



Thinking with Generated Images

Ethan Chern^{1,4*}† Zhulin Hu^{1,4*} Steffi Chern^{4*} Siqi Kou¹ Jiadi Su^{3,4} Yan Ma^{3,4}
 Zhijie Deng¹ Pengfei Liu^{1,2,4‡}
¹Shanghai Jiao Tong University ²SII
³Fudan University ⁴Generative AI Research Lab (GAIR)

Abstract

We present *Thinking with Generated Images*, a novel paradigm that fundamentally transforms how large multimodal models (LMMs) engage with visual reasoning by enabling them to *natively* think across text and vision modalities through *spontaneous* generation of intermediate visual thinking steps. Current visual reasoning with LMMs is constrained to either processing fixed user-provided images or reasoning solely through text-based chain-of-thought (CoT). *Thinking with Generated Images* unlocks a new dimension of cognitive capability where models can actively construct intermediate visual thoughts, critique their own visual hypotheses, and refine them as integral components of their reasoning process. To implement *Thinking with Generated Images*, we introduce the *native long-multimodal thought process*, which enables unified LMMs to seamlessly generate intermediate visual thoughts, establish visual subgoals, and iteratively critique their visual hypotheses within a single, coherent reasoning process. This approach naturally performs test-time scaling across modalities. We demonstrate the effectiveness of our approach through two complementary mechanisms: (1) vision generation with intermediate visual subgoals, where models decompose complex visual tasks into manageable components that are generated and integrated progressively, and (2) vision generation with self-critique, where models generate an initial visual hypothesis, analyze its shortcomings through textual reasoning, and produce refined outputs based on their own critiques. Our experiments on vision generation benchmarks show substantial improvements over baseline approaches, with our models achieving up to 50% (from 38% to 57%) relative improvement in handling complex multi-object scenarios. Beyond these immediate performance gains, *Thinking with Generated Images* opens transformative possibilities for AI systems across diverse real-world applications. From biochemists exploring novel protein structures, and architects iterating on spatial designs, to forensic analysts reconstructing crime scenes, and basketball players envisioning strategic plays, our approach enables AI models to engage in the kind of visual imagination and iterative refinement that characterizes human creative, analytical, and strategic thinking. This work establishes a foundation for future research in multimodal cognition and complex visual reasoning tasks. We release our open-source suite at <https://github.com/GAIR-NLP/thinking-with-generated-images>.

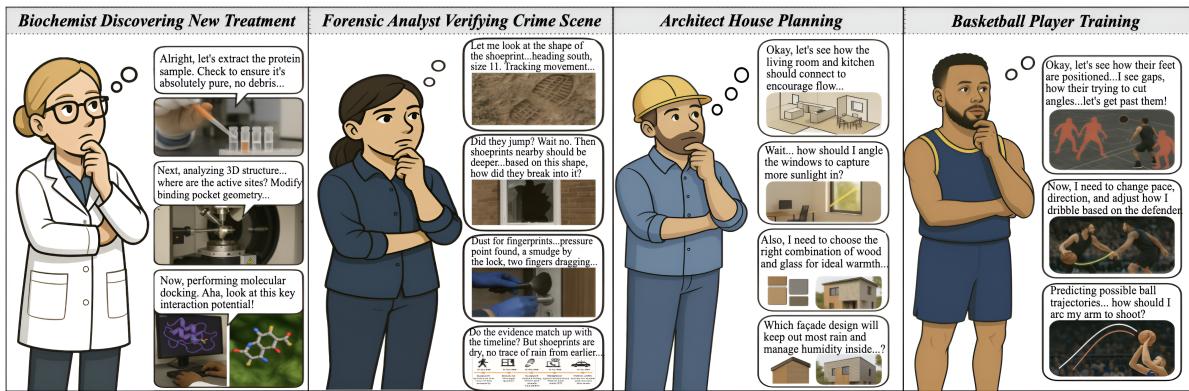


Figure 1: Real-world tasks that require *Thinking with Generated Images*. These tasks often require visual foresight and imagination, which text-based thought alone cannot fully accomplish.

*Equal Contribution.

†Partial work done at Bytedance Seed.

‡Corresponding author.

1 Introduction

“Thorough thought leads to victory; insufficient thought leads to defeat.”

—Sun Tzu

Human goal-oriented cognition (Simon and Newell, 1971; Newell, 2014; Xia et al., 2025) can be characterized as a goal-directed *search* through a *problem space*¹—a landscape of intermediate states connected by operators (action or thinking steps). Large language models (LLMs) exhibit this same pattern: when prompted to write a chain-of-thought (CoT) (Wei et al., 2022), they traverse intermediate states, with performance improving with additional inference compute—a.k.a. test-time scaling (Brown et al., 2024; Snell et al., 2024). Yet, the text-only CoT process of LLMs captures only a partial view of cognitive search.

Crucially, human cognition is natively *multimodal*. Biochemists explore protein structures to discover new treatment approaches; forensic analysts verify crime scene reconstructions to establish evidence connections; architects revise spaces and light patterns to optimize building designs. Visual thinking creates unique combinations and novel connections between concepts (Hao et al., 2024; Barrault et al., 2024). This helps us glimpse new possibilities in our thinking—ideas and connections we wouldn’t have discovered through text reasoning alone.

An AI system that thinks or reasons only in the text modality is therefore constrained:² it cannot explore alternative therapeutic pathways as freely, verify criminal evidence at a glance, or revise design configurations in place³—capacities that underlie much of human creativity and analysis. Extending the thoughts of AI systems to multiple modalities unlocks new forms of interactivity: a model can generate not only intermediate text but also visual thoughts, decompose tasks into both text and visual subgoals, critique these intermediate steps, and iterate—ideally, all within a single, coherent thought process.

While standard large multimodal models (LMMs) or vision-language models (VLMs) can process visual information, these models are merely *seeing* images—processing them once during the forward pass rather than *thinking* with images more deeply. To address this limitation, previous works have explored leveraging multi-step or tool-augmented approaches (Zhang et al., 2023; Gupta and Kembhavi, 2023; Surís et al., 2023; Hu et al., 2024b; OpenAI, 2025; Su et al., 2025) to enable models to revisit or conduct simple operations on user-input images as intermediate steps toward solutions. However, these works on *Thinking with Images* only enable models to perform limited operations on images, constraining the thinking space of the model.

To establish a generalized, scalable model that can naturally think and reason across modalities while avoiding the limitations mentioned above, we propose the idea of “*Thinking with Generated Images*”: An AI model (system) that *spontaneously* thinks and reasons across (e.g., text and vision) modalities. Rather than entirely relying on user-supplied images, which would constrain its thinking, the model proactively produces its own visual steps or subgoals toward solving problems when needed.

Recent works have attempted to take the first steps toward *Thinking with Generated Images*, using either agentic (Guo et al., 2025; Fang et al., 2025; Ni et al., 2024) or single-model approaches (Li et al., 2025). We argue that improving AI system performance via agentic approaches is parallel to single-model approaches, but integrating various capabilities into a single, end-to-end generalist model represents the continued evolution toward superintelligence (Bubeck et al., 2023). More recent work (Li et al., 2025) attempted to leverage unified LMMs (models that can generate multimodal content simultaneously) to solve spatial reasoning tasks such as maze navigation. While taking a first step, their approach is limited to maze problem-solving, which can be adequately solvable without visual intermediate steps (Dao and Vu, 2025). We argue that *Thinking with Generated Images* is best applied to complex tasks (as shown in Fig. 1), where their intermediate steps *cannot* be adequately represented in textual form alone. Thus, to demonstrate the potential of *Thinking with Generated Images*, in this paper, we focus on vision (image) generation tasks.

To effectively realize *Thinking with Generated Images* with an end-to-end, single-model approach and demonstrate its strong potential on vision generation tasks, we introduce the **native long-multimodal thought process** on unified autoregressive LMMs (Team, 2024a; Chern et al., 2024). The native long-multimodal thought process comprises interleaved multimodal tokens—words or subwords for text, patches for vision, frames for audio, and

¹A problem space is the basic unit of goal-oriented symbolic activity (Newell, 2014). It consists of states (the possible configurations of the system), operators (actions that transform one state into another), an initial state, one or more goal states, and optional path constraints that restrict legal moves. Take the *Tower of Hanoi* as an example: a state is any arrangement of the disks on three pegs; the operators are “move the top disk from one peg to another” and “recognizing the arrangement of the disks”; the initial state has all disks stacked on peg1; the goal state has them stacked on peg3; and the path constraint forbids placing a larger disk atop a smaller one.

²We argue that vision-language models that are only capable of generating text are still reasoning only in the text modality—they cannot construct intermediate visual thinking.

³Although visual content can sometimes be encoded into pure text descriptions, this process is usually inevitably lossy, except for certain tasks like maze problem solving where states and steps can be fully encoded with code or text.

2. Thinking with Generated Images

other domain-specific representations⁴. The native long-multimodal thought process not only naturally enable models to spontaneously generate images in the thinking process, but also natively perform test-time scaling for better model capabilities.

In summary, our contributions are as follows:

- We propose the concept of *Thinking with Generated Images*, demonstrating how a single AI model can learn to spontaneously think and reason across text and vision modalities.
- We explore test-time scaling for native unified LMMs, shedding light on future research directions.
- We introduce the *native long-multimodal thought process* to instantiate *Thinking with Generated Images*, enabling models to think and reason across modalities, and naturally performs test-time scaling. We demonstrate that the native long-multimodal thought process is beneficial for vision generation and can potentially be applied to more diverse and complex tasks with the emergence of stronger base models.

2 Thinking with Generated Images

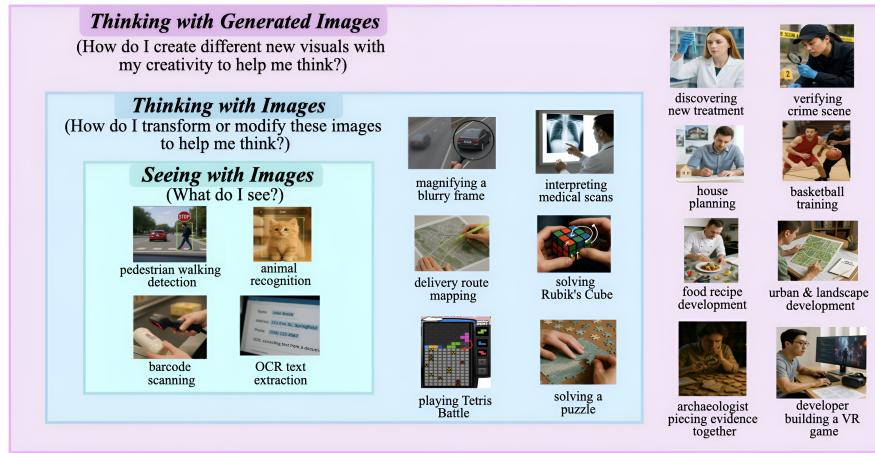


Figure 2: Examples of “seeing with images” vs. “thinking with images” vs. “thinking with generated images.”

2.1 “Seeing” vs. “Thinking” with Images

Standard LMMs or VLMs only *see* an image once. During a single forward pass, the image is converted into visual tokens, and the model then autoregressively generates text tokens conditioned on those visual tokens. The thinking process of AI models happens *entirely* in the text modality; the image acts as a fixed prior or conditioning context for a *text-only* chain-of-thought.

To integrate images into the thought process, recent work leverages established multi-step pipelines (Zhang et al., 2023), multi-agent systems (Yang et al., 2023), or external tools such as code executors, OCR, or image manipulators (Gupta and Kembhavi, 2023; Surís et al., 2023; Hu et al., 2024b; OpenAI, 2025). These approaches let the model revisit the (user-given) images multiple times or generate transformed versions (e.g., cropped, rotated, zoomed) as intermediate steps toward a solution.

Thinking with images—rather than merely *seeing* them once—enables models to better tackle multi-step visual reasoning tasks such as spatial puzzles, complex visual question answering, and chart or diagram interpretation.

2.2 Thinking with Images v.s. Thinking with “Generated” Images

Thinking with images alone is not sufficient—models are still constrained when limited to reasoning with fixed user input images or slightly “transformed” versions of these images. The reasoning steps for tasks like design planning and physical-world reasoning would be lossy if they entirely relied on text and user-input image transformations.

To attempt the first steps toward thinking with generated images, previous works have primarily relied on multi-component agentic approaches (Guo et al., 2025; Fang et al., 2025; Ni et al., 2024) that require multiple model passes to achieve pipelined planning, (re)captioning, generation, critique, and reward components. Though demonstrating great potential, the complexity of these systems makes further generalizability and scaling to diverse scenarios complicated. Additionally, established test-time scaling and post-training scaling methods like long-form CoT and reinforcement learning must be engineered separately for each module, a burden that adds unwarranted

⁴We focus on text and vision modalities in this paper.

complexity. These limitations ultimately place a ceiling on the performance potential of such fragmented systems. More recent works (Li et al., 2025) have attempted single-model, end-to-end approaches. However, their approach limited its tasks to spatial reasoning such as maze problem-solving, where the maze itself and the intermediate states and steps are essentially discrete and can be entirely encoded with pure text descriptions (Dao and Vu, 2025). Such tasks are thus solvable without visual intermediate steps and don’t fully unlock the potential of *Thinking with Generated Images*.

To fully unlock the potential of *Thinking with Generated Images*, we focus on tasks that fundamentally require intermediate thoughts expressed in visual space. In this paper, we concentrate specifically on vision (image) generation, while demonstrating how our approach (the native long-multimodal thought process) enables models to inherently and deeply integrate vision into their reasoning processes. Unlike approaches that rely on fixed, user-inputted images, *Thinking with Generated Images* allows models to dynamically create and manipulate visual representations throughout their reasoning process. This fundamental shift unlocks new forms of interactivity and enables capabilities such as 3D modeling, visual foresight, design, and intuitive physics—domains where visual intermediate steps are not merely helpful but essential for effective problem-solving. A summary of the differences between our work and previous related works can be found in Fig. 3.

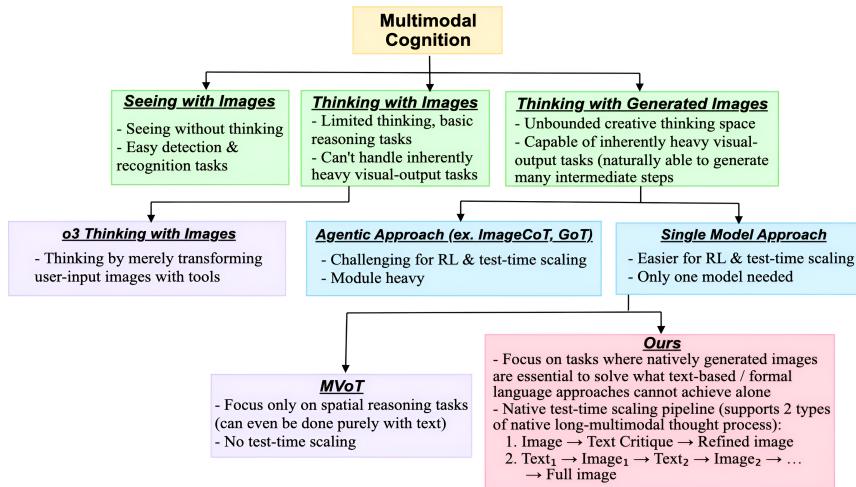


Figure 3: Comparison between related works on multimodal cognition.

2.3 Thinking with Generated Images with the Native Long-Multimodal Thought Process

We implement *Thinking with Generated Images* through the *native long-multimodal thought process* that leverages unified autoregressive LLMs via effective supervised fine-tuning on carefully curated high-quality samples. The native long-multimodal thought process offers multi-faceted benefits that transcend previous approaches: (1) enables models to “think” across modalities by generating modality-specific tokens with “natively” embedded multimodal generation capabilities via a single forward pass; (2) performs diverse multimodal tasks natively via the generative paradigm; (3) provides natural test-time scaling across modalities via the generated long-thought; and (4) facilitates future integration with post-training scaling techniques like reinforcement learning.

2.3.1 The Native Long-Multimodal Thought Process

The native long-multimodal thought process is a sequence of interleaved multimodal tokens. Formally, if we have m modalities with vocabulary spaces of $\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_m$, we define the unified vocabulary as $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \dots \cup \mathcal{V}_m$. Importantly, these vocabulary spaces need not be finite sets of discrete tokens (i.e., not constrained to autoregressive next-token-prediction LMMs); they could also be infinite sets of continuous tokens. A sequence of multimodal tokens is represented as $\mathbf{x} = (x_1, x_2, \dots, x_n)$ where $\forall i, x_i \in \mathcal{V}$.

Several length or token constraints may be introduced to ensure reasonable interaction within the multimodal thought process. For example, vision tokens might be required to appear in fixed-size blocks (e.g., exactly 1024 consecutive tokens for each image representation), and special delimiter tokens may be introduced to explicitly mark the boundaries and transitions between different modality spaces within the unified vocabulary \mathcal{V} .

2.3.2 Unified Auto-Regressive Next-Token-Prediction LMMs

Developing the native long-multimodal thought process on top of unified auto-regressive next-token-prediction LMMs is a natural approach, since the model can intuitively generate the thought process token-by-token. However, the main technical challenge in implementing the native long-multimodal thought process on unified LMMs stems primarily from the developmental stage of the unified LMMs field itself. Such implementation requires models to

3. Experiments

inherently support interleaved multimodal generation—a frontier that remains less developed. Many current open-source models claiming to be “unified vision understanding and generation LMMs” typically lack native interleaved multimodal generation support, instead supporting only text or image outputs in isolation. Recent successes in closed-source models like GPT-4o and Gemini have demonstrated the tremendous potential of interleaved multimodal generation in real-world applications.

In this paper, we employ Anole (Chern et al., 2024; Team, 2024a) as our base model for generating native long-multimodal thought process for test-time scaling on LMMs. Anole is a unified auto-regressive next-token-prediction LMMs that are trained to directly predict the next multimodal (text or image) token. Anole presents several compelling advantages over alternative LMMs. First, Anole is pre-trained and post-trained directly on interleaved text-image tokens, enabling inherent capabilities for interleaved generation of multimodal tokens. Second, Anole’s relatively efficient image representation encodes each image with 1024 tokens, making test-time scaling with the native long-multimodal thought process viable within reasonable inference compute budget. Third, Anole’s modeling strategy closely resembles the state-of-the-art LLMs, making it capable to leverage existing training and inference infrastructure of LLMs.

2.3.3 Methodology for Data Curation

We deliberately curate our set of supervised fine-tuning (SFT) dataset, which consists of diverse vision (image) generation prompts to ensure high-quality alignment (Team, 2024a). To elicit LMMs to perform the native long-multimodal thought process, we carefully design and construct the solution multimodal reasoning chain to elicit LMMs’ capabilities to spontaneously (1) critique their own generated visual steps and (2) generate intermediate visual subgoals. For more details, please refer to Section 3.

3 Experiments

To instantiate *Thinking with Generated Images* on vision generation tasks, we meticulously design two types of native long multimodal thought processes: *vision generation with intermediate visual subgoals* and *vision generation with self-critique*. We explore how these two processes can help LMMs to better conduct vision (image) generation via generative visual thinking.

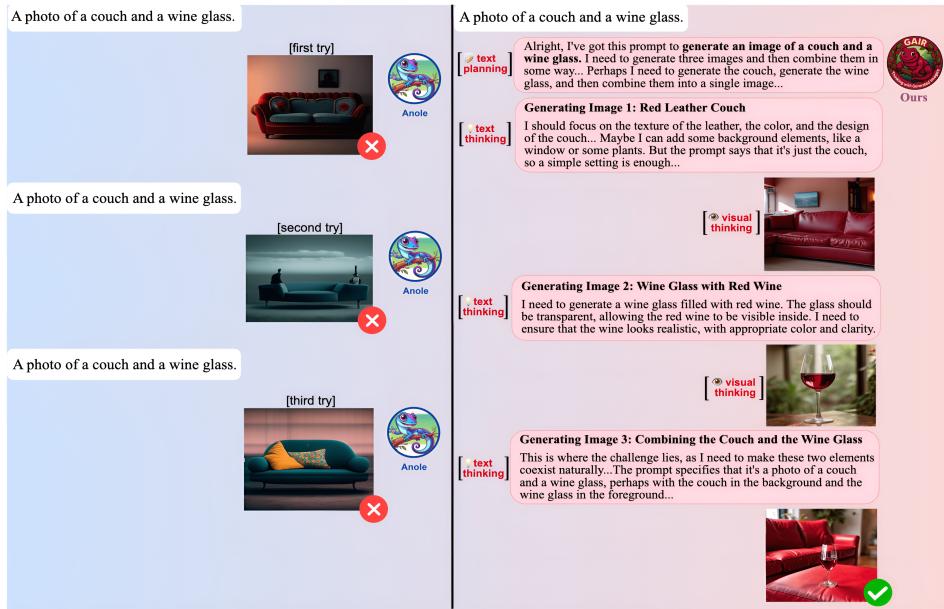


Figure 4: Demonstration of our long-multimodal thought process on GenEval.

3.1 Vision Generation with Intermediate Visual Subgoals

Through the *vision generation with intermediate visual subgoals* process, LMMs are able to plan and achieve visual subgoals as intermediate steps toward the final solution, as presented in equation 1. Specifically, in the image generation task, when presented with a complex prompt containing multiple objects, the model generates each object individually and then integrates them into the final output image to fulfill the requirements of the multi-object prompt. An example shown in Fig. 4 for demonstration.

3.2 Vision Generation with Self-Critique

$$[\text{Text Planning}] \rightarrow [\underbrace{\text{Visual SubGoal } i} \rightarrow [\text{Text Planning and Reflection}] \rightarrow [\text{Final Visual Output}]] \quad (1)$$

iteratively

3.2 Vision Generation with Self-Critique

Through the *vision generation with self-critique* process, LMMs are able to set a visual hypothesis, then perform critiquing and reflection the hypothesis, and finally make refinements toward better solution, as presented in equation 2. Specifically, in the image generation task, after generating an image based on a prompt, the model spontaneously reflects on the image to assess whether it meets the prompt's requirements. It then attempts to improve the image further, generating a higher-quality version that better aligns with the prompt by incorporating both the previously generated image and its own reflections. An example shown in Fig. 5 for demonstration.

$$[\text{Initial Visual Hypothesis}] \rightarrow [\text{Text Planning and Reflection}] \rightarrow [\text{Final Visual Output}] \quad (2)$$

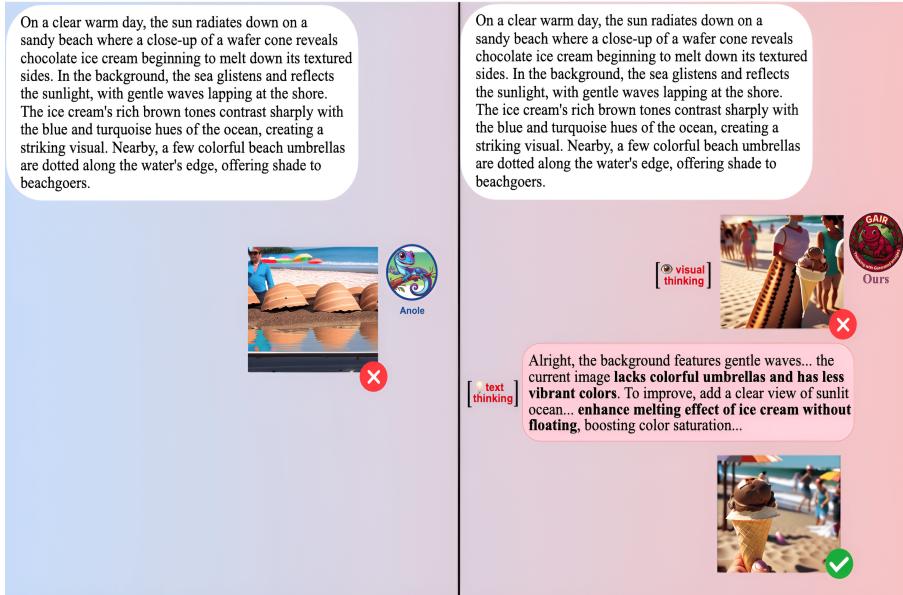


Figure 5: Demonstration of our long-multimodal thought process on DPG-Bench.

3.3 Dataset

We carefully designed our synthetic SFT data-construction workflow that produces high-quality training samples to enable trained models to generate two types of long-multimodal thought processes. Specifically, we developed a pipeline that constructs synthetic examples illustrating long-multimodal reasoning. Our pipeline can generate the two types of long multimodal-thought process mentioned above. After fine-tuning on this data, the LMMs become capable of producing end-to-end long-multimodal chains of thought.

Since there are no single off-the-shelf LMM that supports dynamic test-time scaling, traditional distillation techniques are not applicable. Our full dataset-curation pipeline is illustrated in Fig. 6.

We follow the below rules of thumb when collecting our synthetic data:

- **High-quality image-generation prompts:** To ensure diverse, high-quality outputs, we leverage several state-of-the-art models (including Deepseek-V3 ([DeepSeek-AI et al., 2024](#)), GPT-4o ([Hurst et al., 2024](#)), Claude3.7-Sonnet ([Anthropic, 2025](#)), and Qwen2.5-72B-Instruct ([Yang et al., 2024](#))) to brainstorm complex prompts and then apply rule-based filtering to guarantee both variety and proper formatting.
- **High-quality reflection-reasoning chains:** We use QVQ-72B-Preview ([Team, 2024b](#)) to construct the core components of critique and reflection, conditioned on images.
- **High-quality intermediate visual thoughts:** We employ Flux1-dev ([Labs, 2024](#)) (for vision generation with intermediate subgoals) or Anole-7b ([Chern et al., 2024](#)) (for vision generation with self-critique) to generate initial visual hypotheses from the prompts. We then refine or update these with Flux1-Redux ([Labs, 2024](#)) (a variant of Flux1-dev that accepts both image and text inputs).

3.4 Training

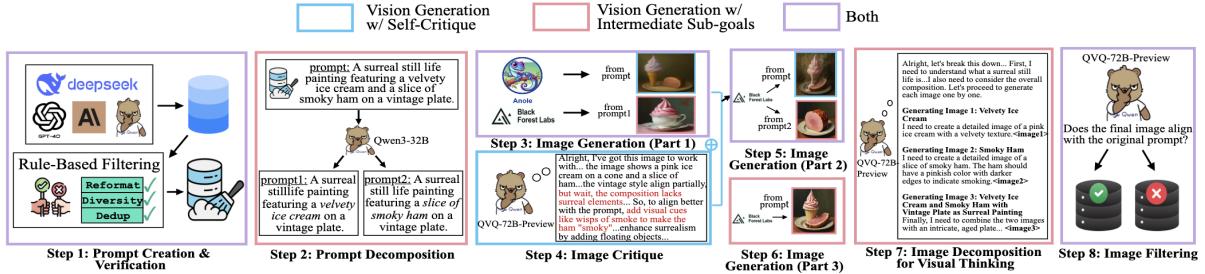


Figure 6: Our data collection pipeline for *Thinking with Generated Images*.

3.3.1 High-Quality Image-Generation Prompts

We curate diverse, complex prompts that go beyond simple object depiction to include variations in color, orientation, and shape with multiple elements. We deduplicate prompts⁵ for data efficiency and reformat them into a unified structure optimized for our “*thinking with generated images*” paradigm. For vision generation with intermediate subgoals, we use Qwen3-32B (Yang et al., 2024) to break down complex visual prompts into simpler components that can be processed independently.

3.3.2 High-Quality Reflection-Reasoning Chains

The image critique step employs QVQ-72B-Preview (Team, 2024b) for its strong long-CoT reasoning capabilities. For vision generation with self-critique, it analyzes each prompt-image pair, assessing accuracy, identifying discrepancies, and suggesting improvements. For vision generation with self-critique, it takes a holistic view of the entire image sequence, outputting structured reasoning chains that tries to reconstruct how a model could arrive at the final image through iterative decomposition. This process enables training our model to “think visually” by using images as a medium for solving tasks through reasoning.

3.3.3 High-Quality Intermediate Visual Thoughts

Our three-part image generation process creates visual representations of prompts. Initially, we use Anole-7b (Chern et al., 2024) (for vision generation with self-critique) or Flux1-dev (Labs, 2024) (for vision generation with intermediate subgoals) to generate first-round images. In the second stage, Flux1-Redux (Labs, 2024) refines images using a multimodal approach that incorporates the original prompt, first-stage image, and critique feedback to refine the image for the self-critique thought process. For the intermediate subgoals thought process, we instead use Flux1-Redux (Labs, 2024) conditioned on the first-round images (but based on the second component from the original prompt) to generate the second-round images. The third part of the image generation process continues by using Flux1-Redux (Labs, 2024) as the model (conditioned on both first and second round images) to generate a final image. Finally, QVQ-72B-Preview (Team, 2024b) performs quality control, filtering out examples where final generated images significantly deviate from the image-generation prompts.

3.4 Training

Loss Function Design There remains considerable room for optimization when training unified LMMs for visual generation tasks using only cross-entropy loss. To address image integrity issues arising from raster-based prediction, we introduce a reconstruction loss at the visual feature level. Specifically, we project the hidden states of the generated images back into the visual feature space and compute the mean squared error (MSE) loss against the corresponding features of the ground-truth image. This encourages the model to produce outputs with greater visual coherence and structural integrity.

Our ablation studies demonstrate that the model achieves optimal performance on the GenEval benchmark (Ghosh et al., 2023) when the reconstruction loss is combined with the cross-entropy loss, using a weight of 1 for the reconstruction term. Thus, in subsequent training, we adopt a composite loss function that combines cross-entropy with the reconstruction loss (weighted at 1) to enhance the visual quality of the generated images. Further details on the auxiliary loss design and the effectiveness of this approach can be found in the Appendix.

Training Stages The training is conducted in two stages. In the first stage, we performed continued training on Anole-7b with the JourneyDB dataset (Sun et al., 2023) to enhance the model’s fundamental vision generation capabilities. In the second stage of training, we train the model with the synthetic dataset constructed as described in section 3.3. We fine-tuned two models: TwGI-Anole-7b-Obj., using the vision generation with intermediate

⁵Prompt deduplication is crucial when leveraging language models for prompt generation, as these models sample from probability distributions that inherently lead to clustering around common outputs. When repeatedly sampling from these models, we observed occasional cases of repetition.

3.5 Inference

subgoals dataset, and TwGI-Anole-7b-Crit., using the vision generation with self-critique dataset. These two trained models are capable of generating visual intermediate subgoals and self-critiquing visual hypothesis, respectively.

3.5 Inference

Unlike standard VLMs or LLMs that don't need to think much about inference issues, for vision generation tasks on unified LMMs, classifier-free guidance (Ho and Salimans, 2022) will significantly improve model performance on vision generation. Specifically, on top of the full conditions, unconditions, and the image conditions as mentioned in (Team, 2024a), we add extra "original prompt conditions" and "negative conditions" to facilitate the intermediate visual steps to be more faithful without being disturbed too much by the generated long-text thought. By balancing between these conditions, we can enable the model to benefit from the long-text thought, without being distracted too much by potential noise inside the long thoughts.

3.6 Main Results

We benchmark the performance of TwGI-Anole-7b-Obj. and TwGI-Anole-7b-Crit. on GenEval (Ghosh et al., 2023) and DPGBench (Hu et al., 2024a). We benchmark TwGI-Anole-7b-Obj. against the base model Anole-7b (which is also continued pretrained on the JourneyDB dataset), exploring whether generating visual subgoals can be beneficial for vision generation on complex prompts (i.e., in GenEval and DPGBench, this refers to image generation prompts that require generating at least two objects). Note that since some categories of GenEval require generating only single object, we omit the comparison on these categories. Then, we benchmark whether the TwGI-Anole-7b-Crit. model can correct its initial visual hypothesis (i.e., TwGI-Anole-7b-Crit. (visual hypo.) in Tab. 1 and Tab. 2), and generate a better image generation result (i.e., TwGI-Anole-7b-Crit. (final) in Tab. 1 and Tab. 2).

Model	GenEval						
	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
LlamaGen (Sun et al., 2024)	0.71	0.34	0.21	0.58	0.07	0.04	0.32
LDM (Rombach et al., 2022)	0.92	0.29	0.23	0.70	0.02	0.05	0.37
Chameleon-7b (Team, 2024a)	–	–	–	–	–	–	0.39
SDv1.5 (Rombach et al., 2022)	0.97	0.38	0.35	0.76	0.04	0.06	0.43
Intermediate Visual SubGoals							
Anole-7b (Chern et al., 2024)	–	0.38	–	–	0.04	0.04	–
TwGI-Anole-7b-Obj.	–	0.57	–	–	0.11	0.16	–
Self-Critique on Visual Thoughts							
TwGI-Anole-7b-Crit. (visual hypo.)	0.98	0.51	0.24	0.84	0.06	0.08	0.45
TwGI-Anole-7b-Crit. (final)	0.98	0.59	0.23	0.87	0.12	0.10	0.48

Table 1: Performance on GenEval.

3.7 Analysis

Generating intermediate visual thoughts is beneficial for vision generation We show in Tab. 1 and Tab. 2 that TwGI-Anole-7b-Obj. consistently outperforms the baseline Anole-7b model across both GenEval and DPGBench. On GenEval, Anole-7b-Obj. achieves a significant boost in the "Two Obj." category (0.57 vs. 0.38), indicating its improved ability to handle complex prompts involving multiple entities. It also shows notable improvements in position and color attribute alignment, suggesting a stronger capacity for precise spatial and visual compositional reasoning. Similarly, on DPGBench, TwGI-Anole-7b-Obj. yields substantial gains in the "Entity", "Attribute", and "Relation" categories, reflecting its enhanced understanding of fine-grained visual semantics. These improvements

Model	Global	Entity	Attribute	Relation	Other	Overall↑
Intermediate Visual SubGoals						
Anole-7b (Chern et al., 2024)	74.16	70.31	69.42	80.12	43.60	58.32
TwGI-Anole-7b-Obj.	74.47	78.06	76.17	83.21	42.00	68.44
Self-Critique on Visual Thoughts						
TwGI-Anole-7b-Crit. (visual hypo.)	77.81	71.45	69.24	78.34	58.00	62.83
TwGI-Anole-7b-Crit. (final)	76.90	75.70	73.88	81.86	55.60	67.14

Table 2: Performance on DPG-Bench.

4. Conclusion

validate our hypothesis that breaking down visual tasks into intermediate subgoals enables LMMs to reason more systematically and generate higher-quality outputs.

The native long-multimodal thought process enables models to correct and refine their own visual hypothesis
The results on *self-critique on visual thoughts*, as shown in Tab. 1 and Tab. 2, demonstrate the effectiveness of enabling models to reflect on and revise their own visual outputs. The TwGI-Anole-7b-Crit. model achieves a notable increase in performance after the self-critique step, improving the overall GenEval score from 0.45 to 0.48, and the DPG-Bench score from 62.83 to 67.14. This indicates that the ability to introspectively analyze generated images—through textual reasoning chains grounded in visual feedback—allows the model to identify mismatches, hallucinations, or missing elements, and subsequently correct them. The effectiveness of this visual feedback loop reflects a form of inter-modality synergy, where visual and textual modalities iteratively guide each other. We further include qualitative examples in the Appendix to illustrate how critique-driven revisions result in visually and semantically improved generations.

4 Conclusion

In this paper, we introduce the concept of *Thinking with Generated Images*, and implement it with the *native long-multimodal thought process* on autoregressive next-token-prediction LMMs. While we focus on text and vision modalities, our core idea can be extended to diverse modalities. As more capable LMMs continue to emerge, particularly in the open-source domain, with the *Thinking with Generated Images* paradigm, we anticipate future AI models will explore protein structures or revise building designs as naturally as they write a poem.

5 Limitations and Future Directions

5.1 Limitations

In this paper, we demonstrate *Thinking with Generated Images* on Anole-7b (Chern et al., 2024). As mentioned in the introduction, the developmental stage (especially in the open-source realm) of unified LMMs is still evolving. With stronger unified LMMs emerging in the future, we anticipate more powerful or even emergent capabilities via the *Thinking with Generated Images* paradigm. Also, while in this paper we demonstrate our approach on autoregressive next-token-prediction LMMs, we anticipate our core idea can be applied to diffusion-based LMMs or mixed autoregressive/diffusion LMMs. We leave the exploration on different architectures to future work.

5.2 Future Directions

Better Benchmarking on Thinking with Generated Images Current vision generation benchmarks for unified LMMs focus on standard image generation tasks. In this paper, we also use standard image generation benchmarks (Ghosh et al., 2023; Hu et al., 2024a). However, with the increased inherent capabilities of LMMs and emergent capabilities coming to light, real-world tasks such as those illustrated in Fig. 1 and Fig. 2 will become increasingly feasible for LMMs via *Thinking with Generated Images*. We need more realistic benchmarks to evaluate these models.

Test-Time and Post-Training Scaling on Unified LMMs Our work represents the first step for test-time scaling on unified LMMs. As stronger unified LMMs emerge, we believe that test-time scaling and post-training scaling will become more viable, effective, and worth exploring.

Efficient Vision Representation for LMMs Efficient vision representation will be essential for achieving scalable test-time and post-training scaling in the vision modality. Recent research demonstrates that images can be effectively represented with as few as 32 or even 16 tokens/patches (Yu et al., 2025; Zhou et al., 2024). We believe this line of work is a promising direction for future research.

6 Acknowledgment

We express our sincere gratitude to Mingxuan Wang for his valuable support on this work. This project is supported by SJTU SEIEE - ByteDance Large Language Model Joint Laboratory and SII.

References

- [1] Anthropic. 2025. *Claude 3.7 sonnet system card*. Technical report, Anthropic.
- [2] Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastrauey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. 2024. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*.

- [3] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.
- [4] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.
- [5] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*.
- [6] Alan Dao and Dinh Bach Vu. 2025. *Alphamaze: Enhancing large language models' spatial intelligence via grp*.
- [7] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuan Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. 2024. *Deepseek-v3 technical report*. *CoRR*, abs/2412.19437.
- [8] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. 2025. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*.
- [9] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.
- [10] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. 2025. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*.
- [11] Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.
- [12] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.
- [13] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- [14] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024a. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.
- [15] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024b. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*.
- [16] Zhulin Hu, Yan Ma, Jiadi Su, Ethan Chern, and Pengfei Liu. 2025. Fine-tuning token-based large multimodal models: What works, what doesn't and what's next. In *The Fourth Blogpost Track at ICLR 2025*.
- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- [18] Black Forest Labs. 2024. Flux. <https://github.com/black-forest-labs/flux>.
- [19] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.
- [20] Allen Newell. 2014. Reasoning, problem solving, and decision processes: The problem space as a fundamental category. In *Attention and performance VIII*, pages 693–718. Psychology Press.

References

- [21] Fei Ni, Jianye Hao, Shiguang Wu, Longxin Kou, Jiashun Liu, Yan Zheng, Bin Wang, and Yuzheng Zhuang. 2024. Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13991–14000.
- [22] OpenAI. 2025. [Thinking with images](#).
- [23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- [24] Herbert A Simon and Allen Newell. 1971. Human problem solving: The state of the theory in 1970. *American psychologist*, 26(2):145.
- [25] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- [26] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. 2025. [Openthinkimg: Learning to think with images via visual tool reinforcement learning](#).
- [27] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. 2023. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36:49659–49678.
- [28] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.
- [29] Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898.
- [30] Chameleon Team. 2024a. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.
- [31] Qwen Team. 2024b. Qvq: To see the world with wisdom.
- [32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- [33] Shijie Xia, Yiwei Qin, Xuefeng Li, Yan Ma, Run-Ze Fan, Steffi Chern, Haoyang Zou, Fan Zhou, Xiangkun Hu, Jiahe Jin, Yanheng He, Yixin Ye, Yixiu Liu, and Pengfei Liu. 2025. [Generative ai act ii: Test time scaling drives cognition engineering](#).
- [34] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- [35] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.
- [36] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. 2025. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966.
- [37] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
- [38] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

A Experiment Details

A.1 Training Loss

This section details our loss function design, with particular emphasis on the visual modality in multi-modal training samples. Given a batch of sequences that may contain multiple images interleaved with text, our model processes each sequence through a backbone transformer to obtain hidden representations:

$$\mathbf{h} = f_{\text{trans}}(\mathbf{x}; \theta)$$

where \mathbf{x} represents the tokenized input sequence and $\mathbf{h} \in \mathbb{R}^{B \times L \times D}$ denotes the hidden states for batch size B , sequence length L , and hidden dimension D .

The model employs a unified multimodal vocabulary space where image and text tokens share the same embedding space. Specifically, vocabulary indices 4 through 8195 (8192 tokens total) correspond to visual codebook entries from the VQ-VAE, while the remaining indices represent text tokens. A multimodal language modeling head produces logits over this joint vocabulary:

$$\mathbf{z} = f_{\text{mmlm}}(\mathbf{h}; \theta_{\text{mmlm}})$$

where $\mathbf{z} \in \mathbb{R}^{B \times L \times V}$ and V is the total vocabulary size encompassing both modalities.

Our training objective combines two loss components:

Autoregressive Multimodal Loss: We compute the standard cross-entropy loss over all tokens (both text and image) in the sequence:

$$\mathcal{L}_{\text{mm}} = -\frac{1}{|\mathcal{T}|} \sum_{(b,i) \in \mathcal{T}} \log p(x_{b,i+1} | \mathbf{x}_{b,1:i})$$

where $(b, i) \in \mathcal{T}$ denotes valid token positions with $b \in \{1, \dots, B\}$ being the batch index and $i \in \{1, \dots, L_b - 1\}$ being the token position within sequence b . The probability $p(x_{b,i+1} | \mathbf{x}_{b,1:i})$ is computed by applying softmax to the logits $\mathbf{z}_{b,i} \in \mathbb{R}^V$ at position i , which are generated by the transformer conditioned on all previous tokens $\mathbf{x}_{b,1:i}$. The set \mathcal{T} excludes padded positions and positions where ground-truth labels are masked. This loss naturally handles both text token prediction and image token generation within the unified vocabulary space.

Visual Reconstruction Loss: For image tokens specifically, we extract visual features and compute a reconstruction objective. Each image in the input is demarcated by special tokens $\langle \text{BOI} \rangle$ (beginning-of-image, token ID 8197) and $\langle \text{EOI} \rangle$ (end-of-image, token ID 8196), with exactly 1024 tokens between them. For each image j in the batch:

1. Extract the corresponding hidden states: $\mathbf{h}_{\text{img}}^{(j)} \in \mathbb{R}^{1024 \times D}$
2. Project to codebook space: $\mathbf{p}^{(j)} = f_{\text{proj}}(\mathbf{h}_{\text{img}}^{(j)}; \theta_{\text{proj}})$, where f_{proj} is an additional trainable linear projection layer that maps from hidden dimension D to codebook dimension D' , i.e., $f_{\text{proj}} : \mathbb{R}^D \rightarrow \mathbb{R}^{D'}$
3. Retrieve original codebook features: $\mathbf{c}^{(j)} = f_{\text{vq}}(\text{image}_j; \theta_{\text{vq}})$, where $\mathbf{c}^{(j)} \in \mathbb{R}^{1024 \times D'}$ represents the quantized codebook features for the j -th image

The visual reconstruction loss is computed as the mean squared error (MSE) between the projected hidden states and the original codebook features:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N_{\text{img}}} \sum_{j=1}^{N_{\text{img}}} \frac{1}{1024 \cdot D'} \|\mathbf{p}^{(j)} - \mathbf{c}^{(j)}\|_2^2$$

where N_{img} is the total number of images across all sequences in the batch. Note that if a batch contains no images, $\mathcal{L}_{\text{rec}} = 0$.

The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mm}} + \lambda \mathcal{L}_{\text{rec}}$$

Previous works have made initial attempts to train autoregressive next-token-prediction LMMs beyond cross-entropy loss (Li et al., 2025; Hu et al., 2025). To rigorously benchmark the effectiveness of incorporating the visual reconstruction loss, we fine-tuned Anole-7b on 50,000 images from JourneyDB and evaluated on the GenEval benchmark. We investigated different values of λ and compared against the token discrepancy loss \mathcal{L}_{td} proposed by (Li et al., 2025). Note that we did not apply classifier-free guidance to ensure fair and simple comparison. Results in Tab. 3 demonstrate that $\lambda = 1$ yields optimal performance, improving GenEval scores by approximately 3 points. Neither \mathcal{L}_{td} alone nor the combination $\mathcal{L}_{\text{mm}} + \lambda \mathcal{L}_{\text{rec}} + \mu \mathcal{L}_{\text{td}}$ surpassed our proposed objective. Thus, we adopt $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mm}} + \mathcal{L}_{\text{rec}}$ with $\lambda = 1$ for all experiments in this paper.

A.2 Training Stages Details

Model	GenEval						
	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Baseline							
L_{mm}	0.84	0.20	0.11	0.59	0.02	0.04	0.30
Reconstruction Loss Variants							
$L_{mm} + 0.5 \cdot L_{rec}$	0.86	0.21	0.15	0.63	0.04	0.06	0.32
$L_{mm} + L_{rec}$	0.86	0.26	0.12	0.61	0.04	0.08	0.33
$L_{mm} + 5 \cdot L_{rec}$	0.86	0.22	0.12	0.58	0.03	0.04	0.31
Token Discrepancy Loss Variants							
$L_{mm} + 0.5 \cdot L_{td}$	0.84	0.24	0.11	0.58	0.02	0.05	0.31
$L_{mm} + L_{td}$	0.84	0.21	0.13	0.60	0.03	0.05	0.31
$L_{mm} + 5 \cdot L_{td}$	0.80	0.18	0.10	0.56	0.02	0.04	0.28
Combined Loss							
$L_{mm} + L_{td} + L_{rec}$	0.88	0.22	0.13	0.56	0.06	0.03	0.31

Table 3: Evaluation results on GenEval with different loss functions.

A.2 Training Stages Details

Tab. 4 shows the hyperparameters used in our training experiments. For stage-1 training, we train on approximately 4M text-image pairs. For stage-2 training, we use approximately 5k samples for intermediate visual subgoals and 40k samples for self-critique on visual thoughts, respectively.

Hyperparameters	Intermediate Visual Subgoals		Self-Critique on Visual Thoughts	
	Stage 1	Stage 2	Stage 1	Stage 2
Learning rate	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
LR scheduler	Linear	Linear	Linear	Linear
Gradient clip	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)		AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)	
Training steps	5K	2K	5K	26K
Batch size	1536	8	1536	8

Table 4: Hyperparameters for training *vision generation with intermediate visual subgoals* and *vision generation with self-critique*. Note that *vision generation with intermediate visual subgoals* and *vision generation with self-critique* share the same stage-1 model. The stage-1 model is then separately fine-tuned for generating different types of native long-multimodal thought process.

A.3 Inference Details

Classifier-free guidance (CFG) is crucial for enhanced image generation quality. We apply the following CFG settings to ensure balanced conditioning, enabling the model to concurrently leverage text-based thoughts, previous visual hypothesis, visual subgoals, and original prompts. Specific configurations are shown in Tab. 5.

CFG scale	Intermediate Visual Subgoals		Self-Critique on Visual Thoughts
	subgoals	final	
Full conditions	5.0	2.0	1.5
Image conditions	0.0	1.2	0.8
Negative conditions	3.0	3.0	3.0
Prompt conditions	0.0	5.0	5.0

Table 5: CFG scale for *vision generation with intermediate visual subgoals* and *vision generation with self-critique*.

B Case Study

B.1 Vision generation with intermediate visual subgoals

Fig. 7 shows cases of vision generation with intermediate subgoals. The model demonstrates its ability to decompose complex visual generation tasks into manageable sub-components. When generating *broccoli and vase*, the model first generates a high-quality image of a broccoli, focusing on realistic textures and vibrant green colors with detailed florets. It then generates a simple vase. In the final step, it combines these elements to create the complete image as requested by the prompt. Similarly, when tasked with generating both a *pizza and a bench*, the model first creates each object separately, producing a detailed pizza with visible toppings, and then a wooden bench. The final image successfully integrates both elements, demonstrating the model’s ability to handle multiple distinct objects in a single composition. Generating intermediate visual subgoals allows the model to focus on completing each subgoal before producing the final output, resulting in a higher-quality solution.

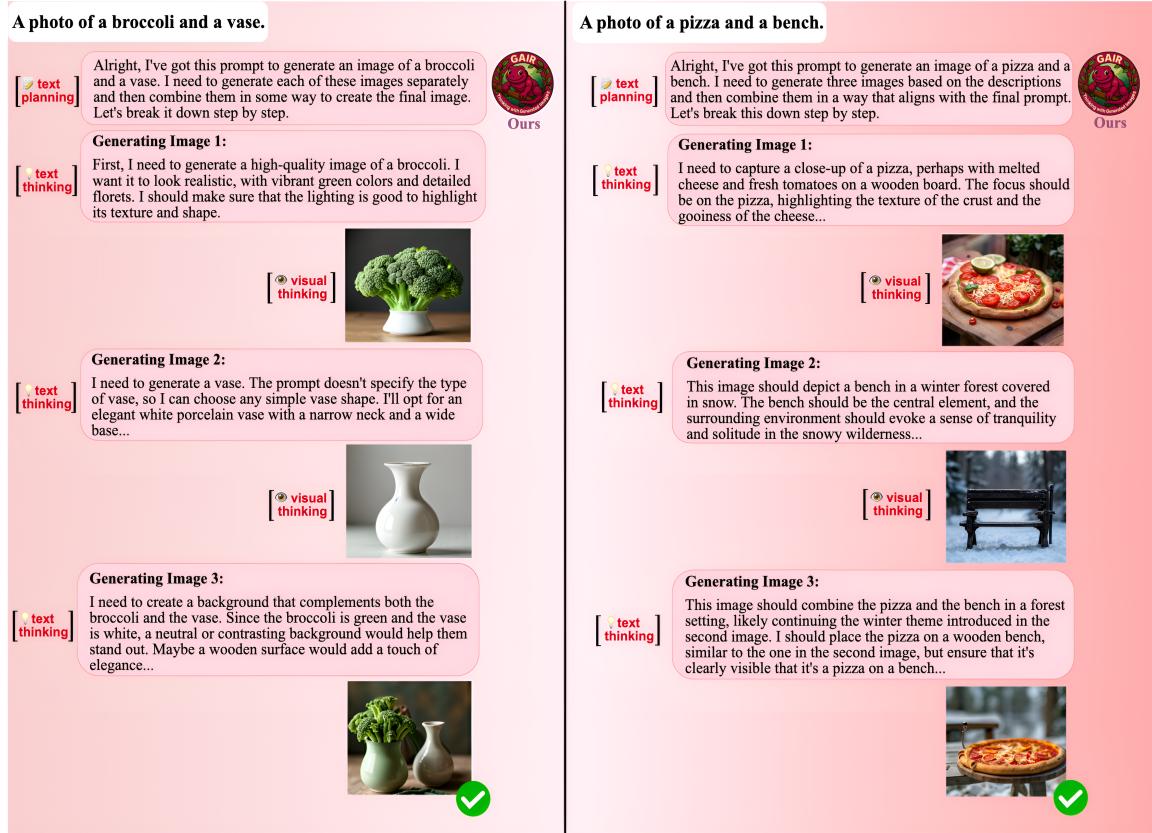


Figure 7: Vision generation with intermediate visual subgoals on GenEval.

B.2 Vision generation with self-critique

Fig. 8 shows cases of vision generation with self-critique. The model demonstrates its ability to analyze initial visual hypothesis and iteratively improve them through reflection. For generating a yin-yang symbol with a tiger head, the model initially generates an imperfect yin-yang symbol featuring a tiger. Through self-critique, it recognizes the need to emphasize the tiger faces to be more stylized and integrated. In the final version, the model successfully implements the corrections, with the tiger head now seamlessly blend into the yin-yang design. For generating a 4K-resolution portrait of “Blind Ambition,” the model first produces a split-view sketch with rough annotations, uneven shading, and a cluttered background that breaks the required central symmetry and leaves the character’s gaze only loosely obscured. Upon self-critique, it recognizes that the composition must be unified and centrally framed, the metaphorical blindness needs either a clear blindfold or shadowed hollows, and the background should shift from sketchy walls to a neutral, atmospheric void while textures demand deeper, more subtle rendering. The final image successfully centers a lone figure in a balanced stance, with warm steel armor carved by refined gradients, and a misty, minimalist backdrop that underscores the character’s solitary, driven ambition.

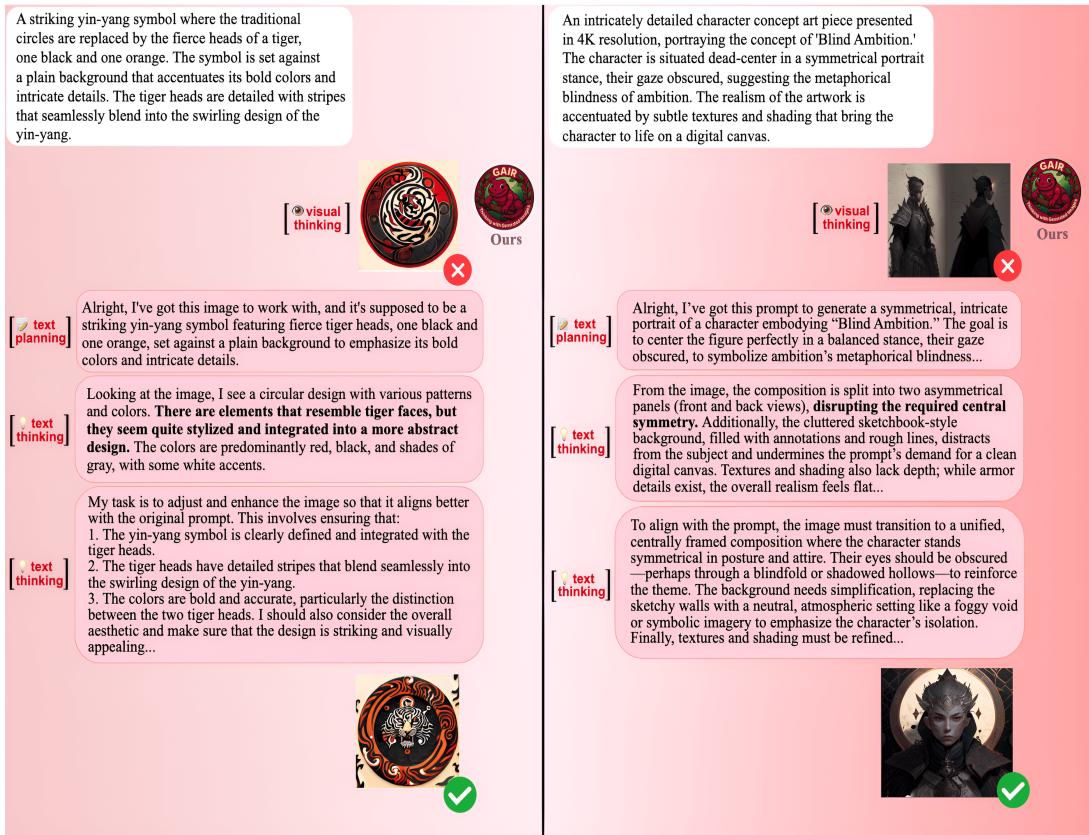


Figure 8: *Vision generation with self-critique* on DPG-Bench.

B.3 Failure Cases

B.3 Failure Cases

The left-half of Fig. 9 shows a failure case of vision generation with intermediate subgoals. While the model successfully generates individual components, it fails to combine them in the final output. When tasked with generating *microwave and bench*, the model first produces a microwave placed on a kitchen counter, then generates a bench in a park setting. However, in the final step, the model fails to meaningfully combine both elements. This failure illustrates that while the model can correctly generate individual subgoals (the microwave and the bench separately), it sometimes struggles with the spatial reasoning required to integrate disparate objects that don't naturally belong together. We anticipate this issue to be mitigated with stronger base models or further post-training.

The right-half of Fig. 9 shows a failure case of vision generation with self-critique. When prompted to generate *a photo of a tv remote*, the model initially produces a retro TV set instead, completely misinterpreting the request. Through self-critique, the model correctly identifies this error and provides detailed suggestions: remove the TV, focus on the remote either held in hand or on a surface, and ensure the remote is the main subject. Despite this accurate analysis and reasonable corrective suggestions, the second attempt still shows a TV set rather than a TV remote. This failure demonstrates a disconnect between the model's analytical capabilities and generative execution—while the self-critique mechanism can accurately identify problems and propose solutions, the model sometimes fails to implement its own corrections in subsequent generations. The model can reason about what's wrong but cannot translate this understanding into proper image generation. Similarly, we anticipate this issue to be mitigated with stronger base models or further post-training.

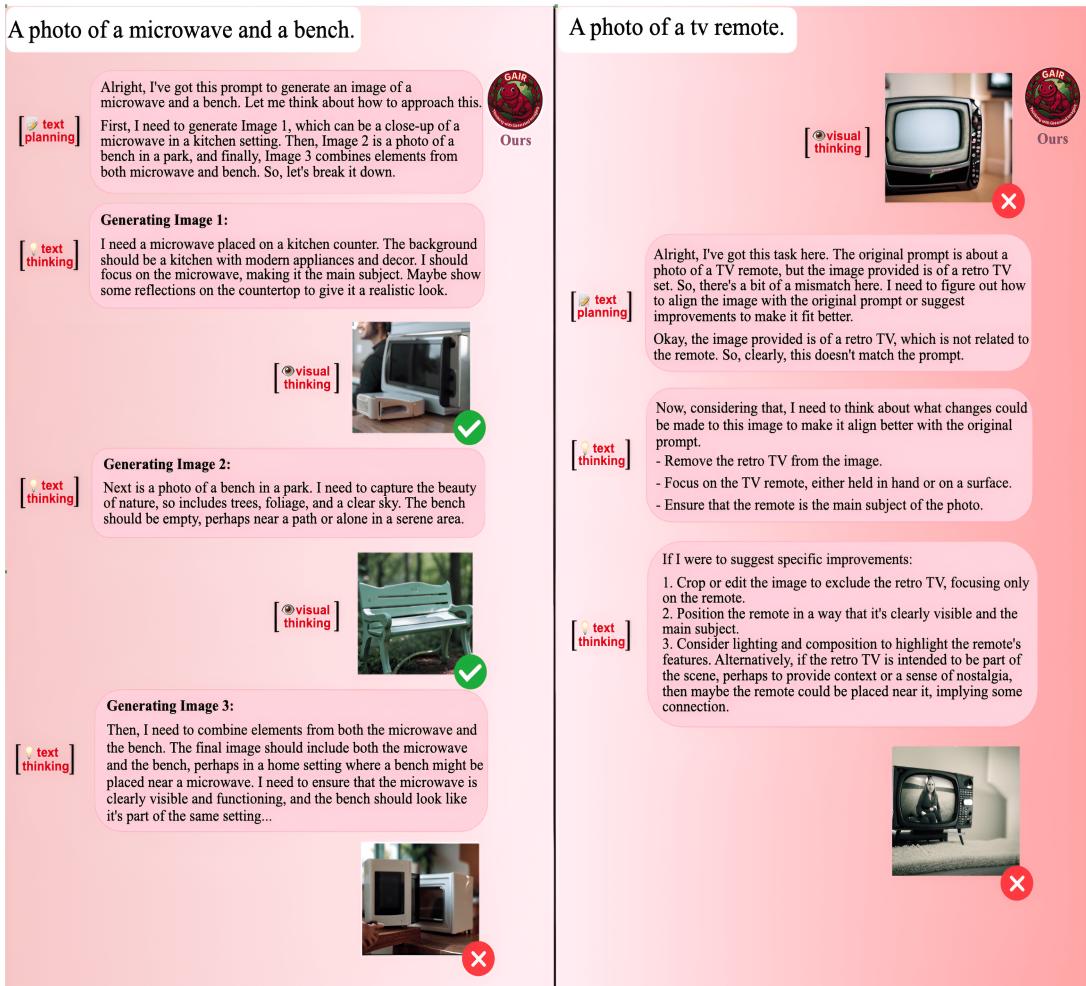


Figure 9: Example failure cases of using *thinking with generated images*. Either the intermediate images generated are correct but fails to combine them together in the end, or the critique is correct but didn't generate the corresponding correct image.