processing remain underexplored.

Existing UMMs generally fall into two categories. The first paradigm, represented by BAGEL Deng et al. (2025) (Figure 1(a)), integrates both tasks within a single framework while maintaining a partially shared internal structure. While preserving task specialization to some extent, its cross-task coordination is restricted to sparse self-attention sharing or simple token concatenation. Alternatively, the second paradigm Xie et al. (2026); Liu et al. (2026) (Figure 1(b)) further pursues a fully shared design, enforcing more thorough unification across representation space and backbone architecture. Although this paradigm improves overall consistency, full parameter sharing introduces not only capacity competition but also optimization interference, making it difficult to simultaneously preserve the performance upper bound of both tasks Luo et al. (2026). Consequently, existing methods unify tasks structurally but fail to systematically address beneficial sharing mechanisms.

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

These observations motivate us to explore what to share and where to share within a unified multimodal backbone. Visual understanding and generation contain complementary semantic modeling signals, while requiring task-specific processing for their heterogeneous feature transformations and output objectives. Therefore, we propose Hourglass, a funnel-shaped UMM that realizes selective sharing along both the network-depth and transformation dimensions. As illustrated in Figure 1(c), Hourglass adopts a hierarchical decoupled-shared-decoupled organization: shallow layers preserve task-specific feature processing, intermediate layers introduce selective parameter sharing, and deep branches accommodate the distinct output objectives of understanding and generation. This structured design avoids both uniformly isolated and fully shared parameterization, striking a better balance between cross-task interaction and task specialization.

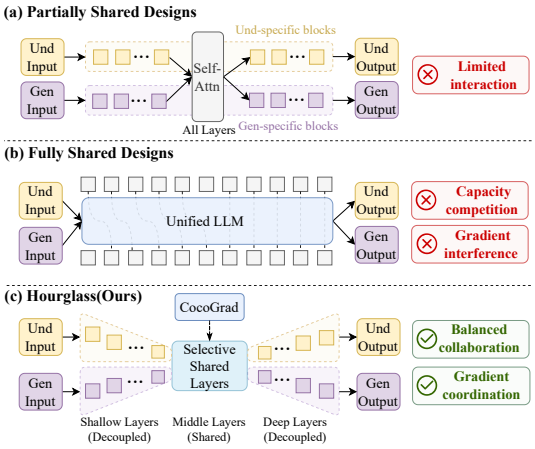

Figure 1: Comparison of representative unified architectures.

However, hierarchical sharing alone does not guarantee effective joint learning. As shown in Figure 2(a), joint optimization yields a higher loss on its corresponding objective than optimizing each task independently, suggesting that multi-task training introduces optimization interference. Figure 2(b) explains this phenomenon: inter-task gradients with opposing sign frequently outweigh the same counterparts, inducing interference during training. This motivates us to separate conflicting and collaborative gradient components, preserving the potentially beneficial common first-order descent directions contributed by collaborative components, while introducing targeted additional coordination for conflicting components that induce a trade-off between the two tasks. Therefore, we introduce CocoGrad, a conflict-aware gradient surgery method for unified models. CocoGrad dynamically partitions the shared-parameter space into collaborative and conflicting subspaces according to the directional relationship between task gradients. Within the collaborative subspace, cross-task synergy is preserved. Meanwhile, CocoGrad projects a new update direction to better balance understanding and generation in conflicting subspaces, as illustrated in Figure 2(c).

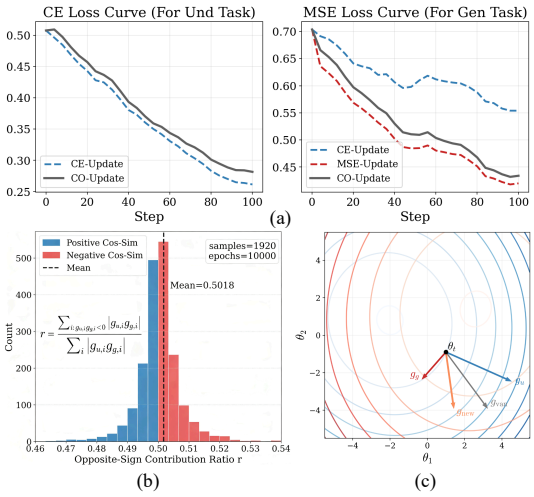

Figure 2: (a) Actual CE and MSE loss curves under different update strategies. "CE-Update" and "MSE-Update" denote updating with CE loss gradients g_u and MSE loss gradients g_g, while "CO-Update" denotes joint updates using both gradients g_{van}. (b) Actual distribution of opposite-sign contribution ratio between g_u and g_g during training, where i indexes dimensions. A ratio above 0.5 means a negative cosine similarity and thus gradient conflict. (c) Hypothetical contour plot of the optimization landscape, with the optimized update direction g_{new} from CocoGrad.

Ultimately, Hourglass integrates the hierarchical structure with conflict-aware optimization, establishing a comprehensive paradigm for unified visual understanding and generation. We conduct systematic experiments on multiple benchmarks for visual understanding and generation. The results show that our method achieves remarkable performance on both understanding and generation. Our contributions are summarized as follows:

- We propose Hourglass, a UMM with a hierarchical decoupled-shared-decoupled architecture tailored to the complementary yet heterogeneous nature of visual understanding and generation, balancing cross-task collaboration and task specialization.
- We introduce CocoGrad, a conflict-aware gradient surgery strategy. By dynamically partitioning the gradient space, CocoGrad applies subspace-specific updates to preserve consistent optimization signals while mitigating inter-task interference.
- We conduct extensive experiments across multiple benchmarks and achieve state-of-the-art or highly competitive performance, validating the effectiveness of our designs.

2 Related Work

2.1 Unified Multimodal Models

Existing UMMs can be broadly categorized into two classes. The first maintains partially shared internal structures. Mogao Liao et al. (2025) and BAGEL Deng et al. (2025) adopt a Mixture-of-Transformers design, where sharing is introduced in the self-attention layers. JoyAI-Image Song et al. (2026) and InternVL-U Tian et al. (2026) unify by connecting MLLMs with MMDiT, using a shared multimodal interface to interact. Although these approaches preserve task specialization, cross-task synergy remains shallow and limited to module coupling.

The second paradigm pursues full sharing. Transfusion Zhou et al. (2025) and Show-o Xie et al. (2025) employ a single Transformer to jointly support autoregressive understanding and discrete diffusion-based generation. Harmon Wu et al. (2025b), Show-o2 Xie et al. (2026) and TUNA Liu et al. (2026) further adopt unified visual representations. Compared with partially shared approaches, these models achieve stronger unification and compactness. However, overly strong unification may cause tasks to fight for parameter allocation and introduce gradient interference.

Despite these advances, prior UMMs do not systematically address how to achieve effective sharing. To tackle this challenge, we propose a hierarchical architecture to more appropriately balance commonality and difference across different layers, enabling effective collaboration between understanding and generation.

2.2 Gradient Surgery Strategies

Joint optimization of shared parameters is a central issue in multitask learning. In multitask settings, shared parameters are updated simultaneously by gradients from multiple objectives, often leading to gradient conflict, negative transfer, and unstable training Sener & Koltun (2019). To alleviate these issues, a variety of general optimization methods have been proposed. PCGrad Yu et al. (2020) projects each task gradient onto the orthogonal direction of other task gradients and averages them after clipping conflicting components. GradNorm Chen et al. (2018) dynamically adjusts loss weights according to task difficulty or gradient magnitude. CAGrad Liu et al. (2021) searches for an update direction that reduces the average loss while ensuring descent for every task, enjoying a theoretical guarantee of convergence to a minimum. CAME-Grad Zhang et al. (2026a) mitigates gradient conflict and optimization energy decay in shared parameters through conflict-avoiding direction correction, magnitude-enhancing energy injection, and adaptive gradient fusion.

However, existing methods overlook fine-grained collaborative and conflicting patterns within shared gradients, and some require iterative solvers, incurring extra computational overhead. In contrast, CocoGrad partitions shared gradients into collaborative and conflicting subspaces, introduces a subspace-aware, task-coordinated gradient update formula and derives a closed-form solution.

3 Method

3.1 Overall Architecture of Hourglass

Figure 3 illustrates the overall architecture of Hourglass, which supports both visual understanding and generation. The understanding branch processes text tokens and image features extracted by a ViT encoder. The generation branch takes VAE image latents as input, conditioned on the text

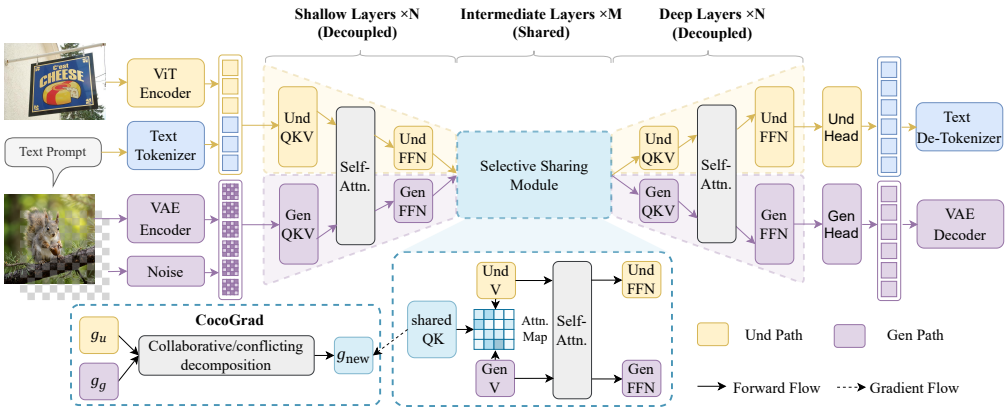

Figure 3: Overall framework of Hourglass. Hourglass unifies visual understanding and generation through a hierarchical decoupled-shared-decoupled backbone, together with CocoGrad for coordinating the joint optimization of shared parameters through collaborative and conflicting subspaces.

or images from the understanding branch according to the generation task. All tokens are packed into a unified sequence and encoded with MRoPE Bai et al. (2025b) to preserve their modality-specific positional structures. Hourglass adopts a hybrid attention strategy. Specifically, for visual and textual tokens in the understanding branch, we employ causal attention Vaswani et al. (2017); for generated image tokens, we use bidirectional attention while maintaining sample-wise isolation. Structurally, the backbone is organized into three hierarchical tiers: a shallow decoupled stage, an intermediate shared stage, and a deep task-adaptive stage.

Shallow Decoupled Stage. Understanding and generation are modeled with independent attention projections and feed-forward neural networks (FFNs). This design preserves the distinct low-level transformations required by the two tasks and avoids premature parameter coupling in the early representation learning stage.

Intermediate Shared Stage. The understanding and generation branches share the query-key transformations, while retaining independent value projections. The shared query-key transformations provide a common mechanism for modeling attention relationships, whereas the branch-specific components preserve the distinct feature processing required by each task.

Deep Task-Adaptive Stage. Independent branches are designed for understanding and generation to accommodate their different objective forms. For the understanding branch, the model outputs text tokens autoregressively. The training objective adopts the next-token prediction (NTP) loss:

$$\mathcal{L}_{\text{und}} = -\frac{1}{T}\sum_{t=1}^{T}\log p_\theta(y_t \mid y_{<t}, x_{\text{und}}), \tag{1}$$

where x_{und} denotes the input for the understanding branch, y_t is the t-th token in the target sequence, and T represents the length of the target sequence.

For the generation branch, we adopt flow matching to model the visual generation process. Specifically, let z_1 denote the clean VAE latent, $z_0 \sim \mathcal{N}(0, I)$ denote noise. The intermediate latent state is defined as $z_t = tz_1 + (1-t)z_0$, where $t \sim \mathcal{U}(0, 1)$. Along this linear interpolation trajectory, the model learns to predict the corresponding velocity field Lipman et al. (2023). The training objective is formulated as:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{z_0, z_1, t}\left\| v_\theta(z_t, t, x_{\text{gen}}) - (z_1 - z_0)\right\|_2^2, \tag{2}$$

where x_{gen} denotes the text or image conditioning input to the generation branch.

The final overall training loss is a weighted sum of the understanding loss and generation loss:

$$\mathcal{L} = \lambda_{\text{und}}\mathcal{L}_{\text{und}} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}}. \tag{3}$$

Algorithm 1 CocoGrad Algorithm

Require: Dataset $\mathcal{D}$, shared parameters θ_s, weights λ_{und} and λ_{gen}, radius c

while not converged **do**
  Sample a mini-batch $\mathcal{B} \sim \mathcal{D}$
  Compute task losses $\mathcal{L}_{\text{und}}$ and $\mathcal{L}_{\text{gen}}$ on $\mathcal{B}$
  Compute gradients $g_u \leftarrow \lambda_{\text{und}} \nabla_{\theta_s} \mathcal{L}_{\text{und}}$, $g_g \leftarrow \lambda_{\text{gen}} \nabla_{\theta_s} \mathcal{L}_{\text{gen}}$
  Construct masks $\mathbf{S} \leftarrow \mathbb{I}[g_u \odot g_g \geq 0]$, $\mathbf{C} \leftarrow \mathbf{1} - \mathbf{S}$
  Extract conflicting components $g_u^C \leftarrow \mathbf{C} \odot g_u$, $g_g^C \leftarrow \mathbf{C} \odot g_g$
  Compute β_0 and ρ
  Clip coefficient $\beta^* \leftarrow \text{clip}(\beta_0, -\rho, \rho)$
  $\delta_C \leftarrow g_u^C + g_g^C + \beta^*(g_u^C - g_g^C)$
  Compose gradient $g_{\text{new}} \leftarrow \mathbf{S} \odot (g_u + g_g) + \mathbf{C} \odot \delta_C$
  Update θ_s with g_{new}
end while

3.2 Conflict-Aware Gradient Surgery

The shared parameters in the intermediate layers are jointly updated by gradients from both objectives. However, these gradient updates are not always uniformly compatible. To mitigate inter-task interference, we propose CocoGrad, a conflict-aware gradient surgery method to optimize shared parameter updates, as summarized in Algorithm 1.

Denoting the shared parameters as θ_s, the weighted task gradients with respect to θ_s are defined as

$$g_u = \lambda_{\text{und}} \nabla_{\theta_s} \mathcal{L}_{\text{und}}, \qquad g_g = \lambda_{\text{gen}} \nabla_{\theta_s} \mathcal{L}_{\text{gen}}. \tag{4}$$

Under vanilla joint optimization, the shared parameters are updated using the aggregated gradient $g_{\text{van}} = g_u + g_g$. However, this uniform aggregation does not distinguish between mutually consistent and conflicting gradient signals, which causes gradient cancellation and degrades the optimization progress of both tasks. Therefore, we partition the coordinate space of the shared parameters into two mutually exclusive subspaces: a collaborative subspace, where the task gradients are directionally consistent, and a conflicting subspace, where their directions disagree. For the i-th parameter coordinate, the corresponding collaborative and conflicting masks, which satisfy $\mathbf{S} + \mathbf{C} = 1$, are defined as

$$\mathbf{S}_i = \mathbb{I}\left[g_{u,i} g_{g,i} \geq 0\right], \qquad \mathbf{C}_i = \mathbb{I}\left[g_{u,i} g_{g,i} < 0\right]. \tag{5}$$

The shared parameters in the intermediate layers are jointly updated by gradients from both objectives. However, these gradient updates are not always uniformly compatible. To mitigate inter-task interference, we propose CocoGrad, a conflict-aware gradient surgery method for optimizing shared parameter updates.

Based on this decomposition, the original aggregated gradient is retained in the collaborative subspace to preserve consistent optimization signals. In the conflicting subspace, an adjustable direction is introduced around the aggregated gradient to mitigate interference. Consequently, we modify the vanilla update gradient to the following form:

$$g_{\text{new}} = \mathbf{S} \odot (g_u + g_g) + \mathbf{C} \odot \left[(g_u + g_g) + \beta\,(g_u - g_g)\right], \tag{6}$$

For the conflicting subspace, let $g_u^C = \mathbf{C} \odot g_u$ and $g_g^C = \mathbf{C} \odot g_g$. The coordinated gradient δ_C is parameterized as

$$\delta_C = g_u^C + g_g^C + \beta\left(g_u^C - g_g^C\right), \tag{7}$$

where β denotes the coordination coefficient, which controls the adjustment toward either task gradient.

For a sufficiently small update step, the first-order decreases of the understanding and generation losses are proportional to the inner products between their gradients and the update direction, $(g_u^C)^\top \delta_C$ and $(g_g^C)^\top \delta_C$, respectively. We determine β by maximizing the smaller descent gain of the two tasks while constraining the coordinated update to remain close to the original aggregated direction:

$$\begin{aligned} \max_{\beta} \quad & \min\left\{(g_u^C)^\top \delta_C,\, (g_g^C)^\top \delta_C\right\}, \\ \text{s.t.} \quad & \left\|\delta_C - (g_u^C + g_g^C)\right\| \leq c\left\|g_u^C + g_g^C\right\|, \end{aligned} \tag{8}$$

where $c \geq 0$ denotes the conflict-tolerance radius. This one-dimensional problem admits the closed-form solution

$$\beta^* = \mathrm{clip}(\beta_0, -\rho, \rho), \tag{9}$$

where $\mathrm{clip}(\beta_0, -\rho, \rho)$ restricts β_0 to the interval $[-\rho, \rho]$, with β_0 and ρ given by

$$\beta_0 = \frac{\|g_g^C\|^2 - \|g_u^C\|^2}{\|g_u^C - g_g^C\|^2 + \epsilon}, \qquad \rho = \frac{c\|g_u^C + g_g^C\|}{\|g_u^C - g_g^C\| + \epsilon}, \tag{10}$$

where ϵ is a small constant for numerical stability. The complete derivation and the convergence analysis are provided in Appendix A and B.

3.3 Training Recipes

We initialize Hourglass from Qwen3-VL-Instruct Bai et al. (2025b) and design a progressive three-stage strategy. In pretraining (PT), we freeze the understanding branch and train only the generation branch on generation and editing samples to inject foundational generation capabilities. In continued pretraining (CT), with the understanding branch still frozen, we scale up input resolution and leverage high-aesthetic image-text corpora recaptioned by Qwen3.5 Team (2026a) alongside synthetic samples to enhance generation quality and compositional reasoning. In supervised fine-tuning (SFT), we unfreeze the entire model for joint training on high-precision supervised data, applying CocoGrad to mitigate inter-task interference. Overall, Hourglass is trained on 67M mostly open-source samples across all stages. Detailed training settings and datasets are provided in Table 1 and Appendix C, respectively.

Table 1: Hyperparameter settings for PT, CT, and SFT stages. LW: loss weight; G: text-to-image data; E: editing data; U: understanding data.

Configuration	PT	CT	SFT
Training part	Gen	Gen	Und & Gen
Optimizer	Adam	Adam	Adam & CocoGrad
Scheduler	Constant	Cosine	Cosine
LR	1e-4	1e-4	2.5e-5
Timestep Shift	1.0	3.0	3.0
LW (CE:MSE)	–	–	0.25:1
Gen Resolution	256×256	512×512	512×512
Training tokens	1.3B	0.8B	0.2B
Training steps	130K	200K	30K
Sample Ratio	G:E=4:1	G:E=1:1	U:G:E=6:1:1

4 Experiments

4.1 Experimental Setup

We evaluate two model variants, Hourglass-3B and Hourglass-7.5B, initialized from Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct respectively, while preserving their native text tokenizers and conversational formats. We adopt the pretrained FLUX VAE Labs (2024) as the image generation encoder, and keep all visual encoders frozen throughout training. The downsampling factor is set to 16 for the generation task and 32 for the understanding task. 12 transformer blocks serve as intermediate shared layers. We apply classifier-free guidance (CFG) Ho & Salimans (2022) to both image and text conditions. During training, text and image conditions are randomly dropped with probabilities of 10% and 30%, respectively. During inference, we use 50 sampling steps for image generation. The CFG scales for dropping the image condition and text condition are set to 1.0 and 4.0, respectively, and the unguided interval is [0.4, 1.0]. For understanding tasks, we set top-k of 20, top-p of 0.8, and temperature of 0.7.

4.2 Overall Performance

As shown in Table 2, we compare Hourglass with both task-specific and state-of-the-art unified models across visual understanding, text-to-image generation, and image editing benchmarks. Detailed evaluation results and more visual examples are provided in Appendices D and F.

Image Understanding. We evaluate Hourglass on MME Fu et al. (2026), MMBench Liu et al. (2024), and MMMU Yue et al. (2024). Hourglass-7.5B achieves the highest MME-P(1709), MME-S(2397) and MMBench(88.4), surpassing unified baselines in visual perception and reasoning. Hourglass-3B also performs strongly among similarly sized models. The qualitative examples in

Table 2: Comprehensive comparison of models across understanding, generation and editing benchmarks. Among unified models, bold denotes the best performance, and underlined denotes the second-best performance. "*" denotes the reproduced results, using the same hyperparameters as Hourglass. All other results are taken from official sources.

Model	Date	Params	Image Understanding				Text-to-Image		Image Editing
			MME-P	MME-S	MMBench	MMMU	GenEval	DPGBench	ImgEdit
Understanding-Only Models									
InternVL1.5 Chen et al. (2024b)	2024.4	2.2B	–	1900	70.9	34.6	–	–	–
LLaVA-OV Li et al. (2024)	2024.8	7B	1580	1998	80.8	48.8	–	–	–
InternVL3.5 Wang et al. (2025a)	2025.8	2B	1552	2123	78.4	59.0	–	–	–
Qwen3-VL-Instruct Bai et al. (2025b)	2025.9	2B	1523*	2069*	77.8	53.4	–	–	–
Generation-Only Models									
SD3-Medium Esser et al. (2024)	2024.6	2B	–	–	–	–	0.74	84.08	–
FLUX.1-dev Labs (2024)	2024.8	12B	–	–	–	–	0.82	84.00	–
GPT-Image-1 [High]	2025.4	–	–	–	–	–	0.84	85.15	4.20
Qwen-Image Wu et al. (2025a)	2025.8	20B	–	–	–	–	0.87	88.32	4.27 (Edit)
Z-Image Team et al. (2026)	2026.1	6B	–	–	–	–	0.84	88.14	4.30 (Edit)
Unified Multimodal Models									
SEED-X Ge et al. (2025)	2024.4	13B	1457	–	70.1	35.6	0.51	–	–
Emu3 Wang et al. (2024)	2024.9	8B	–	–	58.5	31.6	0.66	80.60	–
Janus-Pro Chen et al. (2025b)	2025.1	7B	1567	–	79.2	41.0	0.80	84.19	–
BAGEL Deng et al. (2025)	2025.5	14B	<u>1687</u>	<u>2388</u>	<u>85.0</u>	<u>55.3</u>	0.82	85.07	3.20
Show-o2 Xie et al. (2026)	2025.6	7B	1620	–	79.3	48.9	0.76	**86.14**	–
UniWorld-V1 Lin et al. (2025)	2025.6	20B	–	–	–	**58.6**	0.80	81.38	3.26
OmniGen2 Wu et al. (2026)	2025.6	7B	–	–	79.1	53.1	0.80	83.57	3.44
Tar Han et al. (2025)	2025.6	7B	1571	1926	–	39.0	0.84	84.19	–
Cheers Zhang et al. (2026b)	2026.3	1.5B	–	–	70.4	36.0	0.78	83.48	–
InternVL-U Tian et al. (2026)	2026.4	3.7B	1608	–	–	54.7	0.85	<u>85.18</u>	**3.67**
Hourglass	**2026.9**	**3B**	1624	2135	79.6	47.4	<u>0.87</u>	83.30	3.47
Hourglass	**2026.9**	**7.5B**	**1709**	**2397**	**88.4**	54.2	**0.89**	84.23	<u>3.55</u>

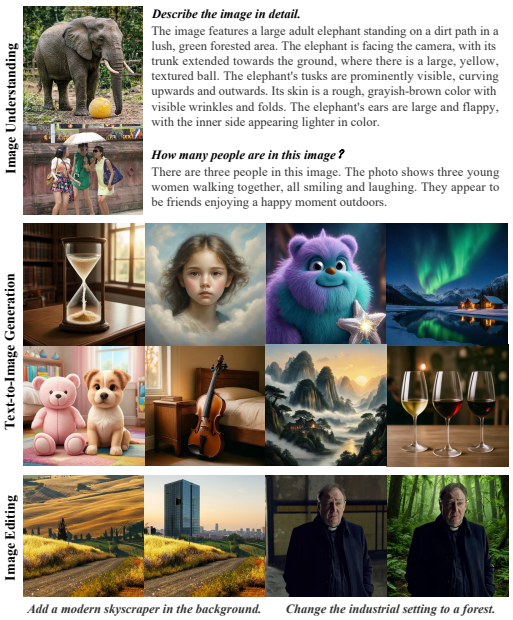

Figure 4: Qualitative results of Hourglass on multimodal understanding and generation.

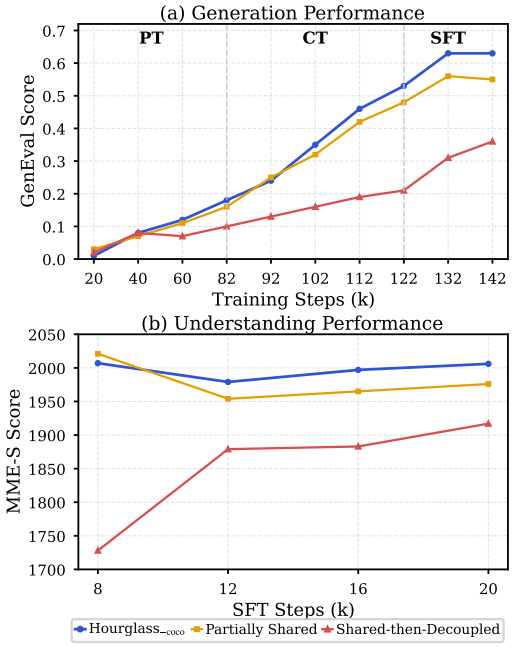

Figure 5: Comparison of sharing architectures across training stages.

Figure 4 show that Hourglass can accurately recognize objects and attributes, describe complex visual content, and perform visual counting.

Text-to-Image Generation. Evaluated on GenEval Ghosh et al. (2023) and DPG-Bench Hu et al. (2024), Hourglass-7.5B and Hourglass-3B achieve the highest and second-highest GenEval scores of 0.89 and 0.87, respectively. Notably, the 3B model already matches the 20B Qwen-Image, while

the 7.5B model further surpasses it, demonstrating strong capabilities in complex compositions and spatial relations. Both also deliver competitive DPG-Bench scores, reflecting robust dense-prompt understanding and generation. As qualitative examples in Figure 4 show, Hourglass generates high-quality images across diverse artistic styles and complex compositions, faithfully rendering fine details, spatial relations, and object attributes.

Image Editing. Tested via ImgEdit Ye et al. (2026), Hourglass-7.5B ranks second among UMMs with a score of 3.55, indicating robust and reliable performance in instruction-driven editing. As shown in Figure 4, the model achieves strong instruction following while maintaining high fidelity to the original image.

4.3 Ablation Study

Except for the scaling-up analysis, all ablation studies are conducted on Hourglass-3B. To disentangle the contributions of the proposed architecture and CocoGrad from those of the large-scale training recipe, all variants within each ablation use the same training data, initialization, training stages, and other experimental settings, with only the component under investigation being changed. Detailed configurations are provided in Appendix E.

Effect of Sharing Placement. To investigate where beneficial sharing should occur, we consider four possible configurations: no shared layers (*Partially Shared*), sharing in the shallow layers (*Shared-then-Decoupled*), sharing in the intermediate layers (our *Hourglass*$_{-\text{coco}}$), and sharing in the deep layers. Since the deep layers need to accommodate the increasingly distinct output objectives of understanding and generation, placing shared layers at the end is less reasonable. We therefore evaluate the first three configurations, as shown in Figure 5. Here, Hourglass$_{-\text{coco}}$ denotes the decoupled-shared-decoupled backbone without CocoGrad. To ensure a fair comparison, CocoGrad is disabled across all variants. All models are trained with shortened PT, CT, and SFT stages to improve efficiency. Among the evaluated variants, Hourglass$_{-\text{coco}}$ achieves the best final performance on both GenEval and MME-S. The shared-then-decoupled variant performs worse, demonstrating the importance of preserving task-specific processing in the shallow layers. The partially shared variant is competitive in early training but improves more slowly in later stages. These results support the proposed placement of shared layers between task-specific shallow and deep stages, which provides a more effective balance between cross-task collaboration and task specialization.

Effect of Sharing Depth. We vary the number of intermediate shared layers from 0 to 16 and symmetrically allocate the remaining layers to the shallow and deep decoupled stages. The results in Table 3 show that 12 shared layers achieve the best final GenEval (0.63) and MME-S (2006) scores, whereas both fewer and more shared layers produce lower final scores. This result indicates that the extent of parameter sharing is an important architectural factor. Fewer shared layers may limit cross-task collaboration, whereas excessive sharing may weaken task-specific specialization, supporting the 12-layer configuration adopted in Hourglass.

Table 3: GenEval scores at the end of each training stage and MME-S scores across SFT checkpoints with different numbers of intermediate shared layers.

# Shared	GenEval			MME-S			
	PT	CT	SFT	8k	12k	16k	20k
0	0.16	0.48	0.55	**2021**	1954	1965	1976
4	0.16	0.48	0.56	1952	1956	1951	1917
8	0.16	0.45	0.60	1882	1963	1843	1812
12	0.18	**0.53**	**0.63**	2007	**1979**	**1997**	**2006**
16	**0.19**	0.51	0.60	1851	1892	1923	1896

Effect of Subspace Partition. With the same data usage and all hyperparameters fixed, we compare CocoGrad with a variant that uses the same update rule without partitioning subspaces. We evaluate both variants after 10K SFT steps. As shown in Table 4, removing this partition consistently leads to higher training losses and worse benchmark performance.

Table 4: Effect of conflict-aware subspace partitioning. SP: Subspace Partition. $\overline{\text{MSE}}$: Averaged MSE Loss across training. $\overline{\text{CE}}$: Averaged CE Loss across training. $\overline{\text{TL}}$: Averaged Total Loss across training. ↓ indicates lower value is better; ↑ indicates higher value is better.

Method	MME-S↑	GenEval↑	$\overline{\text{MSE}}$↓	$\overline{\text{CE}}$↓	$\overline{\text{TL}}$↓
CocoGrad	**1293**	**0.8440**	**0.3818**	**0.8527**	**0.5950**
w/o SP	1233	0.8400	0.3824	0.8544	0.5960

This suggests that extending the gradient coordination mechanism to all parameter coordinates can disrupt the naturally aligned descent signals in the collaborative subspace and introduce unnecessary optimization noise. Moreover, estimating β over the entire parameter space reduces its effectiveness in coordination. These results support our design choice of confining gradient coordination exclusively to the conflicting subspace, which achieves a better empirical balance between task performance and training stability.

Sensitivity to Conflict-Tolerance Radius. We further evaluate the conflict-tolerance radius c using $c \in \{0.5, 1, 2\}$ for 30K SFT steps, with all other settings fixed. As shown in Table 5, performance improves and plateaus with increasing c, suggesting that an overly small c constrains the feasible gradient-update space. Nevertheless, CocoGrad exhibits mild sensitivity to c, with marginal performance variation on benchmarks in three settings, and consistently outperforms the baseline without CocoGrad, demonstrating the robustness of our method.

Table 5: Performances of different settings of c.

c	MME-S↑	GenEval↑	$\overline{\text{MSE}}$↓	$\overline{\text{CE}}$↓	$\overline{\text{TL}}$↓
0.5	1403	0.8450	**0.3389**	0.8484	0.5510
1	**1419**	**0.8460**	**0.3389**	**0.8321**	**0.5469**
2	**1419**	0.8450	0.3392	0.8325	0.5473

Comparison of Gradient Surgery Methods. Using 30K SFT steps, we compare CocoGrad with other gradient surgery methods, including CAGrad Liu et al. (2021) and PCGrad Yu et al. (2020), as well as a vanilla baseline without any gradient modification. Based on the previous experiment, we set $c = 1$ for CocoGrad and $c = 0.5$ for CAGrad for the best performance. As shown in Table 6, CocoGrad achieves the best benchmark performance among the compared gradient surgery methods, obtaining the highest MME-S score and GenEval score, while also yielding the lowest average CE Loss and Total Loss. This advantage stems from the subspace-aware design and performing coordination on the conflicting subspace. In contrast, PCGrad can be overly aggressive by directly removing conflicting components, whereas CAGrad only seeks a global worst-task compromise, which may reduce their effectiveness. Compared with CAGrad, CocoGrad also incurs lower computational overhead, since it directly computes the update coefficient in closed form rather than relying on an additional solver. These results support both the effectiveness and efficiency of CocoGrad.

Table 6: Performances of different gradient surgery strategies. $\overline{\text{Time}}$: Average function computation time, reported in seconds per step.

Method	MME-S↑	GenEval↑	$\overline{\text{MSE}}$↓	$\overline{\text{CE}}$↓	$\overline{\text{TL}}$↓	$\overline{\text{Time}}$↓
CAGrad	1357	0.8370	0.3388	0.8487	0.5510	0.0831
PCGrad	1378	0.8390	0.3389	0.8491	0.5512	0.0135
Vanilla	1152	0.8140	**0.3386**	0.8516	0.5515	**0.0029**
CocoGrad	**1419**	**0.8460**	0.3389	**0.8321**	**0.5469**	0.0617

CocoGrad or Not. Under the same settings as Table 2, we conduct a full SFT run of Hourglass$_{-\text{coco}}$ and compare it with Hourglass. As shown in Table 7, CocoGrad consistently improves performance across all evaluated benchmarks. These results support the effectiveness of CocoGrad in further improving cross-task coordination on top of the proposed hierarchical sharing architecture.

Table 7: Ablation study of CocoGrad. Higher is better for all metrics. Edit: ImgEdit.

Method	MME-S	MMB	MMMU	GenEval	DPG	Edit
Hourglass	**2135**	**79.6**	**47.4**	**0.870**	**83.30**	**3.47**
Hourglass$_{-\text{coco}}$	2096	79.1	44.3	0.866	82.93	3.40

Scaling Up. Table 2 shows that Hourglass achieves stronger performance when scaled to a larger model size, validating its scalability.

5 Conclusion

We present Hourglass, a unified framework for visual understanding and generation. Its hierarchical decoupled-shared-decoupled architecture promotes cross-task collaboration while preserving task-specific capacity. To stabilize joint training, we propose CocoGrad, a conflict-aware gradient surgery method for shared parameters. Extensive experiments and ablations confirm that Hourglass achieves strong, balanced performance across tasks. We hope Hourglass can serve as a solid step toward unified multimodal modeling.

AI USE STATEMENT

In this work, we used generative AI tools for improving the clarity, grammar, and style of the English writing. These tools were not used to generate research ideas, develop methods, or conduct experiments. All AI-assisted edits were carefully reviewed and verified by the authors. We take full responsibility for the final content of this work.

REFERENCES

Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing, 2025a. URL https://arxiv.org/abs/2412.04280.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025b. URL https://arxiv.org/abs/2511.21631.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, Hao Ai, Siying Bai, Yang Chen, Zhihui Chen, Fengbin Gao, Ying Guo, Dong Li, Zhen Shen, Leilei Shi, Jing Wang, Siyu Wang, Yimeng Wang, Rui Zheng, Ting Yao, and Tao Mei. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer, 2026. URL https://arxiv.org/abs/2605.11061.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021. URL https://arxiv.org/abs/2102.08981.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025a. URL https://arxiv.org/abs/2505.09568.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025b. URL https://arxiv.org/abs/2501.17811.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024b. URL https://arxiv.org/abs/2404.16821.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing, 2024. URL https://arxiv.org/abs/2405.04007.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation, 2025. URL https://arxiv.org/abs/2404.14396.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations, 2025. URL https://arxiv.org/abs/2506.18898.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024. URL https://arxiv.org/abs/2403.05135.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing, 2024. URL https://arxiv.org/abs/2404.09990.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Ming Li, Xin Gu, Fan Chen, Xiaoying Xing, Longyin Wen, Chen Chen, and Sijie Zhu. Superedit: Rectifying and facilitating supervision for instruction-based image editing, 2025. URL https://arxiv.org/abs/2505.02370.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation, 2025. URL https://arxiv.org/abs/2505.05472.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation, 2025. URL https://arxiv.org/abs/2506.03147.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing, 2025. URL https://arxiv.org/abs/2504.17761.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15740–15751, 2026.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Yinyi Luo, Wenwen Wang, Hayes Bai, Hongyu Zhu, Hao Chen, Pan He, Marios Savvides, Sharon Li, and Jindong Wang. Torchumm: A unified multimodal model codebase for evaluation, analysis, and post-training, 2026. URL https://arxiv.org/abs/2604.10784.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing, 2025. URL https://arxiv.org/abs/2510.19808.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. URL https://arxiv.org/abs/2111.02114.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization, 2019. URL https://arxiv.org/abs/1810.04650.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, Guohui Zhang, Wenhu Zhang, Hang Xu, Nan Jiang, Xin Han, Haoze Sun, Maoquan Zhang, Haoyang Huang, and Nan Duan. Joyai-image: Awaking spatial intelligence in unified multimodal understanding and generation, 2026. URL https://arxiv.org/abs/2605.04128.

Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, Chi Zhang, Shifeng Zhang, Ruikai Zhou, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer, 2026. URL https://arxiv.org/abs/2511.22699.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026a.

Qwen Team. Qwen3.5-omni technical report, 2026b. URL https://arxiv.org/abs/2604.15804.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, Zirun Zhu, Ziqian Fan, Leyao Gu, Haomin Wang, Qi Wei, Jinhui Yin, Xue Yang, Zhihang Zhong, Qi Qin, Yi Xin, Bin Fu, Yihao Liu, Jiaye Ge, Qipeng Guo, Gen Luo, Hongsheng Li, Yu Qiao, Kai Chen, and Hongjie Zhang. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing, 2026. URL https://arxiv.org/abs/2603.09877.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a. URL https://arxiv.org/abs/2508.18265.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5m: A million-scale, gpt-generated image dataset, 2025b. URL https://arxiv.org/abs/2507.21033.

Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision, 2025. URL https://arxiv.org/abs/2411.07199.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang,

Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation, 2026. URL https://arxiv.org/abs/2506.18871.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17739–17750. IEEE, 2025b.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

Yan Yang, Haochen Tian, Yang Shi, Wulin Xie, Yi-Fan Zhang, Yuhao Dong, Yibo Hu, Liang Wang, Ran He, Caifeng Shan, et al. A survey of unified multimodal understanding and generation: Advances and challenges. *Authorea Preprints*, 2025.

Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2026.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Erjian Zhang, Yatong Hao, Liejun Wang, and Zhiqing Guo. The double dilemma in multi-task radiology report generation: A gradient dynamics analysis and solution. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a.

Yichen Zhang, Da Peng, Zonghao Guo, Zijian Zhang, Xuesong Yang, Tong Sun, Shichu Sun, Yidan Zhang, Yanghao Li, Haiyan Zhao, Wang Xu, Qi Shi, Yangang Sun, Chi Chen, Shuo Wang, Yukun Yan, Xu Han, Qiang Ma, Wei Ke, Liang Wang, Zhiyuan Liu, and Maosong Sun. Cheers: Decoupling patch details from semantic representations enables unified multimodal comprehension and generation, 2026b. URL https://arxiv.org/abs/2603.12793.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

APPENDIX

A DERIVATION OF THE COCOGRAD COORDINATION COEFFICIENT

Given the conflict-subspace update in Eq. 7 of main text, its alignment with each task gradient characterizes the corresponding first-order reduction in the task loss. Accordingly, the descent gains for understanding and generation are defined as follows:

$$(g_u^C)^\top \delta_C = \|g_u^C\|^2 + (g_u^C)^\top g_g^C + \beta\big(\|g_u^C\|^2 - (g_u^C)^\top g_g^C\big)\,. \tag{11}$$

$$(g_g^C)^\top \delta_C = \|g_g^C\|^2 + (g_u^C)^\top g_g^C + \beta\big((g_u^C)^\top g_g^C - \|g_g^C\|^2\big)\,. \tag{12}$$

Since $(g_u^C)^\top g_g^C < 0$, in the conflicting subspace, the first gain increases with β, whereas the second decreases. Therefore, as shown in Figure 6, the unconstrained max-min solution is attained at their intersection:

$$(g_u^C)^\top \delta_C = (g_g^C)^\top \delta_C. \tag{13}$$

Solving the equality yields

$$\beta_0 = \frac{\|g_g^C\|^2 - \|g_u^C\|^2}{\|g_u^C - g_g^C\|^2 + \epsilon}, \tag{14}$$

where ϵ is a small constant for numerical stability.

The neighborhood constraint in Eq. 8 of main text implies

$$|\beta|\,\|g_u^C - g_g^C\| \le c\|g_u^C + g_g^C\|, \tag{15}$$

and hence $\beta \in [-\rho, \rho]$, where

$$\rho = \frac{c\|g_u^C + g_g^C\|}{\|g_u^C - g_g^C\| + \epsilon}. \tag{16}$$

The constrained solution is obtained by projecting β_0 onto this feasible interval:

$$\beta^* = \mathrm{clip}(\beta_0, -\rho, \rho). \tag{17}$$

B CONVERGENCE ANALYSIS OF COCOGRAD

Definitions. At iteration t, we define the aggregated objective $\mathcal{L}(\theta_t)$, the update direction d_t, its collaborative and conflicting components $d_{t,S}$ and $\delta_{t,C}$, and the gradients g_t, $g_{t,S}$, and $g_{t,C}$, where

$$\begin{aligned} g_{t,S} &= g_{u,t}^S + g_{g,t}^S, \\ g_{t,C} &= g_{u,t}^C + g_{g,t}^C, \\ g_t &= g_{t,S} + g_{t,C}. \end{aligned} \tag{18}$$

Assumptions. We make the following standard assumptions.

- The objective $\mathcal{L}$ is H-smooth, i.e.,

$$\|\nabla\mathcal{L}(x) - \nabla\mathcal{L}(y)\| \le H\|x - y\|, \qquad H > 0. \tag{19}$$

- The objective $\mathcal{L}$ is lower bounded:

$$\mathcal{L}^* = \inf_\theta \mathcal{L}(\theta) > -\infty. \tag{20}$$

- The conflicting-subspace gradient discrepancy is uniformly bounded along the optimization trajectory:

$$\|g_{u,t}^C - g_{g,t}^C\|^2 \le D_{\max}, \qquad \forall t, \tag{21}$$

for some constant $D_{\max} > 0$.

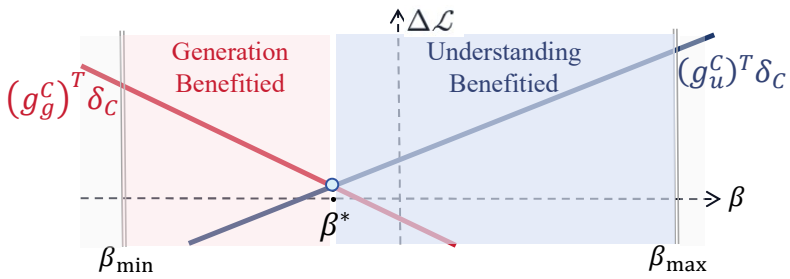

Figure 6: Loss change curves for understanding and generation tasks. The feasible β range is bounded by β_{min} and β_{max}. The optimal $\beta^\star$ lies at the intersection of the two lines, maximizing the minimum task gain.

Lemma 1: Lower Bound for Conflicting-Subspace Descent. Since β_t^* is obtained by clipping $\beta_{0,t}$ to a symmetric interval around zero, there exists $\lambda_t \in [0,1]$ such that

$$\beta_t^* = \lambda_t \beta_{0,t}. \tag{22}$$

For simplicity, define

$$A_t = \|g_{u,t}^C\|^2, \quad B_t = \|g_{g,t}^C\|^2, \quad P_t = (g_{u,t}^C)^\top g_{g,t}^C. \tag{23}$$

We have

$$\begin{aligned} g_{t,C}^\top \delta_{t,C} &= (g_{t,C})^\top \left(g_{u,t}^C + g_{g,t}^C + \beta_t^*(g_{u,t}^C - g_{g,t}^C)\right) \\ &= \|g_{t,C}\|^2 + \beta_t^*(g_{u,t}^C + g_{g,t}^C)^\top (g_{u,t}^C - g_{g,t}^C) \\ &= \|g_{t,C}\|^2 + \beta_t^*(A_t - B_t) \\ &= \|g_{t,C}\|^2 - \lambda_t \frac{(A_t - B_t)^2}{\|g_{u,t}^C - g_{g,t}^C\|^2 + \epsilon}, \end{aligned} \tag{24}$$

Since $\lambda_t \le 1$, it follows that

$$g_{t,C}^\top \delta_{t,C} \ge \|g_{t,C}\|^2 - \frac{(A_t - B_t)^2}{\|g_{u,t}^C - g_{g,t}^C\|^2 + \epsilon} \tag{25}$$

Using

$$\|g_{t,C}\|^2 \|g_{u,t}^C - g_{g,t}^C\|^2 - (A_t - B_t)^2 = 4(A_t B_t - P_t^2), \tag{26}$$

we obtain

$$g_{t,C}^\top \delta_{t,C} \ge \frac{4(A_t B_t - P_t^2) + \epsilon \|g_{t,C}\|^2}{\|g_{u,t}^C - g_{g,t}^C\|^2 + \epsilon}. \tag{27}$$

By *Cauchy–Schwarz*, $A_t B_t - P_t^2 \ge 0$. Therefore,

$$\begin{aligned} g_{t,C}^\top \delta_{t,C} &\ge \frac{\epsilon}{\|g_{u,t}^C - g_{g,t}^C\|^2 + \epsilon} \|g_{t,C}\|^2 \\ &\ge \mu \|g_{t,C}\|^2, \end{aligned} \tag{28}$$

where

$$\mu = \frac{\epsilon}{D_{\max} + \epsilon} > 0. \tag{29}$$

Lemma 2: Upper Bound for Update Direction Norm. By the clipping operation from Eq. 17,

$$|\beta_t^*| \le \rho_t = \frac{c\|g_{u,t}^C + g_{g,t}^C\|}{\|g_{u,t}^C - g_{g,t}^C\| + \epsilon}. \tag{30}$$

Hence,

$$\begin{aligned} \|\beta_t^*(g_{u,t}^C - g_{g,t}^C)\| &\le \frac{c\|g_{u,t}^C + g_{g,t}^C\|}{\|g_{u,t}^C - g_{g,t}^C\| + \epsilon} \|g_{u,t}^C - g_{g,t}^C\| \\ &\le c\|g_{u,t}^C + g_{g,t}^C\|. \end{aligned} \tag{31}$$

Therefore,

$$\begin{aligned}\|\delta_{t,C}\| &= \|g_{u,t}^C + g_{g,t}^C + \beta_t^*(g_{u,t}^C - g_{g,t}^C)\| \\ &\le \|g_{u,t}^C + g_{g,t}^C\| + \|\beta_t^*(g_{u,t}^C - g_{g,t}^C)\| \\ &\le (1+c)\|g_{u,t}^C + g_{g,t}^C\|.\end{aligned} \tag{32}$$

Since the collaborative and conflicting subspaces are orthogonal, the norm of the final update is

$$\begin{aligned}\|d_t\|^2 &= \|d_{t,S}\|^2 + \|\delta_{t,C}\|^2 \\ &\le \|d_{t,S}\|^2 + (1+c)^2\|g_{t,C}\|^2,\end{aligned} \tag{33}$$

Lemma 3: Single-Step Descent. By the smoothness of $\mathcal{L}$, we have

$$\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t) - \alpha g_t^\top d_t + \frac{H\alpha^2}{2}\|d_t\|^2. \tag{34}$$

Using the orthogonal decomposition,

$$g_t^\top d_t = g_{t,S}^\top d_{t,S} + g_{t,C}^\top \delta_{t,C}. \tag{35}$$

Since $d_{t,S} = g_{t,S}$ and $g_{t,C}^\top \delta_{t,C} \ge \mu\|g_{t,C}\|^2$, we obtain

$$g_t^\top d_t \ge \|g_{t,S}\|^2 + \mu\|g_{t,C}\|^2. \tag{36}$$

Combining Eq. 33, Eq. 34 and Eq. 36, we have

$$\begin{aligned}\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) &\le -\alpha\left(\|g_{t,S}\|^2 + \mu\|g_{t,C}\|^2\right) \\ &\quad + \frac{H\alpha^2}{2}\left(\|g_{t,S}\|^2 + (1+c)^2\|g_{t,C}\|^2\right) \\ &= -\alpha\left(1 - \frac{H\alpha}{2}\right)\|g_{t,S}\|^2 \\ &\quad - \alpha\left(\mu - \frac{H\alpha(1+c)^2}{2}\right)\|g_{t,C}\|^2.\end{aligned} \tag{37}$$

For a sufficiently small step size α satisfying

$$0 < \alpha \le \min\left\{\frac{1}{H}, \frac{\mu}{H(1+c)^2}\right\}, \tag{38}$$

we have

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \le -\frac{\alpha}{2}\|g_{t,S}\|^2 - \frac{\alpha\mu}{2}\|g_{t,C}\|^2. \tag{39}$$

Let

$$\kappa = \frac{1}{2}\min\{1, \mu\}. \tag{40}$$

Since

$$\|g_t\|^2 = \|g_{t,S}\|^2 + \|g_{t,C}\|^2, \tag{41}$$

we obtain

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \le -\alpha\kappa\|g_t\|^2. \tag{42}$$

Theorem 1: Convergence and Rate. Summing the Eq. 42 from $t = 0$ to $T - 1$, we have

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta_0) \le -\alpha\kappa\sum_{t=0}^{T-1}\|g_t\|^2. \tag{43}$$

Since $\mathcal{L}(\theta_T) \ge \mathcal{L}^*$, it follows that

$$\alpha\kappa\sum_{t=0}^{T-1}\|g_t\|^2 \le \mathcal{L}(\theta_0) - \mathcal{L}^*. \tag{44}$$

Table 8: Open-source datasets used in our training pipeline.

Data Type	Datasets
Image generation	BLIP-3o Chen et al. (2025a), CC12M Changpinyo et al. (2021), LAION Schuhmann et al. (2021), cheers-SFT Zhang et al. (2026b)
Image editing	ImgEdit Ye et al. (2026), OmniEdit Wei et al. (2025), HumanEdit Bai et al. (2025a), SuperEdit Li et al. (2025), HQEdit Hui et al. (2024), InstructPix2Pix Brooks et al. (2023), UltraEdit Zhao et al. (2024), GPT-Image-Edit Wang et al. (2025b), SEED-Data-Edit Ge et al. (2024), AnyEdit Yu et al. (2025), Pico Banana Qian et al. (2025), UniWorld-V1 Lin et al. (2025)
Visual understanding	cheers-SFT Zhang et al. (2026b), LLaVa-OneVision-Data Li et al. (2024), ShareGPT4V Chen et al. (2024a), TextVQA Singh et al. (2019), OCR-VQA Mishra et al. (2019), GQA Hudson & Manning (2019), VSR-random Liu et al. (2023a), ScienceQA Lu et al. (2022)

Therefore,

$$\sum_{t=0}^{T-1} \|g_t\|^2 \leq \frac{\mathcal{L}(\theta_0) - \mathcal{L}^*}{\alpha\kappa}. \tag{45}$$

Taking $T \to \infty$, we obtain

$$\sum_{t=0}^{\infty} \|g_t\|^2 < \infty. \tag{46}$$

Since each term is nonnegative, it follows that

$$\lim_{t\to\infty} \|g_t\| = 0. \tag{47}$$

That is, CocoGrad converges to a first-order stationary point of the aggregated objective $\mathcal{L}$. Moreover, from Eq. 45, we have

$$\min_{0\leq t<T} \|\nabla\mathcal{L}(\theta_t)\|^2 \leq \frac{\mathcal{L}(\theta_0) - \mathcal{L}^*}{\alpha\kappa T}. \tag{48}$$

Thus, CocoGrad preserves the standard nonconvex convergence rate

$$\min_{0\leq t<T} \|\nabla\mathcal{L}(\theta_t)\|^2 = O\left(\frac{1}{T}\right). \tag{49}$$

C Training Data and Implementation Details

This section provides additional details on the data composition and implementation of the three-stage training pipeline. The optimization settings for each stage are reported in the main paper. The open-source datasets used for image generation, image editing, and visual understanding are summarized in Table 8.

Implementation Details. The training was conducted on 80 H100 GPUs, each with 80 GB of GPU memory.

Pretraining (PT). The pretraining stage uses approximately 37M image-text pairs and 8M image editing open-source samples. The image generation data are mainly collected from BLIP-3o and CC12M, while the image editing data are drawn from multiple sources, such as ImgEdit, OmniEdit, and HumanEdit. This stage aims to establish the initial generation capability through large-scale and heterogeneous generation corpora.

Continued Pretraining (CT). In the continued pretraining stage, we use 6M image generation samples and 1M image editing samples. We perform large-scale aesthetic annotation over the training corpus and select high-quality subsets from both the pretraining data and the LAION data. For the LAION subset, we further apply Qwen3.5 Team (2026a) to produce recaptioned descriptions at different levels of granularity, enabling the model to better follow both short and long-form generation

Table 9: Evaluation of text-to-image generation ability on GenEval and DPG-Bench. Among unified models, bold denotes the best performance, and underlined denotes the second-best performance.

Model	GenEval							DPG-Bench					
	Single obj.	Two obj.	Counting	Colors	Position	Color attr.	Overall	Global	Entity	Attribute	Relation	Other	Overall
Generation-Only Models													
SD3-Medium-2B	0.99	0.94	0.72	0.89	0.33	0.60	0.74	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev-12B	0.98	0.93	0.75	0.93	0.68	0.65	0.82	82.10	89.50	88.80	91.10	89.40	84.00
GPT-Image-1 [High]	0.99	0.92	0.85	0.92	0.75	0.61	0.84	88.89	88.94	89.84	92.63	90.96	85.15
Qwen-Image-20B	0.99	0.92	0.89	0.88	0.76	0.77	0.87	91.32	91.56	92.02	94.31	92.73	88.32
Z-Image-6B	1.00	0.94	0.78	0.93	0.62	0.77	0.84	93.39	91.22	93.16	92.22	91.52	88.14
Unified Multimodal Models													
SEED-X-13B	0.96	0.65	0.31	0.80	0.18	0.14	0.51	–	–	–	–	–	–
Emu3-8B	0.99	0.81	0.42	0.80	0.49	0.45	0.66	–	–	–	–	–	80.60
Janus-Pro-7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80	86.90	88.90	89.40	89.32	89.48	84.19
BAGEL-14B	0.99	0.94	0.81	0.88	0.64	0.63	0.82	88.94	90.37	**91.29**	90.82	88.67	85.07
Show-o2-7B	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76	89.00	**91.78**	89.96	91.81	**91.64**	**86.14**
UniWorld-V1-20B	0.99	0.93	0.79	0.89	0.49	0.70	0.80	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2-7B	**1.00**	0.95	0.64	0.88	0.55	0.76	0.80	88.81	88.83	90.18	89.37	90.27	83.57
Tar-7B	0.99	0.92	**0.83**	0.85	0.80	0.65	0.84	83.98	88.62	88.05	**93.98**	84.86	84.19
Cheers-1.5B	0.98	0.92	0.65	0.86	0.63	0.65	0.78	**90.84**	90.24	89.42	89.71	87.37	83.48
InternVL-U-3.7B	0.99	0.94	0.74	0.91	0.77	0.74	0.85	90.39	90.78	90.68	90.29	88.77	85.18
Hourglass-3B	**1.00**	0.93	0.73	**0.93**	**0.85**	0.78	0.87	85.31	88.44	88.96	92.58	86.47	83.30
Hourglass-7.5B	**1.00**	**0.96**	0.78	**0.93**	0.84	**0.83**	**0.89**	85.13	91.39	90.28	91.44	89.46	84.23

Table 10: Evaluation of image editing ability on ImgEdit. Among unified models, bold denotes the best performance, and underlined denotes the second-best performance.

Model	ImgEdit									
	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall
Generation-Only Models										
AnyEdit Yu et al. (2025)	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
Step1X-Edit Liu et al. (2025)	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
GPT-Image-1 [High]	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
Qwen-Image-Edit-20B	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
Z-Image-Edit-6B	4.40	4.14	4.30	4.57	4.13	4.14	4.85	3.63	4.50	4.30
Unified Multimodal Models										
BAGEL-14B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1-20B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2-7B	3.57	3.06	1.77	3.74	3.20	3.57	**4.81**	2.52	**4.68**	3.44
InternVL-U-3.7B	**4.13**	3.40	2.27	**4.13**	3.39	**3.84**	4.77	**3.03**	4.05	**3.67**
Hourglass-3B	3.75	3.76	**3.28**	3.01	3.34	3.18	4.50	1.68	3.13	3.47
Hourglass-7.5B	3.99	**4.00**	2.62	3.28	**3.61**	3.19	4.66	1.94	3.55	3.55

instructions. In addition, we introduce a small amount of high-quality synthetic data generated by HiDream-O1-Image Cai et al. (2026) and Qwen-Image-Edit Wu et al. (2025a) to further enhance the model's understanding of composition and its compositional reasoning ability.

Supervised Fine-Tuning (SFT). The supervised fine-tuning mixture consists of 9M visual understanding, 5M text-to-image generation, and 1M image-editing data. Compared with the previous two stages, the SFT stage applies stricter filtering criteria to select higher-quality supervised data for image generation and editing. We additionally incorporate the SFT data released by Cheers into the image-generation and visual-understanding subsets. Moreover, we curate several high-quality visual question answering datasets to improve the model's fine-grained visual perception, question comprehension, and multimodal reasoning capabilities.

D Detailed Evaluation Results

Text-to-Image Generation. Table 9 provides a more detailed breakdown of the GenEval and DPG-Bench results, respectively. On GenEval, both Hourglass-3B and Hourglass-7.5B demonstrate superior compositional generation capabilities. Hourglass-7.5B achieves a SOTA overall score of 0.89, while Hourglass-3B achieves the second-highest score of 0.87. These results outperform substantially larger generation-only and unified models, including Qwen-Image-20B (0.87) and BAGEL-14B (0.82). More specifically, both Hourglass variants rank first or second among unified models in the Colors, Position, and Color Attribute sub-tasks. These tasks require accurate recognition of visual attributes and spatial relationships, demonstrating Hourglass's strong ability to capture fine-

grained semantic information and compositional dependencies. On DPG-Bench, Hourglass-7.5B (84.23) attains competitive performance and surpasses most approaches, such as SEED-X (80.60), OmniGen2 (83.57) and Uniworld-V1 (81.38). Hourglass exhibits superior semantic alignment with dense prompts involving multiple objects.

Image Editing. As shown in Table 10, Hourglass-7.5B achieves an overall score of 3.55 on ImgEdit, ranking second among the compared unified models and significantly outperforming some specialized editing models like AnyEdit (2.45) and Step1X-Edit (3.06). Meanwhile, the Hourglass variants achieve leading results on several fundamental editing operations. Hourglass-7.5B ranks first on Adjust and Remove and second on Add, while Hourglass-3B and Hourglass-7.5B rank first and second, respectively, on Extract. The strong performance across these sub-tasks demonstrates the ability of Hourglass to accurately follow diverse editing instructions and perform fine-grained modifications to objects and their attributes.

E Detailed Ablation Experimental Configurations.

To balance computational cost with controlled comparisons, we adopt different experimental configurations for different groups of ablation experiments. Within each ablation group, all variants use identical initialization, training data, optimization settings, and computational resources, except for the factor under investigation.

Sharing placement and depth. The ablation studies on the effect of sharing placement and sharing depth adopt a shortened version of the three-stage training pipeline. Specifically, each model is trained for 82K steps during pre-training, 40K steps during continued pretraining, and 20K steps during supervised fine-tuning. All variants in these two experiments are trained using 16 H100 GPUs.

CocoGrad component analysis. The ablation studies on the effect of subspace partition, sensitivity to the conflict-tolerance radius, and the comparison of gradient surgery methods are initialized from the same checkpoint obtained after completing the full CT stage specified in Table 1 of the main text. Starting from this checkpoint, the subspace partitioning experiment performs 10K SFT steps, while the conflict-tolerance radius and gradient surgery experiments perform 30K SFT steps. All these experiments are conducted using 8 H100 GPUs.

Full-pipeline experiments. The *CocoGrad or Not* and *Scaling Up* experiments follow exactly the same complete training pipeline as the main experiments in Table 1 of the main text, including the training stages, numbers of training steps, and computational resources. Each model is trained using 80 H100 GPUs.

F More Visual Examples

This section presents additional qualitative results of Hourglass across multiple tasks. Figure 7 shows text-to-image generation results covering diverse prompt types. Figure 8 presents image editing results under various instructions, including object addition, removal, replacement, and style transfer. These visualizations highlight Hourglass's strong instruction-following and visual generation capabilities.



Figure 7: Qualitative examples of Hourglass for text-to-image generation.

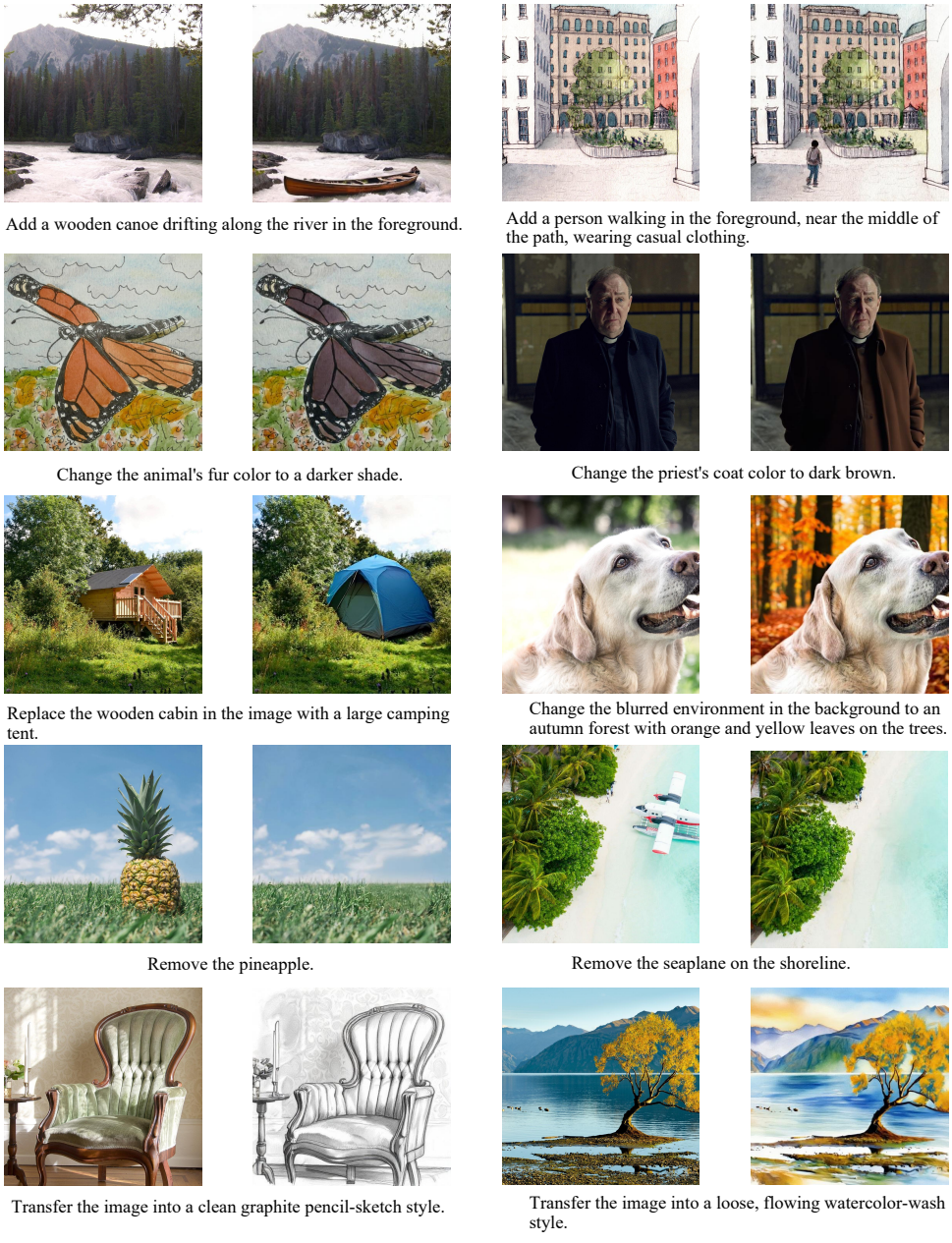

Figure 8: Qualitative examples of Hourglass for instruction-based image editing.