

# SYNERGIZING UNDERSTANDING AND GENERATION WITH INTERLEAVED ANALYZING-DRAFTING THINKING

006 **Anonymous authors**

007 Paper under double-blind review

## ABSTRACT

013 Unified Vision-Language Models (UVLMs) aim to advance multimodal learning by supporting both understanding and generation within a single framework. 014 However, existing approaches largely focus on architectural unification while 015 overlooking the need for explicit interaction between the two capabilities during 016 task solving. As a result, current models treat understanding and generation as 017 parallel skills rather than synergistic processes. To achieve real *synergy*, we intro- 018 duce the interleaved Analyzing–Drafting problem-solving loop (**AD-Loop**), a 019 new think paradigm that dynamically alternates between analytic and drafting op- 020 erations. By interleaving textual thoughts with visual thoughts, AD-Loop enables 021 models to iteratively refine both comprehension and outputs, fostering genuine 022 synergy. To train this mechanism, we design a two-stage strategy: supervised 023 learning on interleaved thought data to initialize alternation, followed by rein- 024 forcement learning to promote adaptive and autonomous control. Extensive ex- 025 periments demonstrate that AD-Loop consistently improves performance across 026 standard benchmarks for both understanding and generation, with strong trans- 027 ferability to various UVLMs architectures. Visual analyses further validate the effec- 028 tiveness of implicit visual thoughts. These results highlight AD-Loop as a prin- 029 cipled and broadly applicable strategy for synergizing comprehension and creation. 030 Code and model will be available later.

## 1 INTRODUCTION

032 Unified Vision–Language Models (Wang et al., 2024; Wu et al., 2025b; Deng et al., 2025; Wu 033 et al., 2025a; Xie et al., 2025) can handle both understanding and generation tasks, attracting 034 significant research attention as they hold the potential to move beyond task-specific solutions 035 toward general multimodal intelligence. Recent advances in multimodal large language models 036 (MLLMs) (Liu et al., 2024a; Bai et al., 2023; Dai et al., 2023) and diffusion-based generative trans- 037 formers (DiTs) (Podell et al., 2023; Yang et al., 2024; AI, 2025) have substantially improved 038 performance in both multimodal comprehension and content creation. Building upon these advances, 039 early efforts (Shen et al., 2023; Wu et al., 2024b) have attempted to integrate existing understand- 040 ing and generation models within a single framework, enabling both capabilities simultaneously. 041 However, such straightforward integration merely co-locates the two paradigms without enabling 042 genuine interaction or mutual reinforcement. In essence, understanding and generation are inher- 043 ently complementary (Paivio et al., 2006; Ellamil et al., 2012). Robust understanding provides the 044 semantic foundation for faithful generation, while successful generation results can serve as tangible 045 evidence of visual comprehension. Therefore, an effective unified model should not only combine 046 the two capabilities but also establish a mutually reinforcing loop.

047 To realize such synergy, early works have sought to build coherent unified architectures, for instance 048 by casting both understanding and generation as autoregressive (AR) next-token prediction (Zhan 049 et al., 2024; Wu et al., 2025a; Wang et al., 2024; Team, 2024a). While this design provides a unified 050 interface from an engineering standpoint, it often suffers from issues such as *parameter competition* 051 and *task interference*. More recent efforts explored (i) decoupled encoders with multi-head outputs 052 to reduce representational conflicts (Chen et al., 2025c; Ma et al., 2025; Kou et al., 2024; Deng et al., 053 2025), and (ii) hybrid AR–diffusion models (Xie et al., 2025; Zhou et al., 2024; Xiao et al., 2025b)

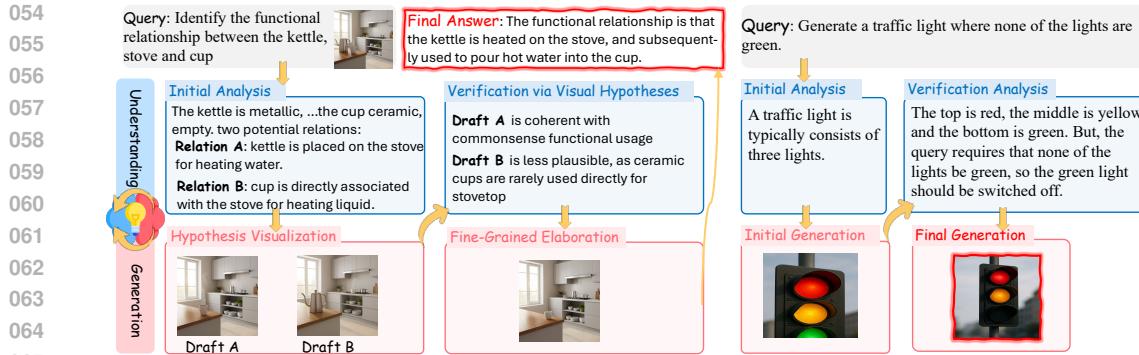


Figure 1: Illustration of the interleaved analytic–drafting problem-solving loop, where understanding and generation interact synergistically to yield accurate solutions.

that take advantage of efficiency with high-fidelity rendering. Overall, these explorations of synergy have predominantly focused on improving architectural design to unify the two abilities, while overlooking a crucial point: during task solving, the understanding and generation modules often lack *close and explicit interaction*, thereby failing to realize genuine synergy between comprehension and generation. To illustrate this, when a user’s instruction is ambiguous, an understanding module could first propose several plausible candidate solutions for a question, then invoke generation to produce sketches or key visualizations as a means of “verifying” these candidates, ultimately yielding the correct answer. Conversely, once the generation module produces initial results, it could query the understanding module for high-level guidance, such as attributes or reasonable spatial layouts, to progressively refine the output (see Fig. 1). This motivates a new perspective: *instead of treating understanding and generation as co-existing skills, we argue they should be interleaved in a problem-solving loop*.

To combat this, we propose a novel thinking paradigm, termed the interleaved *analyzing–drafting problem-solving loop (AD-Loop)*. The core idea is to enable the model to dynamically alternate between understanding and generation, thereby fostering an organic synergy that enhances overall problem-solving capability. Specifically, building upon existing UVLMs, we design to interleave textual thoughts, such as semantic abstraction and reasoning, with visual thoughts, including sketching and spatial layout. This dynamic switching between analytic and synthetic modes enables the model to iteratively refine its reasoning and outputs. Furthermore, we introduce a two-stage training strategy to effectively optimize the model, as shown in Fig. 2. **In the first stage**, we construct supervised datasets of interleaved textual–visual thoughts, which initialize the model with the ability to alternate between analyzing and drafting in a guided manner. **In the second stage**, we employ reinforcement learning with hybrid feedback to encourage the model to intelligently and autonomously decide when to invoke understanding versus generation, thereby fostering adaptive and self-directed synergy. Importantly, our proposed training framework and problem-solving paradigm are architecture-agnostic, making them broadly applicable to a wide range of existing UVLMs. By integrating interleaved analyzing and drafting, our approach substantially enhances their ability to achieve deep synergy between comprehension and creation.

We conduct extensive experiments to evaluate the effectiveness of the proposed AD-Loop. First, across widely used understanding and generation benchmarks, our method consistently delivers broad and significant performance improvements, including an average +2.3% improvement on understanding tasks and an overall score of 86% on GenEval. Next, ablations across thinking strategies demonstrate that interleaving analytic and drafting thinking yields clear advantages. We then conduct transfer studies across different UVLM architectures, showing that our approach can be seamlessly applied to diverse models and significantly enhance both understanding and generation benchmarks compared to the original baselines. Furthermore, visualizations of implicit visual thoughts confirm the rationality of our design. Finally, through detailed case analyses, we present compelling evidence that our model alternates between understanding and generation during problem-solving, leading to superior outcomes. Our contributions can be summarized as follows:

- We propose a fundamentally new strategy for synergizing understanding and generation by introducing the interleaved analyzing–drafting thinking mechanism, which enables tight, explicit interactions between the two capabilities.

- 108 • We develop a novel two-stage learning strategy that equips UVLMs with the ability to intel-  
109 ligently adopt interleaved analytic–drafting problem-solving loops during the task-solving  
110 process. Moreover, the framework is architecture-agnostic across different UVLMs.
- 111 • Extensive experiments on understanding and generation benchmarks demonstrate the ef-  
112 fectiveness of our approach, providing intuitive visual analyses that confirm both the ratio-  
113 nality and efficacy of the proposed synergy.

## 115 2 RELATED WORK

116 The mutual benefits between understanding and generation have long been a central theme Paivio  
117 et al. (2006); Ellamil et al. (2012). Enabled by rapid progress in multimodal large language models  
118 (MLLMs) Liu et al. (2024a); Bai et al. (2023); Li et al. (2024) and diffusion-based generators Podell  
119 et al. (2023); Yang et al. (2024), state-of-the-art systems now achieve strong performance on each  
120 task in isolation. This has sparked growing interest in unified vision–language models (UVLMs) Xie  
121 et al. (2025); Kou et al. (2024); Wu et al. (2025b), which support multimodal understanding and  
122 generation within a single framework. Early explorations (Shen et al., 2023; Wu et al., 2024b)  
123 largely composed powerful specialist models by directly connecting an understanding model with a  
124 generation model. However, such plug-and-play integration merely co-locates the two capabilities  
125 and does not realize mutual assistance. More recent efforts pursue more coherent formulations  
126 that jointly model both tasks, e.g., casting both as autoregressive next-token prediction (Wu et al.,  
127 2025a; Team, 2024a; Chen et al., 2025c;b; Geng et al., 2025b) or adopting unified AR-diffusion  
128 architectures (Xie et al., 2025; Zhou et al., 2024; Deng et al., 2025). Yet even these tighter designs  
129 still treat understanding and generation as parallel, independently callable skills; at inference time,  
130 the two modules rarely engage in close, explicit interaction, and genuine synergy remains elusive.  
131 In contrast, we introduce a novel thinking paradigm that enables true synergic learning between  
132 understanding and generation through an interleaved analyzing–drafting loop.

133 A complementary line of work (Shao et al., 2024; Xu et al., 2024; Shen et al., 2025) focuses on end-  
134 owing MLLMs with strong reasoning capabilities. Early approaches follow the Chain-of-Thought  
135 paradigm (Wei et al., 2022) by decomposing problems into substeps and supervising stepwise ratio-  
136 nales. Subsequent methods scale test-time computation (Snell et al., 2024) via self-consistency or  
137 majority voting and incorporate search-based inference (Yao et al., 2024), while training-time credit  
138 assignment is improved through reinforcement learning (RL) or preference optimization tailored to  
139 reasoning. More recently, RL has been explored to elicit emergent reasoning abilities that improve  
140 multimodal understanding (Huang et al., 2025; Chen et al., 2025a; Yuan et al., 2025) and genera-  
141 tion (Xiao et al., 2025b; Jiang et al., 2025a). In light of this, some explorations (Jiang et al., 2025b;  
142 Yan et al., 2025) attempt to leverage RL to co-optimize understanding and generation, thereby  
143 strengthening their respective capabilities. Beyond textual-only reasoning thoughts, several stud-  
144 ies (Cheng et al., 2025a; Shao et al., 2024) interleave textual reasoning with visual evidence, either  
145 via tool-augmented pipelines (Zheng et al., 2025; Zhang et al., 2025a; Man et al., 2025; Su et al.,  
146 2025) or by directly generating visual traces (Chern et al., 2024; Li et al., 2025; Yang et al., 2025b).  
147 Building on these advances, our work models the synergy between understanding and generation  
148 as an interleaved analyzing–drafting problem-solving loop that jointly produces textual and visual  
149 thoughts, thereby enhancing synergic learning within UVLMs.

## 150 3 METHODOLOGY

151 Building upon a UVLM, we model the synergistic *understanding* and *generation* thinking process  
152 as an **interleaved analyzing–drafting problem-solving loop**. Given an input, the model alternates  
153 between analyzing (producing *textual thoughts*) and drafting (producing *visual thoughts*) before de-  
154 livering the final output. To achieve this, we design a two-stage training pipeline: **Stage 1** performs  
155 supervised training to *imitate interleaved thinking*, and **Stage 2** leverages reinforcement learning to  
156 enable the model to *adaptively decide when to invoke* analysis or drafting.

### 157 3.1 MODEL ARCHITECTURE

158 Let the input be  $x = (q, \mathcal{I})$ , where  $q$  is text and  $\mathcal{I} = \{I_m\}_{m=1}^M (M \geq 1)$  is an optional set of images.  
159 The input image set  $\mathcal{I} = \{I_m\}$  is encoded by a vision encoder into visual embeddings, which are  
160 fused with the textual stream and then processed by an LLM backbone for reasoning. The model

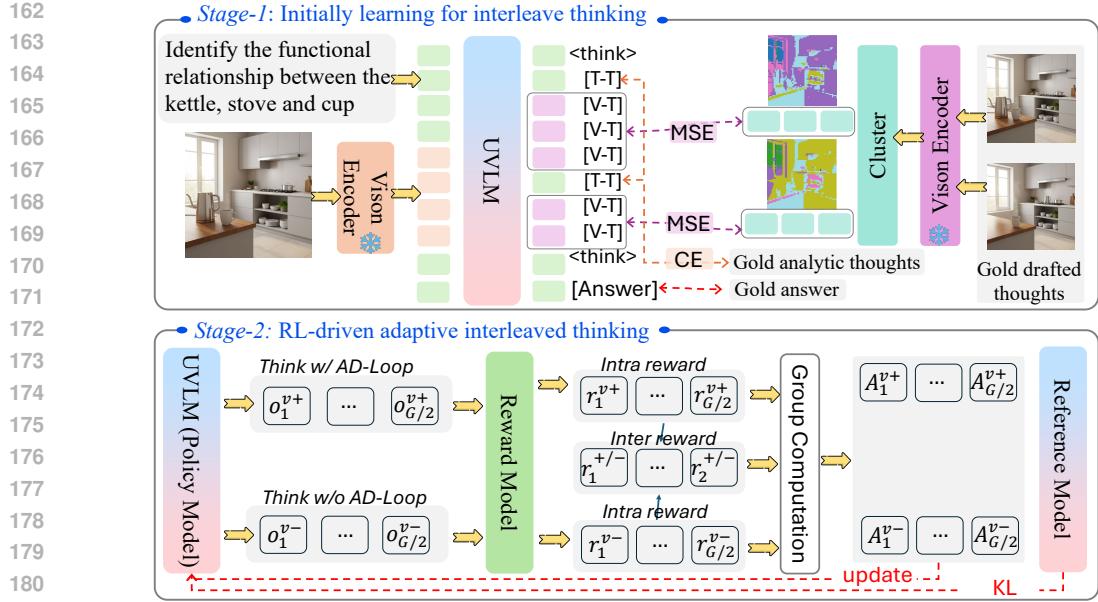


Figure 2: Pipeline of our training framework. **Stage-1**: train the UVLM to emit interleaved thinking traces. **Stage-2**: apply GRPO for hybrid reasoning. The policy samples multiple traces with/without the interleaved AD-Loop for each input. A reward model scores outcomes, and then group-normalized advantages are applied to update the policy, teaching the model when AD-Loop helps.

outputs a thinking trace wrapped in `<think>` and `</think>` tags and a final outcome:

$$<\text{think}> [\text{T-T}] [\text{V-T}] [\text{T-T}] [\text{V-T}] \dots </\text{think}> [\text{Answer}], \quad (1)$$

where `[T-T]` denotes *text thought*, `[V-T]` denotes *visual thought*, which is enclosed by two special tokens marking the beginning and end of the visual thought. `[Answer]` is the final text or image. For image synthesis, prior unified models typically adopt either (i) **discrete** codebooks that predict quantized “image tokens” (Chen et al., 2025c; Xie et al., 2025), or (ii) **continuous** latent regressors that predict low-dimensional vectors consumed by a renderer (Deng et al., 2025). Emitting a *full* visual scene during the thought phase requires long token streams (often hundreds of tokens or prolonged diffusion steps), which increases latency and entangles reasoning with pixel-level details that are irrelevant to the decision. Motivated by cognitive accounts that internal imagery is schematic rather than pixel-accurate (Shepard & Metzler, 1971), we replace full rendering during thinking with a compact set of *latent visual thoughts*,  $\{v_j\}_{j=1}^K$ , which summarizes only the factors *useful for reasoning* under a strict budget  $K \ll$  the tokens required to render an image. This design preserves the sufficiency of visual cues for the downstream decision while avoiding the cost of pixel-level synthesis in the thought process.

### 3.2 STAGE-1: SUPERVISED TRAINING OF INTERLEAVED THINKING

We begin with supervised fine-tuning on interleaved reasoning data. This approach mitigates the instability commonly observed in cold-start RL for reasoning and provides a strong initialization for next-stage reinforcement learning.

**Dataset Construction.** Our interleaved corpus comprises two halves: *i) Understanding*. We leverage the existing multimodal CoT resources (Cheng et al., 2025b; Zhang et al., 2025b; Shao et al., 2024), reorganizing their rationales and cropping the referenced regions according to the provided annotations to construct the interleaved schema. We further synthesize AD-Loop thinking traces by instantiating schematic grid tasks following Li et al. (2025). In total, the understanding portion contains 20K interleaved examples. *ii) Generation*. We use the GoT dataset (Fang et al., 2025) and additionally construct an Inter-T2I set from X-to-image corpora (Xiao et al., 2025a), augmenting each instance with an interleaved thinking trace. This yields 22K interleaved generation traces.

**Training.** To teach the model an AD-Loop style of thinking, we fine-tune the UVLM directly on the above collected corpus, as shown in Fig. 2. A practical issue is that the data sources provide

216 explicit visual thoughts (i.e., pixel images), whereas our reasoning format requires latent visual  
 217 thoughts. We therefore convert each explicit visual-thought image  $I$  into a compact set of tokens  
 218 via a *frozen* generator-side encoder and a deterministic clustering step. Technically, we reuse the  
 219 encoder for image generation from the unified model. Given  $I$ , the encoder yields a grid of latent  
 220 tokens  $\{z_i\}_{i=1}^N$ . Rather than supervising the model to emit the full grid, we reduce  $\{z_i\}_{i=1}^N$  to a  
 221 small, semantically coherent set  $\{v_j\}_{j=1}^K$  with  $K \ll N$ . Following Jin et al. (2024), we adopt a  
 222 density peaks clustering mechanism to construct the token clusters. For each cluster  $\mathcal{C}_j$ , we produce  
 223 a representative latent visual thought token  $v_j$  as the union of members  $v_j = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} z_i$ .  
 224 The clustering mechanism yields stable targets that merge tokens based on semantic proximity in  
 225 the latent space, rather than through naive spatial pooling. We order  $\hat{\mathcal{V}} = \{v_j\}_{j=1}^K$  by the center  
 226 coordinates of their clusters (top-left to bottom-right) to obtain a deterministic sequence. Let  $\mathcal{T}^*$   
 227 and  $\mathcal{V}^*$  denote the gold text-thought and visual thoughts sequences in the `<think>` block, and let  
 228  $o^*$  denote the final output. The final training objective is to optimize:

$$\mathcal{L}_{\text{SI}} = \underbrace{\mathcal{L}_{\text{CE}}(\hat{\mathcal{T}}, \mathcal{T}^*)}_{\text{text thoughts}} + \alpha \underbrace{\mathcal{L}_{\text{vis}}(\hat{\mathcal{V}}, \mathcal{V}^*)}_{\text{latent visual thoughts}} + \underbrace{\mathcal{L}_{\text{out}}(\hat{o}, o^*)}_{\text{task output}}, \quad (2)$$

232 where  $\mathcal{L}_{\text{vis}}$  is mean-squared error, and  $\mathcal{L}_{\text{out}}$  is the original task loss.  $\alpha$  is the coefficient weight.

### 234 3.3 STAGE-2: RL-DRIVEN ADAPTIVE INTERLEAVED THINKING

235 After Stage-1 fine-tuning, the model has acquired an initial interleaved thinking capability. How-  
 236 ever, some queries might be solved confidently with an isolated thinking mechanism, i.e., invok-  
 237 ing only the understanding or generation capability. Accordingly, Stage-2 further strengthens the  
 238 model’s thinking competence and makes the policy *adaptive* for each input. The model should be  
 239 able to decide whether to think with AD-Loop ( $\mathcal{V}^+$ ) or without it ( $\mathcal{V}^-$ ).

240 Drawing on group-relative preference optimization (Guo et al., 2025; Jiang et al., 2025c), we intro-  
 241 duce an *adaptive interleaved thinking* regimen, as illustrated in Fig. 2. For every query, the policy  
 242 explores both thinking modes and receives inter- and intra-group normalized feedback. The resulting  
 243 group-relative advantages drive on-policy updates toward a parsimonious strategy. We next detail  
 244 the sampling scheme in *Adaptive Interleaved Thinking*, the reward design in *Reward Assignment*,  
 245 and the update rule in *Policy Optimization*.

246 **Adaptive Interleaved Thinking.** For each query  $q$ , we sample two *groups* of trajectories from  
 247 the old policy  $\pi_{\text{old}}$ : one with the AD-Loop enabled ( $\{o_i^+\}_{i=1}^{G/2}$ ), and one with the AD-Loop disabled  
 248 ( $\{o_i^-\}_{i=1}^{G/2}$ ). Let  $G$  be the total number of samples per query:

$$\{o_i^+\}_{i=1}^{G/2} \sim \pi_{\text{old}}, \quad \{o_i^-\}_{i=1}^{G/2} \sim \pi_{\text{old}}, \quad O = \{o_i^+\}_{i=1}^{G/2} \cup \{o_i^-\}_{i=1}^{G/2} \quad (3)$$

252 **Reward Assignment.** Following Guo et al. (2025), each trajectory is assigned a scalar reward  
 253 composed of a *format* component and a *content* component. For generation tasks, the content term  
 254 aggregates alignment and quality scores (e.g., preference feedback (Wu et al., 2023), unified scor-  
 255 ing (Wang et al., 2025), or text–image alignment (Geng et al., 2025a)), while for understanding  
 256 tasks, we apply rule-based checks to enforce correctness:

$$r_{\text{base}}(o) = r_{\text{format}} + r_{\text{content}}. \quad (4)$$

257 To explicitly encourage *useful* AD-Loop, we add a margin-based bonus to  $\mathcal{V}^+$  only when it outper-  
 258 forms the strongest  $\mathcal{V}^-$  candidate and the AD-Loop contributes meaningfully:

$$r(o_i^+) = r_{\text{base}}(o_i^+) + \lambda \mathbf{1}(\text{AD-Loop}|a) \max(0, r_{\text{base}}(o_i^+) - \max_j r_{\text{base}}(o_j^-) - \delta), \quad r(o_i^-) = r_{\text{base}}(o_i^-), \quad (5)$$

262 where  $\mathbf{1}(\text{AD-Loop}|a)$  is an indicator that the answer is correct and AD-Loop thinking is employed.  
 263 The margin parameter  $\delta$  requires the  $\mathcal{V}^+$  candidate to exceed the strongest  $\mathcal{V}^-$  baseline by at least  
 264  $\delta$ , filtering out spurious wins and favoring the simpler text-only mode unless a meaningful gain  
 265 is achieved.  $\lambda$  scales this bonus, balancing its impact against the base reward. Furthermore, we  
 266 introduce inter-group reward to indicate which thinking mode performs best for each query:

$$r_{\text{inter}}(o_i^m) = \begin{cases} 1, & \text{if } m = \arg \max_{m' \in \{+, -\}} \{r(o^+), r(o^-)\} \\ 0, & \text{otherwise} \end{cases}, \quad r_{\text{intra}}(o_i^m) = r(o_i^m), \quad m \in \{+, -\} \quad (6)$$

270  
271 Table 1: Comparison with baselines on multimodal understanding benchmarks. “Und.” and “Gen.”  
272 denote *understanding* and *generation*, respectively.  
273

Model	#Params	POPE $\uparrow$	MME-P $\uparrow$	MMB $\uparrow$	SEED $\uparrow$	GQA $\uparrow$	MMMU $\uparrow$	MM-Vet $\uparrow$
<b>• Und. Only</b>								
LLaVA-v1.5 (Liu et al., 2024a)	7B	85.9	1510.7	64.3	58.6	62.0	35.4	31.1
Qwen-VL-Chat (Bai et al., 2023)	7B	-	1487.5	60.6	58.2	57.5	-	-
IDEFICS (Laurençon et al., 2023)	8B	-	-	48.2	-	38.4	-	-
InstructBLIP (Dai et al., 2023)	13B	78.9	1212.8	-	-	49.5	-	25.6
<b>• Und. and Gen.</b>								
Emu3 (Wang et al., 2024)	8B	85.2	1244.0	58.5	68.2	60.3	31.6	37.2
Show-o (Xie et al., 2025)	1.3B	80.0	1097.2	-	-	58.0	26.7	-
Liquid (Wu et al., 2024a)	8B	-	1448.0	-	-	61.1	-	-
MMaDA (Yang et al., 2025a)	8B	86.1	1410.7	68.5	64.2	61.3	30.2	-
Janus-Pro (Chen et al., 2025c)	7B	87.4	1567.1	79.2	72.1	62.0	41.0	50.0
BAGEL (Deng et al., 2025)	7B	-	1687.0	85.0	-	-	55.3	67.2
<b>Ours</b>	7B	<b>90.1</b>	<b>1696.0</b>	<b>87.6</b>	<b>74.4</b>	<b>63.8</b>	<b>57.3</b>	<b>69.7</b>

284 Table 2: Comprehensive generation comparison on GenEval (Ghosh et al., 2023) benchmark. “Und.”  
285 and “Gen.” denote *understanding* and *generation*, respectively.  
286

Model	Single Obj. $\uparrow$	Two Obj. $\uparrow$	Counting $\uparrow$	Colors $\uparrow$	Position $\uparrow$	Attri. $\uparrow$	Overall $\uparrow$
<b>• Gen. Only</b>							
Emu3-Gen (Wang et al., 2024)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
SDXL (Podell et al., 2023)	0.98	0.74	0.39	0.85	0.15	0.23	0.55
FLUX.1-dev (Yang et al., 2024)	0.99	0.81	0.79	0.74	0.20	0.47	0.67
SD3-Medium (AI, 2025)	0.99	0.94	0.72	0.89	0.33	0.60	0.74
<b>• Und. and Gen.</b>							
Show-o (Xie et al., 2025)	0.95	0.52	0.49	0.82	0.11	0.28	0.53
TokenFlow-XL (Xie et al., 2025)	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Janus-Pro (Chen et al., 2025c)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
BAGEL (Deng et al., 2025)	0.99	0.94	0.81	0.88	0.64	0.63	0.82
MindOmni (Xiao et al., 2025b)	<b>0.99</b>	0.94	0.71	0.90	0.71	0.71	0.83
<b>Ours</b>	<b>0.98</b>	<b>0.94</b>	<b>0.83</b>	<b>0.90</b>	<b>0.80</b>	<b>0.74</b>	<b>0.86</b>

297  
298 **Optimization.** Following group-relative preference optimization, we combine an intra-group ad-  
299 vantage with an optional inter-group term:  
300

$$A_i = \underbrace{\left[ \frac{r_{\text{intra}}(o_i) - \text{mean}(r_{\text{intra}}(o_j))}{\text{std}(r_{\text{intra}}(o_j))} \right]}_{\text{GRPO for intra-group advantage } A_{\text{intra}}} + \gamma \underbrace{\left[ \frac{r_{\text{inter}}(o_i) - \text{mean}(r_{\text{inter}}(o_j))}{\text{std}(r_{\text{inter}}(o_j))} \right]}_{\text{GRPO for inter-group advantage } A_{\text{inter}}}, \quad (7)$$

304  
305 This group-relative advantage  $A_i$  of  $i$ -th response encourages the model to prioritize the response  
306 with higher relative quality.  $\gamma$  is the weighted parameter. Finally, following Guo et al. (2025), we  
307 apply a KL-divergence regularization with  $\beta$  hyperparameters and a clip trick to optimize the model.  
308

## 4 EXPERIMENTS

## 4.1 SETTINGS

312 **Dataset.** We construct the training dataset to emerge the synergy between understanding and gen-  
313 eration, as described in Sec. §3.2, encompassing interleaved reasoning for both understanding and  
314 generation. The detailed statistics are presented in Appendix §F.  
315316 **Implementations.** Our backbone is BAGEL-7B (Deng et al., 2025) in which SigLIP2-  
317 so400m/14 (Tschannen et al., 2025) is adopted as the image encoder for understanding, and the  
318 FLUX pre-trained VAE (Yang et al., 2024) is utilized as the image latent encoder/decoder for gen-  
319 eration. We train the model in two stages. At stage-1, we fine-tune on the curated interleaved  
320 reasoning corpus using a global batch size of 256 and a cosine learning-rate schedule with initial  
321 learning rate  $1 \times e^{-5}$ . The maximum number of a latent visual thought is set to  $K=16$  tokens. At  
322 stage-2, we implement RL with the VERL framework (Sheng et al., 2025). We fix the random seed  
323 to 42 to ensure reproducibility. The policy is optimized with AdamW, a constant learning rate of  
324  $2 \times e^{-6}$ , a global batch size of 64, and 8 rollouts per prompt. The objective combines the clipped  
325 policy-gradient loss ( $\epsilon=0.5$ ) with a KL regularization to a frozen reference model weighted by 0.01.  
326

324 Table 3: Comparison with different thinking strategies. *Isolated thinking*: understanding or genera-  
 325 tion performed in isolation;  $T$ : analyzing-only thinking;  $T + I$  (*explicit*): supervised textual–visual  
 326 (*explicit*) interleaving;  $T + \tilde{I}$  (*implicit*): learned textual–visual (*implicit*) interleaving;  $T / T + \tilde{I}$ :  
 327 adaptive interleaved thinking that automatically selects with or without interleaved thinking.  
 328

329 Think Strategy	330 <b>Understanding</b>			331 <b>Generation (WISE)</b>		
	332 MathVista↑	333 LogicVista↑	334 SAT↑	335 Cultural↑	336 Space↑	337 Biology↑
338 Isolated thinking	339 61.5	340 40.2	341 0.63	342 0.44	343 0.68	344 0.44
345 $T$	346 68.3	347 44.1	348 0.74	349 0.67	350 0.69	351 0.56
352 $T + I$ ( <i>explicit</i> )	353 72.9	354 46.6	355 0.81	356 0.73	357 0.74	358 0.64
359 $T + \tilde{I}$ ( <i>implicit</i> )	360 73.6	361 47.2	362 0.84	363 0.75	364 0.77	365 0.65
366 $T / T + \tilde{I}$	367 <b>75.8</b>	368 <b>49.5</b>	369 <b>0.89</b>	370 <b>0.79</b>	371 <b>0.78</b>	372 <b>0.68</b>



341 Figure 3: Qualitative comparison: original prompt (left), self-think mode, interleaved thoughts, and  
 342 text-only thoughts filtered from the interleaved thoughts (right). [V-T] means latent visual thoughts.  
 343

## 344 4.2 MAIN RESULTS

345 **Understanding.** We compare the proposed method against state-of-the-art understanding-only  
 346 and unified models on multimodal understanding benchmarks, including POPE (Li et al., 2023b),  
 347 MME (Zhang et al., 2021), MMB (Liu et al., 2024b), SEED (Li et al., 2023a), GQA (Hudson &  
 348 Manning, 2019), MMMU (Yue et al., 2024), MM-Vet (Yu et al., 2023). As highlighted in Table 1,  
 349 our method consistently yields the best overall results. This improvement stems from interleaving  
 350 visual–text reasoning, which effectively mitigates conflicts between understanding and generation  
 351 while unlocking their synergy. Notably, even compared to MMaDA (Yang et al., 2025a), which  
 352 utilizes RL-based thinking training, our approach still yields significant gains, confirming that gen-  
 353 eration can indeed enhance understanding.

354 **Generation.** We assess visual generation performance on the GenEval benchmark (Ghosh et al.,  
 355 2023). As shown in Table 2, our method achieves the highest overall score of 86%, outperforming  
 356 both generation-only and unified baselines. In particular, it delivers substantial improvements in  
 357 fine-grained attributes in positional accuracy and attribute correctness. Moreover, compared with  
 358 MindOmni (Xiao et al., 2025b) that employs textual thinking only, our approach achieves better  
 359 performance, clearly demonstrating the effectiveness of incorporating visual thoughts.

## 360 4.3 ABLATION ON THINKING TYPES

361 We systematically evaluate reasoning strategies across both understanding and generation bench-  
 362 marks, including MathVista (Lu et al., 2023), LogicVista (Xiao et al., 2024), SAT (Ray et al., 2024),  
 363 and WISE (Niu et al., 2025). Results are summarized in Table 3. First, compared with the no-think  
 364 setting, equipping the model with reasoning capability (*T*-*think*) yields substantial improvements in  
 365 both understanding and generation, confirming the effectiveness of reasoning guidance. Second,  
 366 augmenting textual reasoning with visual thoughts further boosts performance, highlighting that vi-  
 367 sual cues supply fine-grained details complementary to text. When comparing explicit ( $T + I$ -*think*)  
 368 and implicit ( $T + \tilde{I}$ -*think*) visual thoughts, we observe only marginal differences. This can be partly  
 369 owed to the learning paradigm, and explicit supervision can introduce pixel-level noise, whereas  
 370 implicit reasoning captures compact salient cues. Nonetheless, combining the two in our hybrid  
 371 setting achieves the best results across all tasks, suggesting a strong complementarity.

372 Further qualitative evidence is provided in Fig. 3. With the raw prompt alone, the model tends to  
 373 generate superficial, semantically shallow outputs. Adding self-think produces more detailed de-  
 374 scriptions, yet still overly abstract and often misaligned with user intent. By contrast, interleaved  
 375 thoughts guide faithful, detail-oriented outputs (e.g., correct wheels/screens). Finally, filtering in-  
 376 terleaved traces to text-only frequently reintroduces errors (e.g., lighting/positioning), underscoring  
 377 the necessity of visual thoughts for high-fidelity controllability.

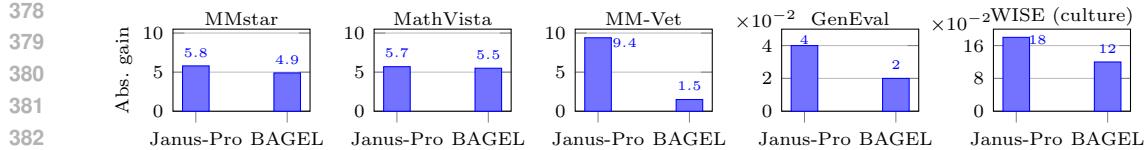


Figure 4: Absolute performance gain after applying **AD-Loop**, comparing Janus-Pro with discrete tokenization and BAGEL with continuous embedding for visual thoughts learning.

Table 4: Comparison of latent visual representations learn by the generation (Gen.) vs. understanding (Und.) encoder.

	MMstar	MathVista	LogicVista	GenEval	WISE (cultural)	WISE (Biology)
From Gen. Encoder	54.9	75.8	47.5	0.86	0.79	0.68
From Und. Encoder	51.6	70.9	44.3	0.84	0.71	0.61



Figure 5: Examples of latent visual thoughts. Each case shows the original image (left) and the corresponding visual thoughts (right), capturing abstract visual structures.

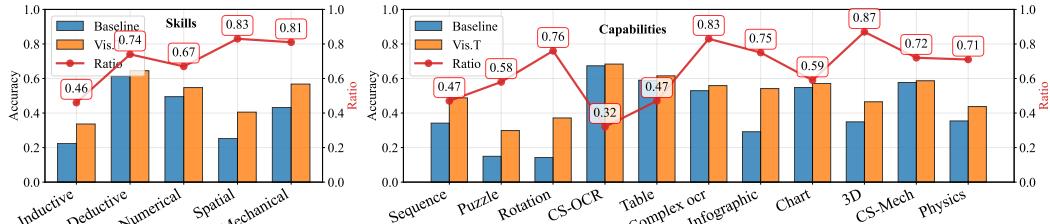


Figure 6: Performance across skills and capabilities on the LogicVista dataset, comparing models with and without visual thoughts, alongside the proportion of visual-thought usage.

#### 4.4 ANALYSES AND DISCUSSION

In this section, we are about to answer the following in-depth questions:

**RQ-1: Can this method extend into different structures of unified MLLMs?** Current unified MLLMs adopt different architectures for visual content generation. One line of work, including our backbone, generates continuous embeddings, while another, exemplified by Janus-Pro (Chen et al., 2025c), produces discrete tokens. We apply our proposed method to both paradigms, as described in Sec. §3.2. As shown in Fig. 4, Unified-R1 consistently improves both understanding and generation across the two settings. These results demonstrate that our approach is architecture-agnostic and can serve as a general mechanism for strengthening emergent task synergy in unified MLLMs.

**RQ-2: Should visual thoughts be derived from understanding or generation?** We compare visual thoughts derived from the understanding versus the generation encoder. As shown in Table 4, using the generation encoder yields consistently better results for both understanding and generation tasks. This can be attributed to two aspects. On the one hand, we observe that models converge faster under this setting, likely because generation-oriented thoughts are already extensively pre-trained. Moreover, generation-based visual thoughts inherently capture both semantic and pixel-level information, making them more informative and beneficial for multimodal reasoning.

**RQ-3: What do the implicit visual thoughts look like?** We visualize the clustering results in Fig. 5. The observations align well with our expectations: latent visual thoughts encode semantically coherent information while preserving coarse pixel-level structures. This allows the model to recover the overall contours of the original image and to unify conceptually similar regions. For example, in case (4), distinct regions depicting watermelons and lemons are consistently represented by the same latent token, reflecting their shared conceptual category.

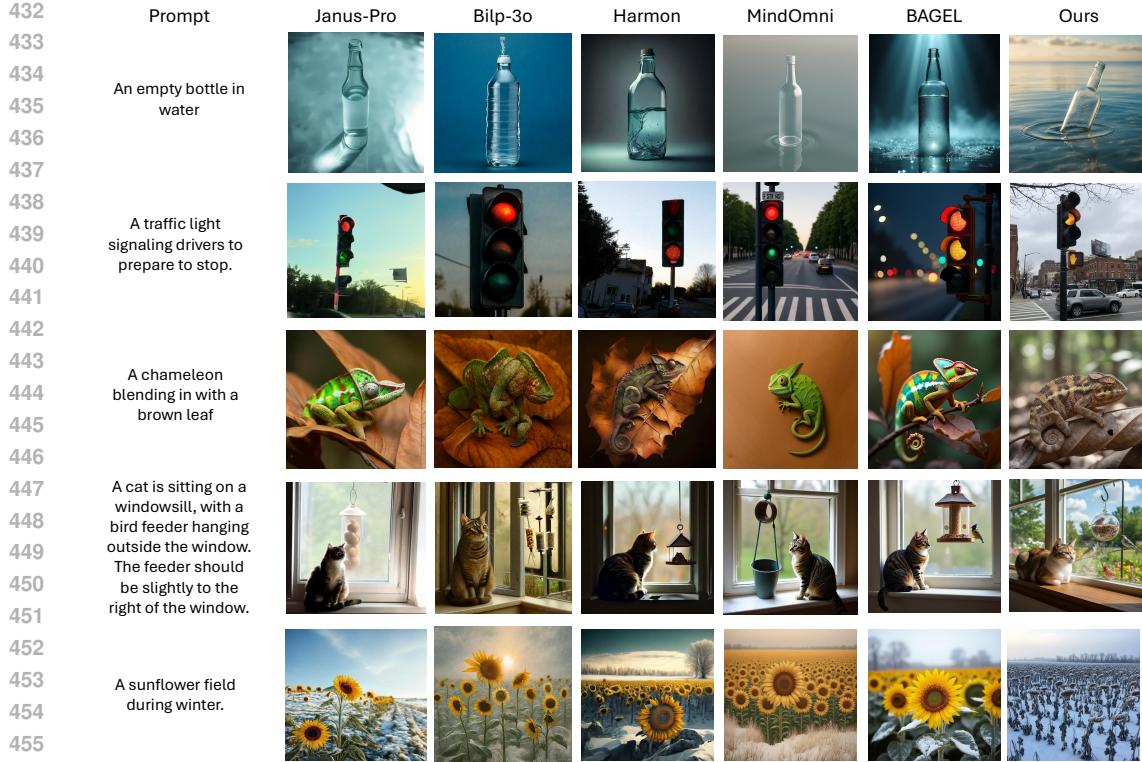


Figure 7: Comparison of unified MLLMs on T2I generation. Existing models often fail on prompts requiring deeper reasoning, whereas our approach yields more faithful outcomes.

**RQ-4: When are the visual thoughts needed?** By comparing performance across different task scenarios and analyzing the proportion of visual-thought usage, we obtain the results shown in Fig. 6. Overall, integrating AD-Loop thoughts improves performance across a wide range of questions, with pronounced gains in spatial and mechanistic reasoning. Fine-grained trends show preferential activation for rotation, complex OCR, and 3D perception, while usage drops for tables, sequences, and symbolic reasoning, where text-only chains suffice. These patterns indicate our adaptive policy that selectively invokes visual thoughts when they offer the greatest benefit.

**RQ-5: Case study.** Finally, we provide qualitative analyses of our model’s capabilities in text-to-image generation. As demonstrated in Fig. 7, baseline models often falter on reasoning-intensive prompts, capturing superficial cues rather than underlying logic—for example, outputting red-green for “a traffic light signaling drivers to prepare to stop” (correct: yellow), rendering a green chameleon on a brown leaf (ignoring adaptive coloration), or producing bright sunflowers for “a sunflower field during winter” (ignoring seasonality). In contrast, our method consistently aligns outputs with prompt semantics and the required reasoning. On the understanding side (Fig. 10), our model exhibits stronger spatial reasoning, correctly localizing objects and motions where baselines fail to do so. These cases highlight how interleaved reasoning facilitates faithful, reasoning-aware generation and a more robust understanding. Additional examples can be found in Appendix §H.

## 5 CONCLUSION

In this work, we introduce a unified interleaved thinking framework, i.e., AD-Loop, for synergizing understanding and generation in UVLMs. Our proposed two-stage learning paradigm first initializes an interleaved AD-Loop thinking through supervised fine-tuning, and then employs hybrid reinforcement learning to enable the model to adaptively invoke interleaved thinking when beneficial. This allows the model to learn *when* and *how* to leverage different capabilities, thereby unlocking genuine synergy across tasks. Extensive experiments on diverse multimodal understanding and generation benchmarks demonstrate the effectiveness of our approach, with particularly strong gains on reasoning-driven tasks. These results highlight the potential of interleaved AD-Loop as a general mechanism for advancing unified multimodal intelligence.

486 REFERENCES  
487

488 Stability AI. Stable diffusion 3 medium: Multimodal diffusion transformer for pho-  
489 torealistic text-to-image generation, 2025. URL <https://stability.ai/news/stable-diffusion-3-medium>.

490

491 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,  
492 Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

493

494 Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and  
495 Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

496

497 Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang  
498 Xie. Sft or r1? an early investigation into training r1-like reasoning large vision-language models.  
499 *arXiv preprint arXiv:2504.11468*, 2025a.

500

501 Juhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi  
502 Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal  
503 models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025b.

504

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

508

509 Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jin-  
510 peng Wang, Zhi Chen, Wanxiang Che, and Libo Qin. Visual thoughts: A unified perspective of  
511 understanding multimodal chain-of-thought. *CoRR*, abs/2505.15510, 2025a.

512

513 Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and  
514 Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language  
515 models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 23678–23686,  
516 2025b.

517

518 Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large mul-  
519 timodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

520

521 Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu.  
522 Thinking with generated images. *CoRR*, abs/2505.22525, 2025.

523

524 Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li,  
525 Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models  
526 with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267,  
527 2023.

528

529 Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao  
530 Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv*  
531 *preprint arXiv:2505.14683*, 2025.

532

533 Melissa Ellamil, Charles Dobson, Mark Beeman, and Kalina Christoff. Evaluative and generative  
534 modes of thought during the creative process. *NeuroImage*, 59(2):1783–1794, 2012.

535

536 Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu  
537 Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li. Got: Unleashing reasoning capability  
538 of multimodal large language model for visual generation and editing. *CoRR*, abs/2503.10639,  
539 2025.

540

541 Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin  
542 Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregres-  
543 sive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025a.

544

545 Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin  
546 Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregres-  
547 sive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025b.

540 Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework  
 541 for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:  
 542 52132–52152, 2023.

543

544 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
 545 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms  
 546 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

547

548 Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and  
 549 Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models.  
 550 *CoRR*, abs/2503.06749, 2025.

551

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

554

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

558

559 Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning  
 560 for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.17534*, 2025b.

561

562 Lingjie Jiang, Xun Wu, Shaohan Huang, Qingxiu Dong, Zewen Chi, Li Dong, Xingxing Zhang,  
 563 Tengchao Lv, Lei Cui, and Furu Wei. Think only when you need with large hybrid-reasoning  
 564 models. *arXiv preprint arXiv:2505.14631*, 2025c.

565

566 Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univ: Unified visual  
 567 representation empowers large language models with image and video understanding. In *Pro-  
 ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–  
 568 13710, 2024.

569

570 Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie  
 571 Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads.  
 572 *arXiv preprint arXiv:2412.00127*, 2024.

573

574 Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov,  
 575 Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open  
 576 web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information  
 577 Processing Systems*, 36:71683–71702, 2023.

578

579 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan  
 580 Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on  
 581 Machine Learning Research*, 2024.

582

583 Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Bench-  
 584 marking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*,  
 585 2023a.

586

587 Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and  
 588 Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint  
 589 arXiv:2501.07542*, 2025.

590

591 Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating  
 592 object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

593

594 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction  
 595 tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,  
 596 pp. 26296–26306, 2024a.

597

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

594 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-  
 595 Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of  
 596 foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

597

598 Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan  
 599 Qi. Unitok: A unified tokenizer for visual generation and understanding. *CoRR*, abs/2502.20321,  
 600 2025.

601 Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-  
 602 Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought.  
 603 In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14268–14280,  
 604 2025.

605

606 Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran  
 607 Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation  
 608 for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

609 Allan Paivio, James M Clark, et al. Dual coding theory and education. *Pathways to literacy achievement for high poverty children*, 1:149–210, 2006.

610

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

614

615 Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan  
 616 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,  
 617 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin  
 618 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,  
 619 Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang,  
 620 Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

621

622 Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha  
 623 Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training  
 624 for multimodal language models. *arXiv e-prints*, pp. arXiv–2412, 2024.

625

626 Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hong-  
 627 sheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and  
 628 benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*,  
 37:8612–8642, 2024.

629

630 Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun  
 631 Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable  
 632 and generalizable r1-style large vision-language model. *CoRR*, abs/2504.07615, 2025.

633

634 Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueling Zhuang. Hugging-  
 635 gpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information  
 636 Processing Systems*, 36:38154–38180, 2023.

637

638 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,  
 639 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings  
 640 of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

641

642 Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*,  
 171(3972):701–703, 1971.

643

644 Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally  
 645 can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

646

647 Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie  
 648 Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via  
 649 visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

648 Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan.  
 649 Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint*  
 650 *arXiv:2406.06525*, 2024.

651 Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint*  
 652 *arXiv:2405.09818*, 2024a.

653 Qwen Team. Qvq: To see the world with wisdom, December 2024b. URL <https://qwenlm.github.io/blog/qvq-72b-preview/>.

654 Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdul-  
 655 mohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff,  
 656 Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language  
 657 encoders with improved semantic understanding, localization, and dense features, 2025.

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

661 Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multi-  
 662 modal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

663 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny  
 664 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*  
 665 *neural information processing systems*, 35:24824–24837, 2022.

666 Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu,  
 667 Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified  
 668 multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern*  
 669 *Recognition Conference*, pp. 12966–12977, 2025a.

670 Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and  
 671 Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators. *arXiv*  
 672 *preprint arXiv:2412.04332*, 2024a.

673 Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multi-  
 674 modal llm. In *Forty-first International Conference on Machine Learning*, 2024b.

675 Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li,  
 676 and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding  
 677 and generation. *arXiv preprint arXiv:2503.21979*, 2025b.

678 Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li.  
 679 Human preference score v2: A solid benchmark for evaluating human preferences of text-to-  
 680 image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

681 Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li,  
 682 Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceed-  
 683 ings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025a.

684 Yicheng Xiao, Lin Song, Yukang Chen, Yingmin Luo, Yuxin Chen, Yukang Gan, Wei Huang, Xiu  
 685 Li, Xiaojuan Qi, and Ying Shan. Mindomni: Unleashing reasoning generation in vision language  
 686 models with rgpo. *arXiv preprint arXiv:2505.13031*, 2025b.

687 Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning  
 688 benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

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

693 Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let  
 694 vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

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

705 Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin,  
 706 Hao Li, Xue Xu, et al. Can understanding and generation truly benefit together—or just coexist?  
 707 *arXiv preprint arXiv:2509.09666*, 2025.

708 Chenglin Yang, Celong Liu, Xueqing Deng, Dongwon Kim, Xing Mei, Xiaohui Shen, and Liang-  
 709 Chieh Chen. 1.58-bit flux. *arXiv preprint arXiv:2412.18653*, 2024.

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

712 Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery:  
 713 Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*,  
 714 2025b.

715 Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang,  
 716 Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning  
 717 and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

718 Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang,  
 719 and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv  
 720 preprint arXiv:2308.02490*, 2023.

721 Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan,  
 722 Deli Zhao, Tingyang Xu, Zhongyu Wei, et al. VI-cogito: Progressive curriculum reinforcement  
 723 learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*, 2025.

724 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens,  
 725 Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-  
 726 modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF  
 727 Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

728 Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo  
 729 Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective  
 730 multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.

731 Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language  
 732 image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*,  
 733 pp. 11975–11986, 2023.

734 Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin  
 735 Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal lilm with discrete sequence  
 736 modeling. *arXiv preprint arXiv:2402.12226*, 2024.

737 Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei  
 738 Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Chain-of-focus: Adaptive visual search and zoom-  
 739 ing for multimodal reasoning via RL. *CoRR*, abs/2505.15436, 2025a.

740 Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei  
 741 Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for  
 742 multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025b.

743 Yunhang Shen Yulei Qin Mengdan Zhang, Xu Lin Jinrui Yang Xiawu Zheng, Ke Li Xing Sun Yun-  
 744 sheng Wu, Rongrong Ji Chaoyou Fu, and Peixian Chen. Mme: A comprehensive evaluation  
 745 benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2021.

746 Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and  
 747 Xing Yu. Deepeyes: Incentivizing “thinking with images” via reinforcement learning. *CoRR*,  
 748 abs/2505.14362, 2025.

749 Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob  
 750 Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and  
 751 diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

## APPENDIX INDEX

This supplementary material includes the following sections:

- Clarification on the Use of Large Language Models (cf. §A).
- Ethics Statement (cf. §B).
- Reproducibility Statement (cf. §C).
- Detailed methodology (cf. §D).
- Comparison with Existing Synergetic Learning Methods (cf. §E).
- Detailed Dataset Construction (cf. §F).
- Detailed Setting (cf. §G).
- Extended Experimental Results (cf. §H).
- **Thinking Strategies Comparison with Existing VLMs via Visual Information (cf. §I).**

## A THE USE OF LARGE LANGUAGE MODELS (LLMs)

In this paper, we employ the large language models (LLMs) to help the dataset construction and improve the clarity and readability of English writing. Specifically, we employ the QVQ-72B-Preview to construct the Inter-T2I dataset. Additionally, LLMs were utilized to refine sentence structure, correct grammatical errors, and enhance the overall presentation of our draft. The technical content, research ideas, experimental design, analysis, and conclusions were entirely conceived, implemented, and validated by the authors without reliance on LLMs.

## B ETHICS STATEMENT

All datasets used in this work are publicly available and open-source. During the process of constructing our experimental data, we employed open-source text-to-image generation models. To mitigate risks of harmful content, discrimination, or bias, all generated samples were manually screened and filtered to ensure suitability for research purposes. Our approach is built on an open-source foundation model that provides strong assurance for generating non-harmful and bias-free content. Nonetheless, as with any generative system, it is impossible to guarantee that unintended or potentially harmful outputs will never occur. We therefore emphasize that users and practitioners should exercise caution when deploying such models in downstream applications, especially in sensitive domains.

We do not involve any human subjects, private or proprietary data, or personally identifiable information (PII) in this research. The work complies with community standards on data usage and research integrity. No legal, ethical, or security violations are posed by the methodologies or experiments presented in this paper. Furthermore, we highlight that our contributions are intended solely for academic and scientific exploration. We explicitly discourage misuse of the proposed methods for generating harmful, misleading, or discriminatory content. Future applications should integrate appropriate safeguards, fairness considerations, and content moderation mechanisms.

## C REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have made a concerted effort to provide all necessary details and materials. We provide comprehensive details of the proposed AD-Loop framework, including its definition, input–output formulation, and implementation (Section §3.1). The model backbone and training methodology are described in detail in Section §3 and Appendix §D. We further report all hyperparameter settings and training configurations in Section 4.1 and Appendix §G, using fixed random seeds to ensure the replicability of the experiments. All datasets used in this study are publicly available open-source resources, and the data construction process, along with the amount of data used at each training stage, is thoroughly documented in Appendix §F. Finally, we will release the full codebase and data processing scripts to the community upon acceptance.

810 D DETAILED METHODOLOGY  
811812 D.1 LATENT VISUAL THOUGHTS CONSTRUCTION.  
813

814 Given an image visual thought  $I$ , the image encoder yields a grid of latent tokens  $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^N$ ,  
815 where the image encoder can be either a VQ tokenizer (Sun et al., 2024; Chen et al., 2025c) or a  
816 VAE encoder (Yang et al., 2024; Deng et al., 2025). We then calculate the local density  $\rho_i$  of the  
817 token  $\mathbf{z}_i \in \mathbf{Z}$  by referring its neighbors:

$$818 \rho_i = \exp\left(-\frac{1}{K} \sum_{\mathbf{z}_m \in \text{KNN}(\mathbf{z}_i, \mathbf{Z})} \|\mathbf{z}_m, \mathbf{z}_i\|^2\right), \quad (8)$$

821 where  $\text{KNN}(\mathbf{z}_i, \mathbf{Z})$  denotes the  $K$ -nearest neighbors of  $\mathbf{z}_i$  in  $\mathbf{Z}$ . We then measure the minimal  
822 distance  $\delta_i$  between the feature  $\mathbf{z}_i$  and other features with higher density:

$$823 \delta_i = \begin{cases} \min_{m: \rho_m > \rho_i} \varphi(\mathbf{z}_m, \mathbf{z}_i), & \text{if } \exists m, n : \rho_m > \rho_i \\ \max_m \varphi(\mathbf{z}_m, \mathbf{z}_i), & \text{otherwise} \end{cases} \quad (9)$$

825 In essence,  $\delta_i$  represents the distance between the given token  $\mathbf{z}_i$  from other high-density tokens. We  
826 summarize the score  $s_i$  of the feature by combining the local density  $\rho_i$  and minimal distance  $\delta_i$  as  
827  $\rho_i \times \delta_i$ . We identify those tokens with relatively high scores,  $s_i$ , as cluster centers and then allocate  
828 other tokens to their nearest cluster center based on the Euclidean distances. Finally, we utilize the  
829 average token within each cluster to represent the corresponding cluster. The latent visual thought  
830 token of the merged patch token is the union of the vision regions within the corresponding cluster.

831 D.2 REWARD MODEL  
832

833 In **Stage 2**, each model response is evaluated with two complementary signals: (i) a *format reward*  
834 that enforces structural validity (e.g., required fields, schema conformity), and (ii) a *content reward*  
835 that measures semantic fidelity and task-specific quality. Below, we describe the content rewards  
836 used for different task types.  
837

838 **Understanding-Task Reward Model** For tasks with deterministic targets, such as multiple-  
839 choice or numeric questions, we employ rule-based matching after normalization (e.g., case folding,  
840 whitespace removal, and unit normalization). This yields a precise, low-variance signal of correctness.  
841 For open-ended understanding questions, we rely on an external, learned reward model as  
842 the judge; specifically, we use InternLM-XComposer2.5-Reward (Zang et al., 2025), which scores  
843 responses by holistic relevance and coherence with the instruction.  
844

845 **Generation-Task Reward Model** For assessing the quality of generated images, we employ two  
846 complementary criteria: 1) **Semantic alignment score**: Measures agreement between the generated  
847 image and the ground-truth prompt via cosine similarity between CLIP image and text embeddings.  
848 2) **Human preference alignment score**: Captures perceived aesthetic quality and prompt adherence  
849 using learned preference models, namely HPS v2 (Wu et al., 2023) and ImageReward (Xu et al.,  
850 2023). These signals are aggregated to provide fine-grained, content-aware feedback that guides the  
851 model toward faithful, high-quality generations while maintaining the required output format.  
852

853 D.3 THE OPTIMIZATION OBJECTIVES IN STAGE-2  
854

855 After obtaining the advantages, we apply the standard objectives as described in Guo et al. (2025)  
856 to optimize our model. In addition, to prevent the optimized policy  $\pi_\theta$  from diverging excessively  
857 from the reference model  $\pi_{\text{ref}}$ , a KL-divergence regularization term  $\mathbb{D}_{\text{KL}}$  is introduced. The overall  
858 optimization objective is:  
859

$$\max_{\pi_\theta} \mathbb{E}_{[q \sim D_{\text{rl}}, \{o_i\}_{i=1}^G \sim \pi_\theta(O|q)]} \left[ \frac{1}{G} \sum_{i=1}^G \min\left(\frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)}, \text{clip}\left(\frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)}, 1 - \epsilon, 1 + \epsilon\right) A_i\right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right], \quad (10)$$

860 where  $\epsilon, \beta$  are the hyper-parameters.  
861  
862  
863

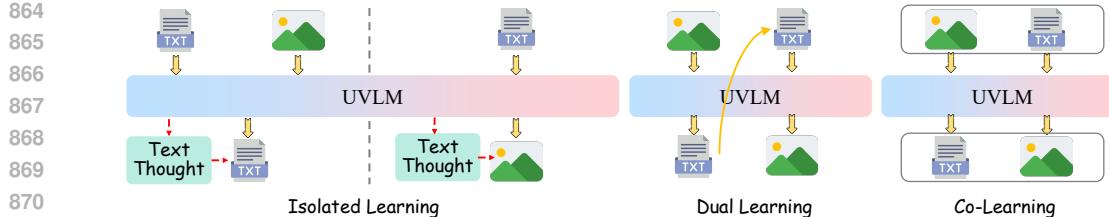


Figure 8: Comparison of existing mechanisms for synergizing understanding and generation, including isolated learning where the two abilities are trained independently, dual learning which leverages cross-modal reconstruction for mutual supervision, and co-learning which jointly optimizes both tasks with paired samples.

## E COMPARISON WITH EXISTING SYNERGETIC LEARNING METHODS

We here position our approach in relation to existing methods for synergistic learning between understanding and generation. As illustrated in Fig. 8, prior work can be broadly grouped into two major directions.

**Unified Models for Synergistic Learning.** A common line of research aims to unify understanding and generation within a single framework. Representative examples (Xie et al., 2025; Wu et al., 2025a; Deng et al., 2025; Xiao et al., 2025b) include models that adopt a purely autoregressive formulation, or hybrid designs that combine autoregressive and diffusion-based paradigms. In these approaches, synergy is encouraged through shared parameters, which enable implicit interaction between the two tasks. However, during training and inference, the two abilities remain largely independent. As a result, such models primarily learn to master understanding and generation in parallel, rather than achieving genuine synergy in task-solving.

**Learning Optimization for Synergistic Learning.** Another approach is to design a learning schema that fosters synergy between understanding and generation. For instance, dual learning methods (Yan et al., 2025) translate visual inputs into textual descriptions and then regenerate visuals from those descriptions, with reconstruction quality serving as the optimization signal. Co-learning strategies (Jiang et al., 2025b) further extend this idea by coupling each sample with both generation prompts and multimodal understanding queries, thereby jointly improving performance on the two tasks. While effective for mutual supervision, these approaches still operate at the stage of co-training skills, rather than enabling active collaboration between them during task execution.

**Our Contribution.** In contrast, our work introduces a fundamentally different perspective. We argue that real synergy should emerge not only during the learning phase but also in the problem-solving process itself. Specifically, we propose the Analyzing-Drafting problem-solving Loop (AD-Loop), a novel problem-solving mechanism in which understanding and generation are interleaved at appropriate moments to jointly address a task. Furthermore, we develop a two-stage training strategy that equips unified multimodal models with this capability. In particular, our second-stage hybrid learning scheme enables the model to adaptively and intelligently decide when to invoke understanding or generation, thereby achieving organic integration of the two abilities and realizing genuine synergy.

## F DETAILED DATASET CONSTRUCTION

Here, we outline the process of constructing training data. The overall statistics are presented in Table 5, and an instance example visualization of each constructed dataset is shown in Fig. 9.

**AD-Loop Dataset for Understanding Task.** To construct the interleaved analyzing–drafting loop for understanding tasks, we draw upon the following representative sources:

- CMoT (Cheng et al., 2025b) is a chain of multimodal-thought dataset that requires multimodal input and multi-output reasoning output. It consists of four categories: (1) Visual

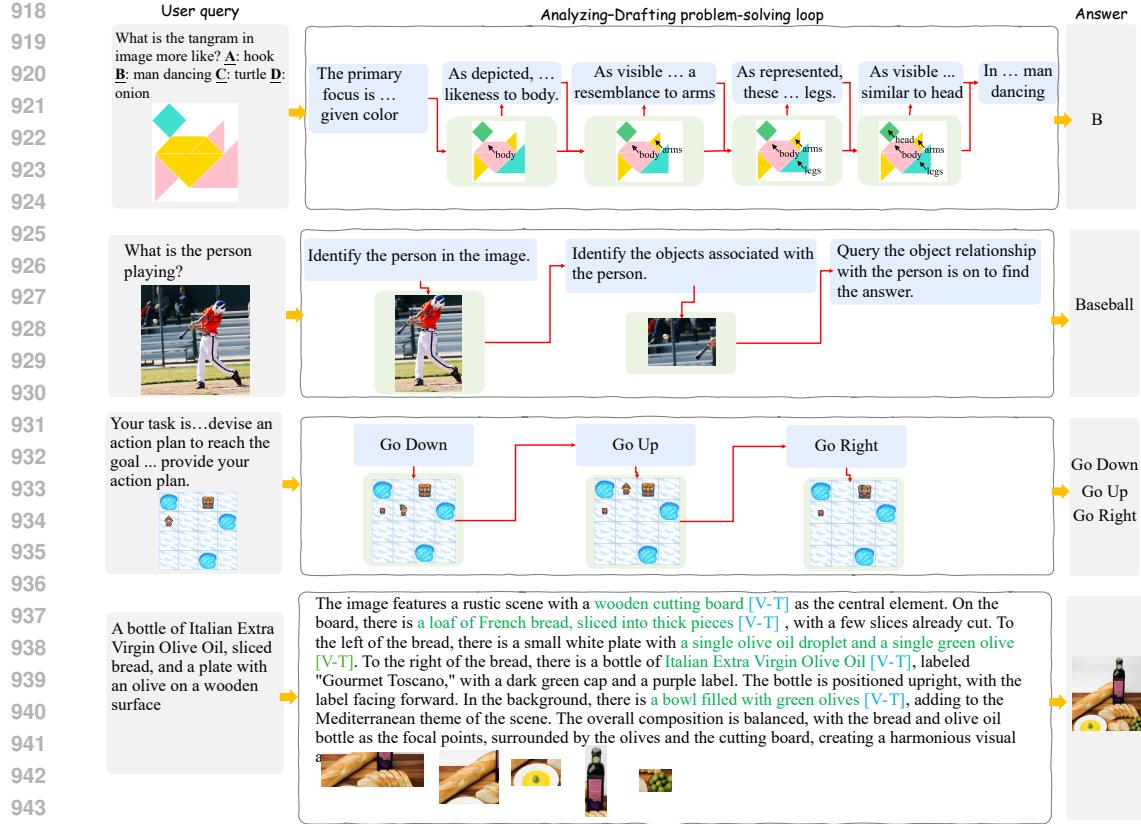


Figure 9: Visualization of the constructed dataset. Given a user query, we establish an analyzing–drafting problem-solving loop in which textual and visual thoughts alternate dynamically, ultimately leading to the final answer.

Creation, (2) Visual Deletion, (3) Visual Update, and (4) Visual Selection to comprehensively explore complex visual operations and concise expression in real scenarios. **The final average number of AD-Loop is 3.**

- **Visual-CoT** (Shao et al., 2024) provides intermediate reasoning annotations with bounding boxes that highlight critical regions necessary for answering visual reasoning questions. We derive explicit visual thoughts by zooming into and cropping the annotated regions, which serve as supervision signals for the visual-thought channel, **resulting in an average of 2 per AD-Loop trajectory**.
- **CoF** (Zhang et al., 2025b) identifies focus regions during question answering. We derive explicit visual thoughts by zooming into and cropping the annotated regions, which serve as supervision signals for the visual-thought channel, **yielding an average of 3 per sample..**
- **Visual Spatial Planning (VSP)** Yang et al. (2025b) formulates schematic grid-based navigation tasks, where an agent must move from a designated start to the destination while avoiding “holes”. Following Li et al. (2025), we render these states using the OpenAI Gym framework (Brockman et al., 2016), with the initial map and action sequence as inputs. **Visual thoughts correspond to the visualization of each movement step, with an average of 3 latent visual thoughts.**

**AD-Loop Dataset for Generation Task.** For generation tasks, we construct interleaved analyzing–drafting resources from the following datasets and models:

- **GoT-T2I** (Fang et al., 2025) provides both reasoning traces and bounding-box annotations of salient objects. We crop the corresponding image regions based on the bounding-box information and append these cropped patches directly after the associated textual descriptions, thereby forming explicit visual representations. **We treat the annotated bounding boxes of salient objects as latent visual thoughts, resulting in an average of 4.**

972 Table 5: The statistics of the constructed dataset for two-stage learning. ‘Avg.’ indicates the average  
 973 number of the latent visual thoughts within the datasets.

974	Task	Dataset	Avg.	Stage-1	Stage-2	Total
976	Understanding	CMoT (Cheng et al., 2025b)	3	1K	1K	2K
977		Visual-CoT (Shao et al., 2024)	2	12K	1K	13K
978		CoF (Zhang et al., 2025b)	3	4K	1K	5K
979		VSP (Yang et al., 2025b)	3	3K	1K	4K
980	Generation	GoT-T2I (Fang et al., 2025)	4	12K	3K	15K
981		Inter-T2I	4	10K	2K	12K

982 Table 6: Hyperparameters and data sampling ratios for Stage-1 and Stage-2.

983		Stage-1	Stage-2	
985	<b>Hyperparameters</b>			
986	Learning rate	$1 \times 10^{-5}$	$2 \times 10^{-6}$	
987	Batch size	256	64	
988	LR scheduler	Cosine	Constant	
989	Optimizer	AdamW	AdamW	
990	Warm-up steps	200	200	
991	<b>Data Number(K)</b>			
992	Only-Text	0.0	4	
993	Understanding Data	20	4	
994	Generation Data	22	5	

996

- 997 • **Inter-T2I** We leverage *X-to-Image* (Xiao et al., 2025a) corpora to obtain high-quality  
 998 prompt–image pairs, and employ QVQ-72B-Preview (Team, 2024b) to construct reflective  
 999 critiques conditioned on visual inputs. To enrich intermediate visual thoughts, we adopt  
 1000 *Fluxl-dev* (Yang et al., 2024) to generate visual hypotheses guided by subgoals extracted  
 1001 from the prompts. We generate visual hypotheses based on sub-goals extracted from the  
 1002 prompts and retain only the first occurrence of each object to avoid duplication. The aver-  
 1003 age number of latent visual thoughts is 4.

## 1004 G DETAILED SETTINGS

### 1005 G.1 TRAINING CONFIGURATION

1006 We adopt two distinct UVLMs as the backbone of our framework. While both models incor-  
 1007 porate dedicated heads for video understanding and generation, they differ fundamentally in their ap-  
 1008 proaches to video synthesis: Janus-Pro employs an LLM backbone that generates *discrete tokens*,  
 1009 whereas BAGEL adopts a design that generates *continuous tokens*. Below, we detail the architectural  
 1010 designs of these two models.

1011 BAGEL (7B) (Deng et al., 2025) is initialized from Qwen2.5 LLM (Qwen et al., 2025). For vi-  
 1012 sual understanding, BAGEL adopts SigLIP2-so400m/14 (Tschanne et al., 2025) with a fixed 384-  
 1013 resolution to convert the raw pixels into tokens. For visual generation, BAGEL utilizes a pre-trained  
 1014 VAE model from FLUX (Yang et al., 2024) to convert images from pixel space to latent space and  
 1015 vice versa. The latent representation is then processed by a  $2 \times 2$  patch embedding layer to reduce  
 1016 the spatial size and match the hidden dimension of the LLM backbone.

1017 Janus-Pro (7B) (Chen et al., 2025c) utilizes the SigLIP (Zhai et al., 2023) to extract high-dimensional  
 1018 semantic features from images. These features are flattened from a 2-D grid into a 1-D sequence,  
 1019 and an understanding adaptor is used to map these image features into the input space of the LLM  
 1020 for understanding. For visual generation tasks, Janus-Pro adopts the VQ tokenizer from Sun et al.  
 1021 (2024), which discretizes images into token IDs for autoregressive generation.

1022 During the training, the maximum number of each latent visual thought is set to 16 tokens. At stage  
 1023 2, we set  $\lambda = 1.0$  and a margin  $\delta = 0.2$  for RL training. We set the weighting hyper-parameters

1026 Table 7: Performance comparison of explicit and latent visual thoughts on the spatial visual planning  
 1027 task at levels 3 and 6.

Type	Model	Level 3	Level 6
Explicit	Anole (Chern et al., 2025)	0.02	0.00
	MVoT (Li et al., 2025)	0.21	0.03
Implicit	Mirage (Yang et al., 2025b)	0.75	0.39
	Ours	<b>0.81</b>	<b>0.47</b>

1035  
 1036  $\alpha = 1$ , and  $\gamma = 1$ . We set the KL coefficient to  $\beta = 0.001$  and  $\epsilon = 0.5$ . Besides the interleaved data,  
 1037 we also leverage 4K text-only reasoning data (Huang et al., 2025; Li et al., 2024) in Stage 2. The  
 1038 other necessary parameters are shown in Table 6.

## 1042 G.2 EVALUATION METRICS

1044 Here, we detail the evaluation metrics employed in the experiments. For understanding tasks, we  
 1045 report *accuracy*, following standard practice Xie et al. (2025); Deng et al. (2025); Ma et al. (2025),  
 1046 to measure the percentage of correctly predicted answers on each dataset. For generation tasks, we  
 1047 conduct evaluations on two benchmarks:

- 1048 • **(1) GenEval.** We adopt the official GenEval toolkit<sup>1</sup> to assess text-to-image generation  
 1049 quality. GenEval comprises six task categories, including single object, two objects, counting,  
 1050 colors, position, and attribute binding. The *GENEVAL score* is formulated as a binary  
 1051 correctness measure, indicating whether all elements specified in the prompt are faithfully  
 1052 rendered in the generated image. This setup directly evaluates the model’s compositional  
 1053 alignment between textual instructions and visual outputs.
- 1054 • **(2) WISE.** We select a subset of WISE domains (e.g., biology, culture, and space) for eval-  
 1055 uation. Following prior work Deng et al. (2025); Chen et al. (2025c), we report *WiScore*,  
 1056 the primary metric of the benchmark, which emphasizes the accuracy of depicted objects  
 1057 and entities grounded in world knowledge. WiScore is computed as a weighted combina-  
 1058 tion of three components: Consistency, Realism, and Aesthetic Quality. Higher WiScore  
 1059 values indicate stronger capability in generating images that correctly represent real-world  
 1060 concepts while maintaining visual plausibility.

## 1061 H EXTENDED EXPERIMENTAL RESULTS

1063 **Comparison of Explicit and Implicit Visual Thoughts.** We compare existing explicit and im-  
 1064 plicit approaches on the spatial visual planning task. As shown in Table 7, explicit methods yield  
 1065 the weakest performance. A likely reason is that when the model is required to explicitly draft  
 1066 intermediate visual thoughts, the quality of these drafts is heavily constrained by the model’s in-  
 1067 herent generation capability. In such cases, the model struggles to construct faithful visualizations  
 1068 from high-level semantic representations, resulting in low-quality visual thoughts. Building upon  
 1069 these imperfect drafts further propagates errors, resulting in cumulative degradation of reasoning.  
 1070 By contrast, latent (implicit) visual thoughts achieve consistently stronger results, as they bypass the  
 1071 limitations of explicit drafting and enable more faithful internal reasoning representations.

1072 **Comparison on Leveraging Understanding for Text-to-Image Generation.** Table 8 reports the  
 1073 performance of different approaches that exploit understanding to enhance text-to-image genera-  
 1074 tion. Existing methods predominantly rely on textual visual thoughts (e.g., MindOmni (Xiao et al.,  
 1075 2025b), T2I-R1 (Jiang et al., 2025a)) or apply reinforcement learning to refine outputs directly.  
 1076 While these approaches yield moderate improvements, their reliance on text-only reasoning or post-  
 1077 hoc optimization limits their ability to capture fine-grained visual semantics. In contrast, our method  
 1078 consistently achieves the best results across GenEval (Ghosh et al., 2023) and WISE Niu et al.  
 1079

<sup>1</sup><https://github.com/djghosh13/geneval>

1080  
1081 Table 8: Performance comparison of different methods on leveraging understanding for the text-to-  
1082 image generation task.

Model	GenEval (Position)	WISE(Cultural)	WISE(Biology)
MindOmni (Xiao et al., 2025b)	0.59	0.60	0.56
X-Omni (Geng et al., 2025b)	-	0.71	0.48
T2I-R1 (Jiang et al., 2025a)	0.71	0.56	0.54
Ours	<b>0.80</b>	<b>0.79</b>	<b>0.68</b>

1083  
1084 (2025), demonstrating that interleaving understanding with generation enables more accurate spa-  
1085 tial reasoning and semantically faithful image synthesis.

1086  
1087 **Visualization on Understanding Tasks.** We further provide qualitative visualizations of the  
1088 model’s performance on various understanding tasks. For instance, in mathematical reasoning prob-  
1089 lems (Fig. 11), the model is first able to analyze the question, then drafts intermediate visual sketches  
1090 analogous to human scratch work, and finally derives the correct solution. Similarly, in more abstract  
1091 scenarios, such as tangram analysis (Fig. 12), the model successfully generates accurate interme-  
1092 diate visualizations that facilitate the correct interpretation, significantly outperforming approaches  
1093 like BAGEL, which also employ self-thinking strategies.

1094  
1095 **Visualization on Generation Tasks.** We also provide additional qualitative results on generation  
1096 tasks. Compared with alternative reasoning-based methods, our model demonstrates superior per-  
1097 formance on prompts that require commonsense reasoning. For example, as shown in Fig. 13, base-  
1098 line models tend to generate kangaroos sitting directly on the ground, whereas our model correctly  
1099 depicts the kangaroo using its tail for support, which is more consistent with reality. In creative syn-  
1100 thesis scenarios, such as the counterfactual prompt “a plant watering a gardener” our model is still  
1101 able to produce outputs that faithfully capture the user’s intent. Furthermore, Fig. 14 contrasts gen-  
1102 erations produced by the base model with and without our thinking mechanism, where our approach  
1103 consistently yields higher-quality images that better align with user instructions.

1104  
1105 **Visualization on Editing Tasks.** We further compare our approach with BAGEL (Deng et al.,  
1106 2025) on diverse image editing tasks, both with and without the “think” process, as shown in Fig. 15.  
1107 The results indicate that our method consistently produces more faithful and semantically aligned  
1108 edits. For instance, when asked to infer the missing element in a visual pattern, BAGEL either fails  
1109 to capture the correct reasoning or generates unrelated content, whereas our model accurately iden-  
1110 tifies the missing orange. In procedural editing tasks such as heating corn kernels until they pop,  
1111 our approach generates realistic popcorn, while BAGEL outputs less plausible textures. Similarly,  
1112 for biological transformations (e.g., albinism in corals) and structural corrections (e.g., fixing un-  
1113 reasonable parts of a bicycle), our method yields coherent results that respect both semantic intent  
1114 and visual fidelity. These comparisons highlight the strength of our interleaved thinking process in  
1115 producing edits that are both accurate and visually credible.

1116  
1117 **Challenging Cases.** Fig. 16 and 17 present several challenging examples where our model still  
1118 exhibits limitations. On the understanding side, the model may occasionally over-draft on simple  
1119 queries. For very basic attribute questions, the introduction of additional reasoning steps can lead  
1120 to unnecessary elaboration and, in some cases, incorrect predictions. Similarly, although the model  
1121 shows improvements on visually grounded mathematical tasks, such as MathVista (Lu et al., 2023),  
1122 it still struggles on problems that require strict symbolic reasoning, where applying AD-Loop may  
1123 introduce deviations from the optimal logical path.

1124  
1125 **On the generation side,** the model continues to face difficulties in rendering fine-grained visual  
1126 details, including scene text, subtle body parts (e.g., fingers), and numerically precise elements.  
1127 A plausible explanation is that the compact latent visual representations may under-encode such  
1128 fine details. We anticipate that incorporating more detailed feedback into the Stage-2 RL reward,  
1129 encouraging the model to attend to fine-grained latent visual thoughts, could further mitigate these  
1130 issues in future work.

1134

1135

1136

## I THINKING STRATEGIES COMPARISON WITH EXISTING VLMS VIA VISUAL INFORMATION

1139

1140 Compared to existing “alternation-based” or “integration-based” visual information argumented  
 1141 multimodal reasoning works (Zhang et al., 2025a; Zheng et al., 2025), the proposed AD-Loop is  
 1142 fundamentally different across several objective and design perspectives.

1143 **(1) Objective: achieving genuine synergy between understanding and generation.** The goal  
 1144 of AD-Loop is not merely to alternate modalities but to enable a mutual strengthening between  
 1145 the model’s understanding and generation capabilities. The entire framework, both in design and  
 1146 training, is centered around this objective, which distinguishes it from prior works that only combine  
 1147 modalities without enabling reciprocal influence.

1148 **(2) Modeling perspective: internal, bidirectional, and iterative interaction rather than passive  
 1149 alternation.** In AD-Loop, both textual and visual information are derived from the model’s internal  
 1150 reasoning without relying on external tools for visual integration. The process is not a fixed alter-  
 1151 nation but a dynamic switching between analytic (textual) and synthetic (visual-drafting) modes,  
 1152 allowing iterative refinement of reasoning. Moreover, unlike prior methods that use explicit images  
 1153 or pixel-level drafts, AD-Loop introduces latent visual thoughts, which serve as concept-aligned  
 1154 abstractions. These representations compress visual regions into semantic clusters and disentangle  
 1155 relevant from irrelevant information, enabling visual reasoning without pixel-level noise or compu-  
 1156 tational overhead.

1157 **(3) Learning perspective: adaptive rather than mandatory or heuristic alternation.** As dis-  
 1158 cussed in Appendix E, existing synergistic or alternating approaches typically rely on fixed sched-  
 1159 ules or heuristics. In contrast, AD-Loop is explicitly adaptive: our hybrid learning scheme with  
 1160 Stage-2 group-relative RL allows the model to intelligently decide when to invoke understanding or  
 1161 drafting. This results in an organic integration of the two capabilities and is key to realizing genuine  
 1162 synergy, rather than superficial alternation.

1163

[Question]: were any of the objects in the initial frame that you can still see in the second frame moved from their original positions?

1164

[Correct answer]: chair was moved left and towards from the camera

1165

[BAGEL w/ Think] To determine if any objects were moved from their original positions between the two frames, we need to carefully compare the positions of the objects in both images.\n1. \*\*Identify Key Objects:\*\*\n - \*\*Chair:\*\* The chair is a prominent object in both frames.\n - \*\*Guitars:\*\* The guitars are also noticeable in both frames.\n - \*\*Desk:\*\* The desk and its contents are another key element.\n2. \*\*Compare Positions:\*\*\n - \*\*Chair:\*\* In the first frame, the chair is positioned towards the left side of the image, closer to the camera. In the second frame, the chair is moved to the right side of the image, further away from the camera.\n - \*\*Guitars:\*\* The guitars appear to be in the same general area in both frames, but their exact positions might have changed slightly.\n - \*\*Desk:\*\* The desk and its contents seem to have shifted slightly, but the main items (like the box and the computer tower) are still in the same general area.\n3. \*\*Conclusion:\*\*\n - The chair has been moved from its original position. It was moved to the right and away from the camera.\nTherefore, the correct answer is: 'chair was moved right and away from the camera'.



1170

[Janus-Pro] Step 1: Observe the initial frame and the second frame.\n- In the initial frame, there is a chair on the left side of the room.\n- In the second frame, the chair is still on the left side but slightly closer to the camera.\nStep 2: Analyze the movement of the objects.\n- The chair appears to have moved slightly towards the camera, but it is still in the same general position relative to the other objects.\nStep 3: Compare the options.\n- Option 1 suggests the chair was moved left and towards the camera.\n- Option 2 suggests the chair was moved right and away from the camera.\nStep 4: Determine the correct answer.\n- The chair's position in the second frame does not match the description of moving left or right and towards or away from the camera.\nFinal answer: **chair was not moved from its original position.**

1171

[Harmon] To determine if any objects were moved, we need to compare the positions of the items in both frames. In the initial frame (left side of the image), there is a chair that appears to be facing towards the camera and has its legs spread out for stability. In the second frame (right side of the image), the same chair seems to have been repositioned with its legs now closer together and angled slightly away from the camera.\nIn the second frame, the chair's position relative to the camera has changed: it was initially on the left but is now more centered in the frame. This indicates that the chair was indeed moved from its original position.\nTherefore, the correct answer is:\n**chair was moved right and away from the camera**

1172

[Ours] To decide whether an object moved between frames, we compare relative positions of key items. In the initial Frame 1 the rolling chair sits beside the guitar case on the floor [V-T], with several guitars nearby and the table close to its right.\nIn the Frame 2, only part of the chair is visible and it appears behind the guitar case [V-T], indicating a shift toward the camera and leftward relative to the original layout; it's also farther from the table than before.\nFinal answer:\n**chair was moved left and towards from the camera**

1173

Figure 10: Qualitative comparison across baseline unified MLLMs on multimodal understanding. Correct answers are marked in green, errors in red, and [V-T] denotes latent visual thoughts.

1180

1181

1182

1183

1184

1185

1186

1187

1188  
 1189 [Question]: AB is the diameter of  $\odot O$ . Point C is a point on AB. Point D is on  $\odot O$ ,  $AD = AC$ . Connect DC and extend it to intersect  $\odot O$  at point E. Connect AD, AE, OE. If  $\angle BAD = 30^\circ$ , the degree measure of  $\angle COE$  is ()  
 1190 [Option]: A:  $30^\circ$  B:  $35^\circ$  C:  $40^\circ$  D:  $45^\circ$   
 1191 [Correct Answer]: A  
 1192 [BAGEL w/o Think] Final answer: B  
 1193 [BAGEL w/ Think] To find the measure of  $\angle COE$ , we can use properties of circles and angles. Let's break down the problem step by step.  
 1194 1. \*\*Identify the given information:\*\* - AB is the diameter of circle  $\odot O$ . - Point C is on AB. - Point D is on circle  $\odot O$  such that  $AD = AC$ . -  $AD = AC$  implies that triangle ADC is isosceles with  $AD = AC$ . -  $\angle BAD = 30^\circ$ .  
 1195 2. \*\*Understand the properties of circles:\*\* - The angle subtended by a diameter in a circle is a right angle ( $90^\circ$ ). Therefore,  $\angle ADB = 90^\circ$ . - Since  $AD = AC$ , triangle ADC is isosceles, so  $\angle ADC = \angle ACD$ .  
 1196 3. \*\*Calculate  $\angle ADC$ :\*\* - Since  $\angle BAD = 30^\circ$  and  $\angle ADB = 90^\circ$ ,  $\angle ABD = 60^\circ$  (since the angles in a triangle sum to  $180^\circ$ ). - In triangle ADC,  $\angle ADC = \angle ACD$ . Let's denote each of these angles as  $x$ . - Therefore,  $2x + 30^\circ = 180^\circ$  (since the sum of angles in triangle ADC is  $180^\circ$ ). - Solving for  $x$ :  $2x = 150^\circ$ , so  $x = 75^\circ$ . - Thus,  $\angle ADC = 75^\circ$ .  
 1197 4. \*\*Determine  $\angle COE$ :\*\* - Since  $\angle ADC = 75^\circ$ , and  $\angle ADC$  is an inscribed angle that subtends arc AC, the measure of arc AC is  $2 * 75^\circ = 150^\circ$ . - The central angle  $\angle COE$  that subtends the same arc AC is equal to the measure of the arc, which is  $150^\circ$ .  
 1198 5. \*\*Calculate the remaining angle:\*\* - Since AB is the diameter, the total angle around point O is  $360^\circ$ . - The angle  $\angle AOE$  is the remaining angle that completes the circle. -  $\angle AOE = 360^\circ - 150^\circ = 210^\circ$ .  
 1199 6. \*\*Determine the correct option:\*\* - The question asks for the measure of  $\angle COE$ , which we found to be  $150^\circ$ . - However, the options provided are A:  $30^\circ$ , B:  $35^\circ$ , C:  $40^\circ$ , D:  $45^\circ$ . - None of the options match  $150^\circ$ , which suggests there might be a misunderstanding in the problem setup or options provided. However, based on the given options and the logical steps, the closest and most reasonable answer would be D:  $45^\circ$ , as it is the only option that could be a part of the correct calculation process.  
 1200 Final answer: D.

1201 [Janus-Pro] Step 1: Identify the key elements in the problem. -  $\odot O$  is the circle. - AB is the diameter of  $\odot O$ . - Point C is a point on AB. - Point D is on  $\odot O$ ,  $AD = AC$ . - Connect DC and extend it to intersect  $\odot O$  at point E. - Connect AD, AE, OE. -  $\angle BAD = 30^\circ$ .  
 1202 Step 2: Analyze the given information. - Since  $AD = AC$ , triangle ACD is isosceles. -  $\angle CAD = 60^\circ$ . - Since  $AD = AC$ ,  $\angle ADE = 60^\circ$ . -  $\angle ADE + \angle COE = 180^\circ$ .  
 1203 Step 3: Calculate  $\angle COE$ . -  $\angle ADE = 60^\circ$ . -  $\angle ADE + \angle COE = 180^\circ$ . -  $60^\circ + \angle COE = 180^\circ$ . -  $\angle COE = 180^\circ - 60^\circ = 120^\circ$ .  
 1204 Step 4: Determine the correct answer. - The degree measure of  $\angle COE$  is  $120^\circ$ .  
 1205 Final answer: 120°

1206 [Ours] <think>To answer this question, the image is depicted below: [V-T] Let's think step by step: 1) Let O(0,0), A(-1,0), B(1,0). Then AB is the diameter. 2)  $\angle BAD = 30^\circ \Rightarrow$  arc BD =  $60^\circ$ , so D =  $(1/2, \sqrt{3}/2)$ . 3) Triangle ABD is a  $30^\circ-60^\circ-90^\circ$  right triangle, hence AD =  $\sqrt{3}$ . Given AC = AD and C lies on AB, we get C =  $(\sqrt{3}/2, 0)$ . 4) Line DC intersects the circle again at E =  $(\sqrt{3}/2, -1/2)$ . 5) Vector OC lies along the x-axis (0°), OE has angle  $-30^\circ$ . Therefore,  $\angle COE = 30^\circ$ . </think><answer>Final answer: A</answer>

1207 Figure 11: Understanding Comparison across UVLMs on the math question answering task. Correct answers are marked in green, errors in red, and [V-T] denotes latent visual thoughts.

1208 [Question]: What does the tangram in the image show a closer likeness to?  
 1209 [Option]: A: a dog's head B: someone flexin C: mailbox D: woman  
 1210 [Correct Answer]: C  
 1211 [BAGEL w/o think] Final answer: B  
 1212 [BAGEL w/ think] To determine what the tangram in the image resembles, we need to carefully analyze the shapes and their arrangement.  
 1213 1. \*\*Identify the Shapes:\*\* The tangram consists of seven pieces: five triangles, one square, and one parallelogram. These pieces are arranged to form a larger shape.  
 1214 2. \*\*Analyze the Arrangement:\*\* - The largest triangle is positioned on the left side.  
 1215 - The square is located at the top center.  
 1216 - The parallelogram is at the bottom center.  
 1217 - The remaining triangles are arranged around these central shapes.  
 1218 3. \*\*Compare with Options:\*\*  
 1219 - \*\*Option A (a dog's head):\*\* The arrangement does not resemble a dog's head. The shapes do not form the typical features of a dog's face.  
 1220 - \*\*Option B (someone flexin):\*\* The shapes do not form a person flexing their muscles. The arrangement lacks the typical features of a human figure.  
 1221 - \*\*Option C (mailbox):\*\* The arrangement does not resemble a mailbox. The shapes do not form the typical rectangular and triangular features of a mailbox.  
 1222 - \*\*Option D (woman):\*\* The arrangement of the shapes does resemble a woman. The largest triangle can be seen as the head, the square as the upper body, and the parallelogram as the lower body. The other triangles can be interpreted as arms and legs.  
 1223 4. \*\*Conclusion:\*\* Based on the analysis, the tangram in the image shows a closer likeness to a woman.  
 1224 Final answer: D.

1225 [Janus-Pro] To answer this question, let's analyze the tangram shapes and their arrangement in the image.  
 1226 1. The image shows a set of geometric shapes that are typically used in tangram puzzles.  
 1227 2. The shapes are arranged in a way that resembles a dog's head.  
 1228 3. The yellow shapes form the outline of the dog's head, with the pink shape representing the body.  
 1229 4. The cyan shape is positioned to resemble the ears or side of the dog.  
 1230 5. Based on this analysis, the tangram in the image shows a closer likeness to a dog's head.  
 1231 Final answer: A

1232 [MindOmni] To determine what the tangram in the image shows a closer likeness to, let's analyze the shapes and their arrangement.  
 1233 1. The image consists of several geometric shapes: triangles, a square, and a parallelogram.  
 1234 2. The shapes are arranged in a way that suggests a recognizable form.  
 1235 3. The large yellow triangle at the top left resembles an ear.  
 1236 4. The two yellow squares and the yellow triangle below them form a shape that looks like a face with a mouth.  
 1237 5. The pink shapes at the bottom resemble a body.  
 1238 Considering these observations, the overall arrangement of the shapes most closely resembles a face with a mouth, which is characteristic of a dog's head.  
 1239 Final answer: A. a dog's head

1240 [Ours] <think>The initial step involves examining regions of the image that are marked by a specific color. As illustrated, these regions resemble the shape of a pole [V-T]. Next, attention shifts to segments highlighted by another color. As shown, these areas take on a form similar to a box [V-T]. Finally, we focus on the portions rendered in the remaining colors. These sections, as depicted, bear a likeness to a flag [V-T]. Taken together, the overall tangram arrangement most closely aligns with the shape of a mailbox [V-T].</think><answer>Final answer: C</answer>

1241 Figure 12: Understanding Comparison across UVLMs on multimodal content reasoning task. Correct answers are marked in green, errors in red, and [V-T] denotes latent visual thoughts.

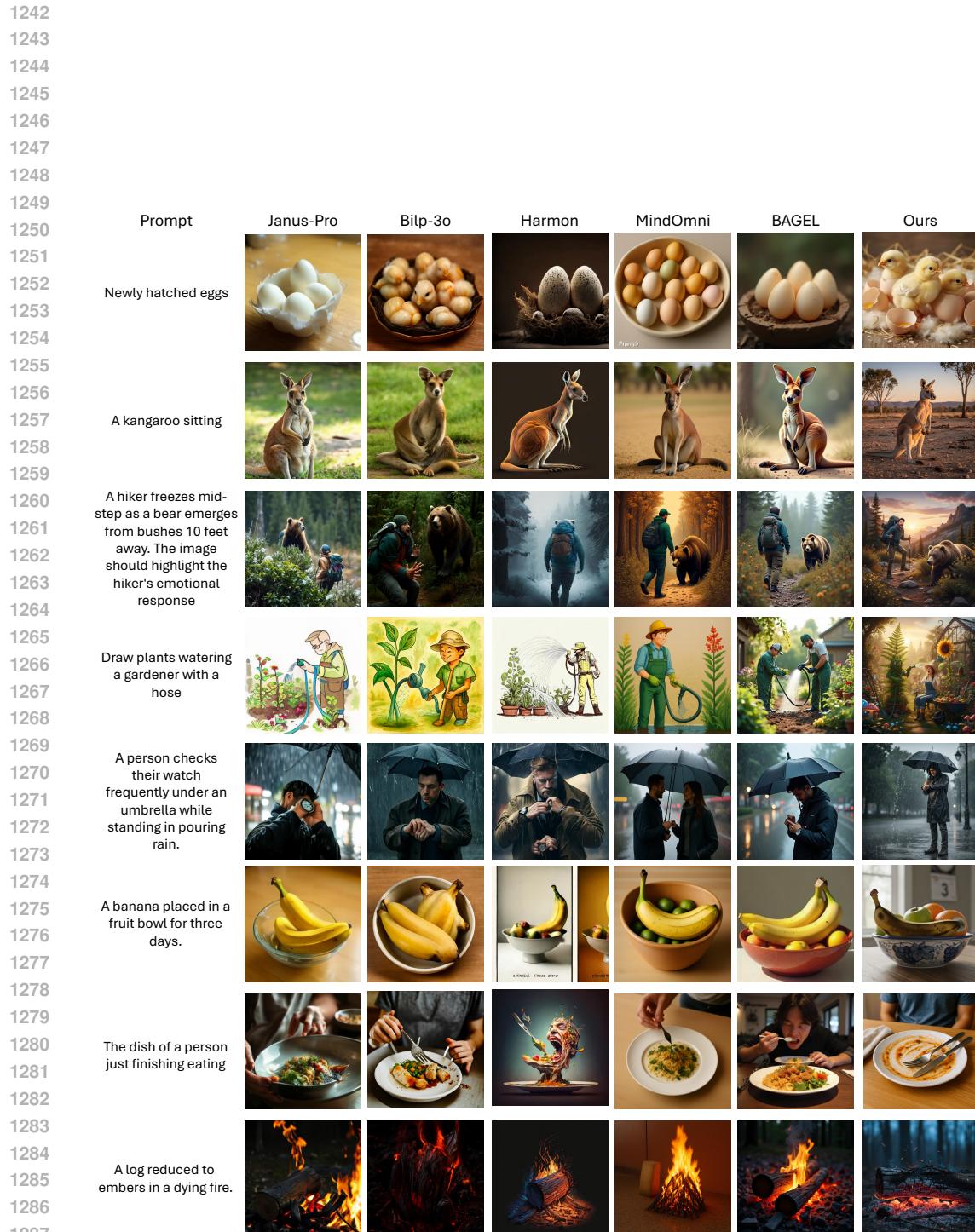


Figure 13: Generation Comparison across UVLMs.

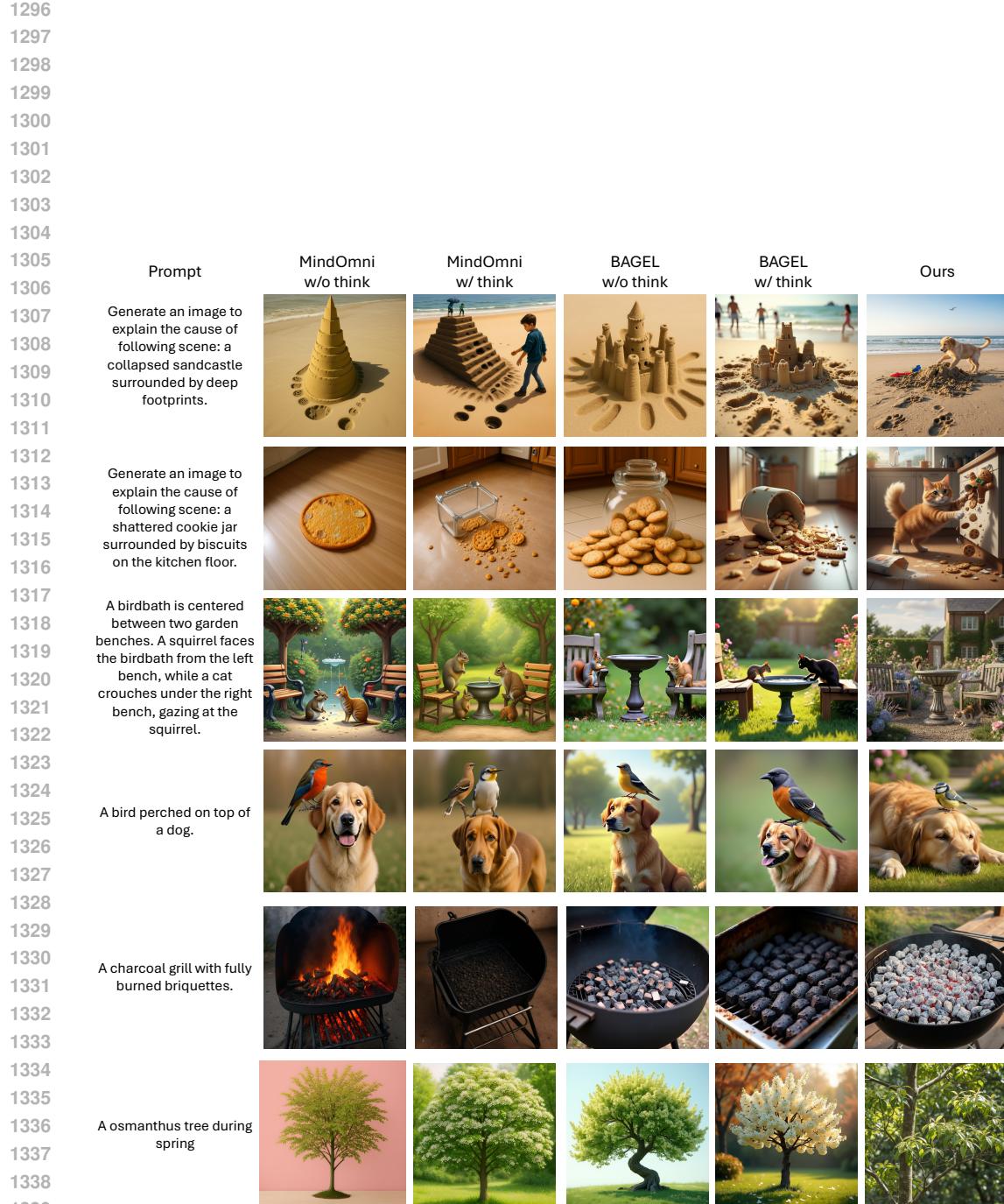


Figure 14: Generation Comparison on MindOmni (Xiao et al., 2025b) and BAGEL (Deng et al., 2025) with/without Think process. Our proposed method achieves more accurate generation results.

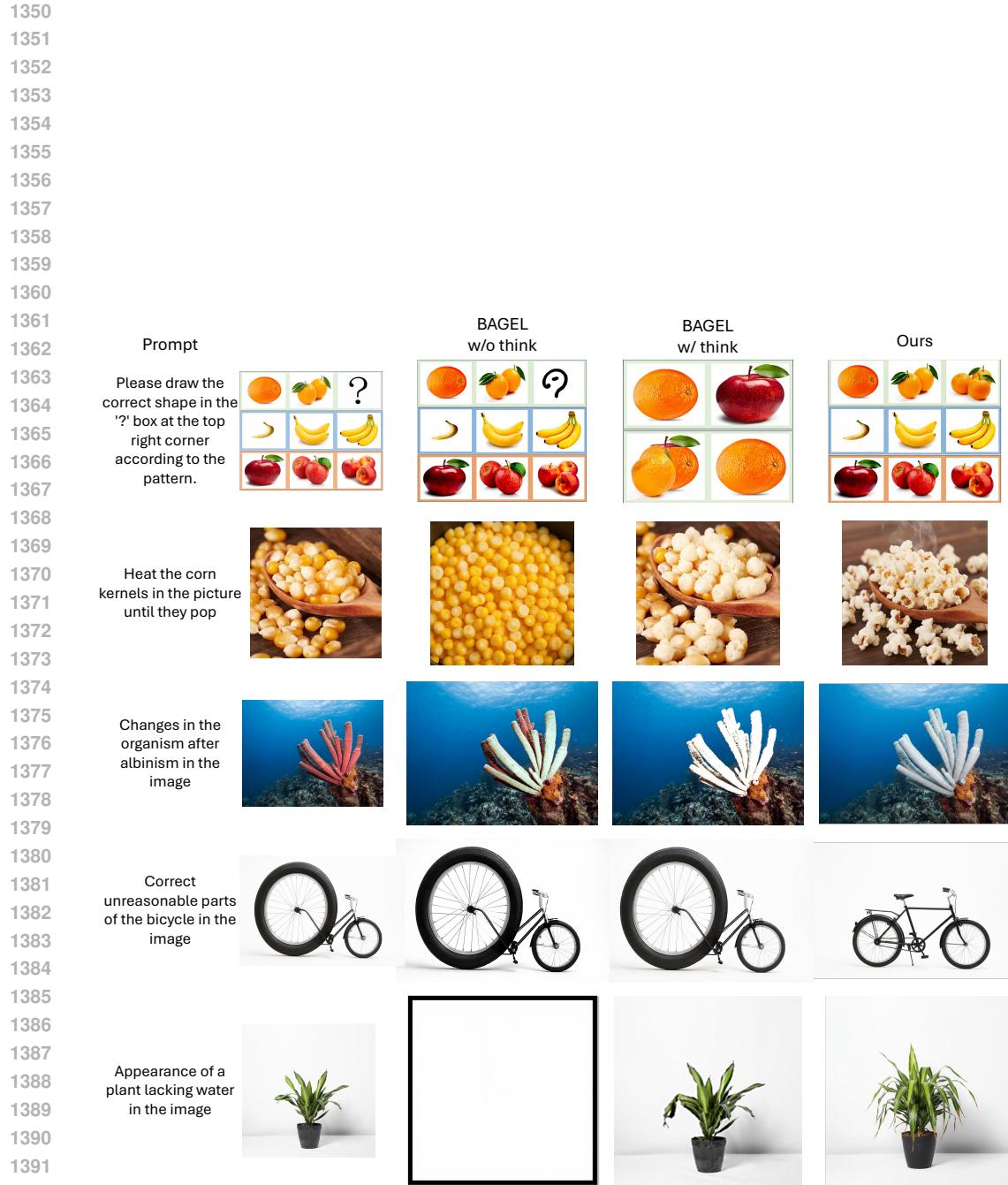


Figure 15: Editing Comparison on BAGEL (Deng et al., 2025) with/without think process.

1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457

**[Question]** What is the most prominent feature in the image?  
**[Option]** A: The skyline, B: The golf course, C: The trees, D: The person

**[BAGEL w/o Think]** Final answer: C.

**[Ours]** The image is taken from a perspective that places the golf course in the foreground [V-T], with the trees and skyline serving as secondary elements [V-T]. The golf course is the central focus, with the green and the golfer being the main subjects [V-T]. The trees frame the scene and add depth, but they are not the primary focus [V-T]. The skyline, while visible, is in the background and does not dominate the image [V-T]. Therefore, the most prominent feature is the golf course. **Final answer: B.**



**[Question]** Find the greatest common divisor of \$3339\$, \$2961\$, and \$1491\$.

**[BAGEL w/ Think]** Let's compute it step-by-step using the Euclidean algorithm.  
1. Find  $\text{gcd}(3339, 2961)$ .  
 $3339 - 2961 = 378$ . So now compute  $\text{gcd}(2961, 378)$ .  
 $2961 \div 378 = 7 \times 378 = 2646$ .  
 $2961 - 2646 = 315$ . Now  $\text{gcd}(378, 315)$ .  
 $378 - 315 = 63$ .  
 $\text{gcd}(315, 63) = 63$ . So,  $\text{gcd}(3339, 2961) = 63$ .  
2. Now compute  $\text{gcd}(63, 1491)$ .  
 $1491 \div 63 = 23 \times 63 = 1449$ .  
 $1491 - 1449 = 42$ . Now  $\text{gcd}(63, 42)$ .  
 $63 - 42 = 21$ . So,  $\text{gcd}(63, 1491) = 21$ .  
**Final Answer:** 21

**[Ours]** To find the greatest common divisor, we first determine the greatest common divisor of the first two numbers:  $\text{gcd}(3339, 2961)$ .  
 $3339 = 1 \cdot 2961 + 378$ .  
 $2961 = 7 \cdot 378 + 318$ .  
 $378 = 1 \cdot 318 + 60$ .  
 $318 = 5 \cdot 60 + 18$ .  
 $60 = 3 \cdot 18 + 6$ .  
 $18 = 3 \cdot 6 + 0$ .  
Therefore,  $\text{gcd}(3339, 2961) = 6$ .  
Next, we find the greatest common divisor of the result from the previous step (6) [V-T] and the third number (1491):  $\text{gcd}(6, 1491)$ .  
 $1491 = 248 \cdot 6 + 3$ .  
 $1491 = 248 \cdot 6 + 3$ .  
Therefore,  $\text{gcd}(6, 1491) = 3$ .  
Thus, the greatest common divisor of all three numbers is 3.  
**Final answer: 3.**

Figure 16: Failure cases illustrating over-drafting on simple tasks, as well as challenges on problems that require strictly logical mathematical reasoning.

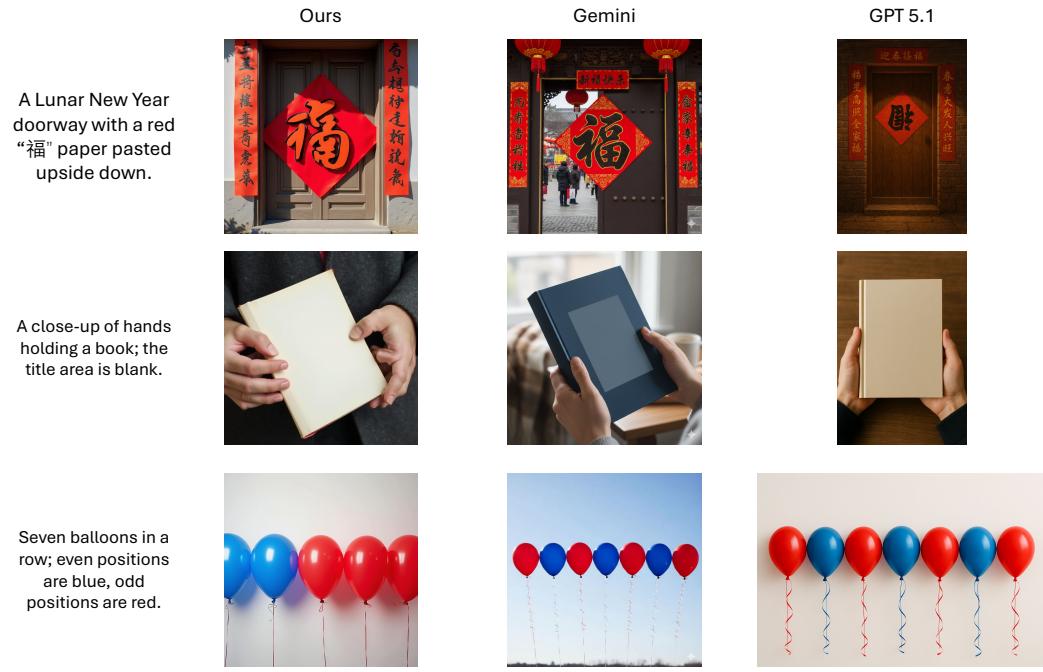


Figure 17: Failure cases showing that the proposed method still struggles with rendering scene text, as well as fine-grained body details and numerically precise content.