SymbOmni: Evolving Agentic Omni Models via Symbolic Concept Learning

Jinxiu Liu[1,2*], Jianru Li[1*], Tanqing Kuang[1*], Xuanming Liu[2], Kangfu Mei[3], Yandong Wen[2†], and Weiyang Liu[4]

[1]South China University of Technology, China [2]Westlake University, China
[3]Johns Hopkins University, USA [4]The Chinese University of Hong Kong, HK SAR

spherelab.ai/symbomni

Abstract. Visual generation is increasingly ubiquitous in diverse domains, from text-to-image/video synthesis to multimodal interactive creation. Yet prevailing monolithic models remain fundamentally constrained by their inability to learn cumulatively and evolve autonomously, which is a limitation we term the "perpetual novice" problem. They lack mechanisms for structuring experience into reusable knowledge and therefore rely on brittle, "from-scratch" reasoning for each task, resulting in poor compositional generalization and inefficient knowledge retention. Motivated by these limitations, we propose SymbOmni, an agentic omni-model designed for cumulative evolution through *Symbolic Concept Learning*. At its core is the Symbolic Concept Box, an optimizable memory module that abstracts low-level operations into reusable Symbolic Workflow Instructions. SymbOmni operates through an *induction-transduction cycle*: experiences are abstracted into symbolic concepts (induction), which are then adaptively composed to solve novel tasks (transduction). The training is done by verbalized backpropagation with language-based feedback to enable continuous self-improvement without gradient-based model fine-tuning. Comprehensive experiments validate that (I) SymbOmni significantly outperforms existing agent-based systems for iterative creation and also surpasses closed-source models (*e.g.*, Nano Banana, GPT-Image-1) in both image quality and task success rates; (II) SymbOmni effectively reduces token consumption by over 40% while maintaining competitive generation quality; and (III) SymbOmni enables effective continual learning by achieving cumulative gains across multiple online-learning benchmarks and setting a new state of the art.

1 Introduction

Visual generation has become essential to a wide range of applications, including text-to-image synthesis [19, 47, 73], text-to-video generation [18, 26], and interactive content generation [1, 12]. While standalone text-to-image or text-to-video models often struggle to meet complex creative demands [69, 73], emerging unified or task-specific frameworks [35, 53, 56] aim to bridge this gap. However, within the open-source community, these unified alternatives still suffer

*Equal contribution; authors may list themselves as first author. †Corresponding author

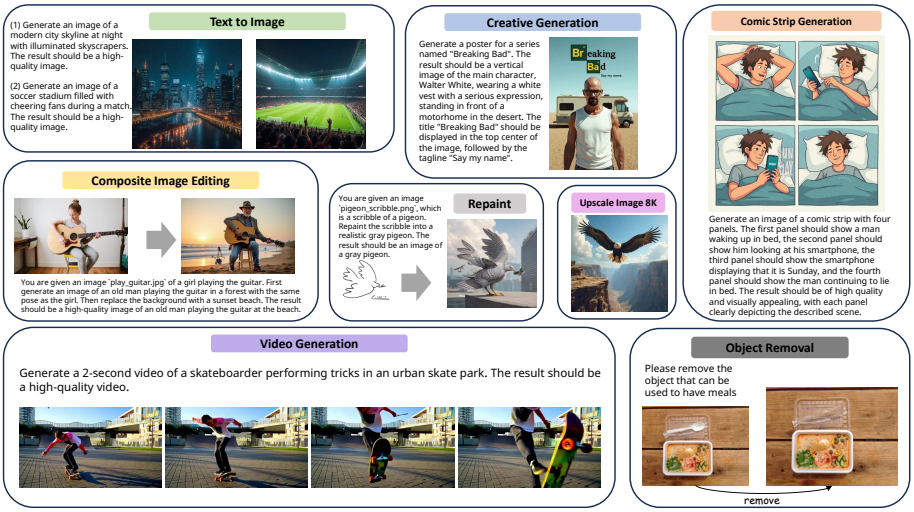


Fig. 1: Overview of the proposed SymbOmni framework. The model achieves versatile visual generation (text-to-image, creative content, image editing, video synthesis) through a unified approach for cumulative symbolic concept learning.

from unstable generation quality and insufficient structured planning [22, 67]. Although proprietary counterparts exhibit remarkably robust unified capabilities [38, 40, 59], their closed nature inherently restricts architectural flexibility, downstream scalability, and broader multimodal adaptation [30, 62].

Another fundamental limitation further challenges current visual generation paradigms. Existing Large language models (LLMs) and agentic systems lack the capacity for cumulative learning (or continual learning) and self-evolution [4, 17, 61]. By treating tasks in isolation, prevailing end-to-end architectures fail to distill useful and transferable knowledge from prior experiences [51, 61]. This results in the *perpetual novice problem*, where the generative system exhibits poor compositional generalization, suboptimal knowledge retention, inefficient reasoning and brittle decision-making [34, 54, 66].

This perpetual novice problem is rooted at the framework level. Current prevailing unified generative-understanding paradigms often lack the mechanisms required to structure experiences into compact and reusable knowledge components. As a result, these systems must rely on from-scratch reasoning within a fixed parametric memory for each new task [39, 40]. While alternative approaches may leverage external tools, they often produce unstable and non-reusable reasoning chains [61, 65, 68]. Furthermore, existing attempts to mitigate these issues via implicit learning or tool-calling agents suffer from poor scalability and decision instability [29, 50]. Even workflow-based frameworks are hampered by rigid architectures that fail to adapt to dynamic user demands [10, 20, 43]. Moreover, this limitation is particularly significant in enterprise scenarios with highly

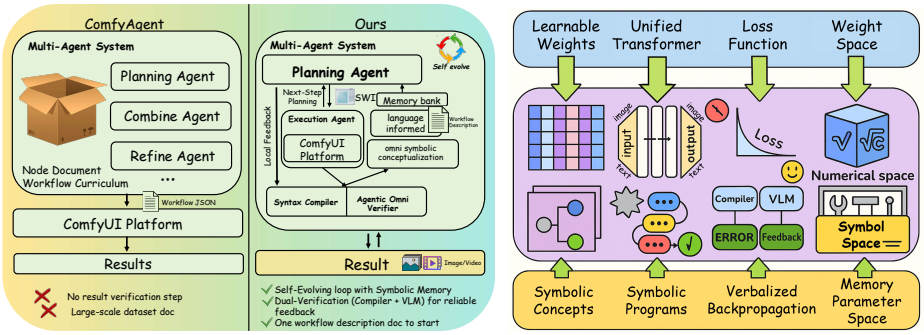
personalized workflows, where long-context methods can incur prohibitive computational costs and elevated hallucination risks [22, 27, 67].

In this work, we introduce **SymbOmni**, an agentic architecture designed for continual evolution through Symbolic Concept Learning. We argue that long-term adaptation requires agents to continuously abstract their experiences into structured, composable knowledge representations. To this end, SymbOmni incorporates a strong inductive bias that combines the flexibility of neural models with the rigor and interpretability of symbolic reasoning. At the core of the architecture is the Symbolic Concept Box (CB), an optimizable memory that stores reusable units of knowledge, termed Symbolic Concepts. Each concept integrates semantic understanding with executable procedures that encode a successful problem-solving strategy. These strategies are further distilled into Symbolic Workflow Instructions (SWIs), which are dynamically retrieved, composed, and executed through a memory-augmented mechanism. This design enables an *Induction-Transduction cycle*. During induction, the agent abstracts past experiences into reusable symbolic concepts. During transduction, it retrieves, composes, and instantiates relevant concepts to solve new tasks. The cycle is completed through verbalized backpropagation [39, 50, 64, 70, 71, 74], which generates linguistic feedback to refine the agent's symbolic knowledge and problem-solving strategies. In this way, SymbOmni supports continual self-improvement without requiring parameter finetuning. Our contributions are threefold:

- **Continual adaptation framework**. SymbOmni is a Symbolic Concept Learning architecture for continual adaptation beyond monolithic models [20, 61, 65].
- **Verbalized learning**. SymbOmni integrates both inductive and transductive reasoning for cumulative learning via verbalized backpropagation [50, 64].
- **Empirical effectiveness**. SymbOmni outperforms existing approaches, reducing token use by over 40% and consistently improving workflows [22, 67, 75].

2 Related Work

Agentic Systems and Workflow Automation. Current AI systems often struggle with *cumulative learning*, as they treat tasks in isolation and suffer from the perpetual novice problem, which leads to poor compositional generalization and inefficient knowledge retention [34, 51, 54, 61]. Recent work has sought to address these limitations through structured knowledge representations and automated workflow construction. Systems such as Deep Research [75] and OpenAI o3 [40] translate natural-language objectives into executable plans, while WorkflowLLM [10] and AFlow [72] emphasize data-driven workflow optimization. Early multi-agent systems [20, 43] enable collaborative problem solving, but often rely on rigid interaction protocols. In specialized environments such as ComfyUI, approaches including ComfyAgent [67] and ComfyGPT [22] support node-based workflow generation, and yet remain constrained by limited domain



(a) Comparison with baseline agent systems. **(b)** Unified models vs. our symbolic model.

Fig. 2: (a) The comparison underscores the advantages of our model: a self-evolving capability through dual-verification feedback and an optimizable symbolic memory, which eliminates the dependency on large-scale pre-existing documentation required by the baseline approach. (b) Contrast between multimodal unified models and our agentic symbolic framework.

coverage and cascading error propagation. More general frameworks such as AutoGen [61] and ReAct [68] enable iterative refinement, but still lack reliable mechanisms for error diagnosis and recovery. Building on these advances, our work formalizes an *Induction-Transduction* cycle together with a dual-verifier architecture to enable genuine cumulative learning.

Visual Generation and Reasoning Systems. Visual generation has progressed from task-specific models [69,73] to general-purpose multimodal systems [35,53], driven by advances in diffusion models [47] and unified architectures such as GPT-Image-1 [38]. In parallel, large reasoning models have substantially improved planning through System-2 reasoning [24] and reinforcement learning [15]. However, these systems often depend on lengthy and computationally expensive contexts, which can amplify hallucinations and limit efficient knowledge reuse. As illustrated in Figure 2b, existing unified multimodal models primarily acquire knowledge through static parametric updates in weight space. In contrast, our architecture introduces an explicit symbolic concept layer that continuously abstracts, consolidates, and reuses knowledge through verbalized learning. SymbOmni unifies symbolic concept learning with visual workflow generation, yielding reusable knowledge components that reduce computational overhead, support continual improvement, and address the challenge of structured knowledge consolidation beyond conventional end-to-end approaches.

3 SymbOmni: Evolving Agentic Visual Generation

3.1 Preliminaries on ComfyUI and Node-based Workflows

ComfyUI [7] is an open-source, node-based interface for visual generation. It can represent generative pipelines as directed acyclic graphs (DAGs) of intercon-

nected nodes, each encapsulating a discrete operation (*e.g.*, model loading, latent sampling, or post-processing). Users compose multi-stage workflows by wiring node outputs to inputs, and the result is serialized as a JSON file specifying all node types, parameters, and connections. This modularity offers fine-grained control of generated content but demands significant expertise in node selection, configuration, and connectivity. To address this challenge, our work focuses on automating the construction of such workflows.

3.2 Overview of the SymbOmni Framework

Our Omni Agentic Model introduces a unified framework that integrates induction and transduction within a self-reinforcing cycle. During the transductive phase, the system draws on reusable symbolic knowledge stored in the Symbolic Concept Box to solve novel tasks. It adapts abstract concepts and constraints to specific problem instances while preserving the underlying knowledge base. The inductive phase serves as the primary learning mechanism, transforming concrete experiences (both successful and unsuccessful queries) into symbolic knowledge by extracting generalizable rules and patterns from individual cases. This process continuously expands and refines the Symbolic Concept Box. Put together, these mechanisms effectively form the coherent "Induction-Memory-Transduction-Experience" cycle illustrated in Figure 3. Induction enriches symbolic memory with generalized knowledge, which subsequently improves the effectiveness of transduction on future tasks. This reciprocal reinforcement supports continuous meta-learning, allowing each new task and experience to further strengthen the system's reasoning capabilities.

3.3 Symbolic Concepts and Concept Box

A key characteristic of our framework is its representation of knowledge in the form of Symbolic Concepts. Each Symbolic Concept, denoted by C_k, encapsulates a reusable unit of expert knowledge and is defined by the quadruple:

$$C_k = (Desc_k, SWI_k, Params_k, Score_k) \tag{1}$$

where $Desc_k$ is a natural-language description that specifies the concept's intended purpose; SWI_k is a Symbolic Workflow Instruction, represented as a parameterized template describing a sequence of operations; $Params_k$ contains optimized parameter configurations learned from prior applications; and $Score_k$ records a dynamically updated measure of the concept's historical effectiveness. During execution, the parameter slots in SWI_k are instantiated using the corresponding values stored in $Params_k$. The collection of all Symbolic Concepts forms the Concept Box, which serves as the system's long-term, optimizable memory. We formalize the CB as a language system, $\mathcal{L} = (\Sigma, \mathcal{R})$, where $\Sigma = \{Desc_k\}$ defines a vocabulary of semantic primitives, and $\mathcal{R} = \{SWI_k\}$ denotes a set of production rules that specify valid workflow compositions. Under this formulation, planning can be viewed as a grammatical derivation in $\mathcal{L}$,

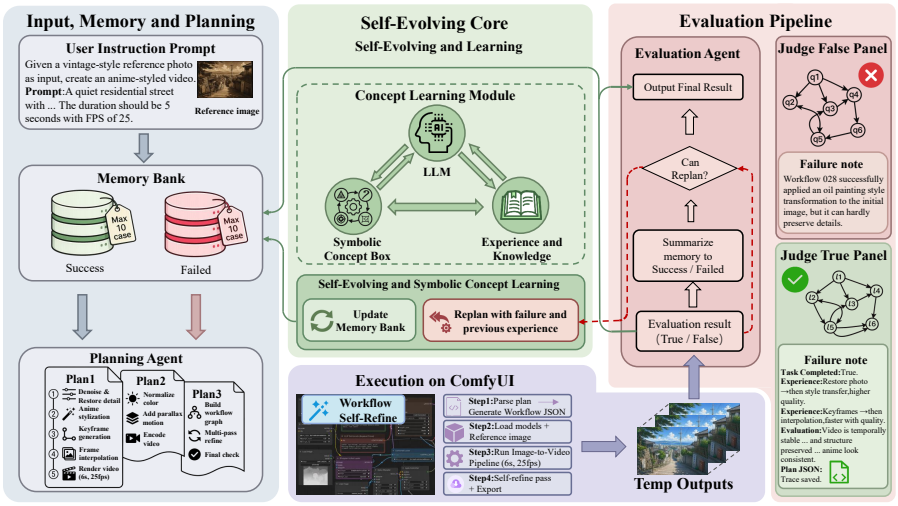


Fig. 3: Illustration of SymbOmni's self-evolving symbolic concept learning mechanism. The framework iteratively transforms user instructions into solutions through a dual-feedback loop: successful outcomes are abstracted into symbolic experiences and stored in the Memory Bank, while failures trigger a replay process that refines future planning. This continuous learning cycle enables the agent to accumulate and leverage transferable knowledge beyond individual tasks.

transforming problem solving from the derivation of solutions from scratch into the composition of previously validated building blocks. The natural-language descriptions $Desc_k$ provide the semantic foundation for assessing the similarity between stored concepts and new tasks, thereby supporting the hierarchical retrieval process detailed in Section 3.4.

3.4 The Self-Evolution Cycle

Equipped with the structured Symbolic CB, SymbOmni operates through a continuous four-phase cycle that seamlessly integrates planning, execution, and learning, as illustrated in Figure 3. In this cycle, we introduce a dual-feedback mechanism that enables the system to accumulate, refine, and reuse knowledge from previous experience. The following sections describe each phase in detail.

Phase 1: Transduction (Planning). The cycle begins with the Transduction phase (*i.e.*, task planning), where the system translates a new task instruction into a concrete, executable plan by leveraging knowledge stored in the CB. Given task instruction I_{new}, the hierarchical retrieval proceeds as follows. First, $ST = \text{Decompose}(I_{\text{new}})$ parses the instruction into semantic subtasks with dependencies. Then, concepts are retrieved through a two-level cascade:

$$C_{\text{ret}} = \text{Rank} \circ \text{Filter}_{\text{struct}} \circ \text{Retrieve}_{\text{sem}}(ST, \text{CB}) \tag{2}$$

where the $\text{Retrieve}_{\text{sem}}$ operation identifies relevant concepts based on the cosine similarity between their embeddings, *i.e.*, $\cos(\text{Embed}(Desc_k), \text{Embed}(ST_m))$. The $\text{Filter}_{\text{struct}}$ operation then enforces dependency constraints, while Rank orders the remaining concepts according to their utility scores. Conditioned on the retrieved concepts and the grammar $\mathcal{G}$, the LLM planner generates workflows:

$$WF_{\text{cand}} = \text{Instantiate}(\text{LLM}(I_{\text{new}}|C_{\text{ret}}, \mathcal{G})) \tag{3}$$

where Instantiate binds the parameters Params_k to the corresponding workflow templates SWI_k, producing executable workflow candidates.

Phase 2: Execution. The Execution phase is quite straightforward. The system executes the planned workflow WF_{cand} step-by-step, invoking the corresponding tools (*e.g.*, image generators, filters) to produce the final output O. Throughout this process, the complete execution trajectory τ is recorded in detail and subsequently used as supervision for the learning phase:

$$\tau = (I_{\text{new}}, WF_{\text{cand}}, O, C_{\text{ret}}, \text{ExecutionLog}). \tag{4}$$

Phase 3: Induction (Learning) and Evaluation. The Induction phase forms the core of the system's learning process. It begins with a critical evaluation step, in which the Evaluation Agent assesses the generated output O against the input instruction I_{new} and produces a binary judgment, $\text{Judge} \in \{\text{True}, \text{False}\}$. This judgment determines the subsequent learning path, thereby leading to the proposed dual-feedback mechanism.

If Judge = True (successful execution), indicating successful execution, the trajectory τ is distilled into positive symbolic experience. Through *symbolic abstraction*, the system extracts generalizable knowledge from the trajectory, either constructing new composite concepts or reinforcing existing ones.

$$\text{CB} \leftarrow \text{CB} \cup \text{Symbolize}(\tau, \text{Positive}) \tag{5}$$

If Judge = False (failed execution), indicating failed execution, the system initiates a detailed diagnostic process. Inspired by [64], the failure is analyzed through *verbalized backpropagation*. Specifically, two complementary verifiers evaluate the execution: $\text{Error}_{\text{hard}}$ assesses syntactic and procedural correctness, while $\text{Feedback}_{\text{soft}}$ evaluates semantic quality. An LLM then analyzes these signals and feeds them into a structured *linguistic loss* L_{lang}, providing an interpretable diagnosis of the failure. Subsequently, chain attribution computes *symbolic gradients* ∇_{sym} to pinpoint refinement targets in the CB, guiding a replay and refinement process.

$$L_{\text{lang}} = \text{LLM}(\mathcal{P}_{\text{loss}}(\tau, \text{Error}_{\text{hard}}, \text{Feedback}_{\text{soft}})) \tag{6}$$

$$\text{CB} \leftarrow \text{Refine}(\text{CB}, \text{Replay}(\tau, \nabla_{\text{sym}})) \tag{7}$$

Phase 4: Memory Optimization. In the *Memory Optimization* phase, the insights acquired during Phase 3 are consolidated into the Symbolic CB. Newly constructed or refined symbolic concepts are incorporated into memory. Successful trajectories are abstracted into reusable knowledge, denoted by C_{success}, whereas failed trajectories yield either negative concepts for avoiding recurring errors, C_{negative}, or targeted parameter updates, C_{refine}. Guided by verbalized learning, this update completes a full self-evolution cycle. The resulting enriched CB enables the system to solve related tasks more effectively in the future, thereby effectively achieving continual learning.

4 Experiments and Results

We evaluate our system's performance on three complementary public benchmarks: ComfyBench for assessing AI agent workflows and automatic workflow generation from natural language descriptions [67], GenEval for assessing semantic consistency and visual fidelity in text-to-image generation tasks [14], and ReasonEdit for examining performance in multi-step reasoning and complex image editing scenarios [23]. Additionally, we analyze the token efficiency and runtime of SymbOmni on large-scale tasks, and examine the performance differences between Online and Offline modes to demonstrate self-evolution capabilities and token efficiency [22, 31, 50, 54, 75]. We further conduct ablation studies to validate the contribution of the symbolic concept memory, a large-scale user preference study to evaluate perceptual quality, and out-of-domain generalization experiments to assess transferability of the learned concepts.

4.1 Experimental Setup

To rigorously evaluate planning and learning capabilities, we enhance the workflow library with improved coordination mechanisms and high-quality workflows, while deliberately retaining several suboptimal ones. All experiments were conducted in the ComfyUI framework [7], using Gemini 2.5 Flash [6] as the reasoning engine. For fair comparison across agentic systems, we standardize key hyperparameters, setting the maximum search depth to 10 and the maximum number of retries to 4 [28, 41, 48, 68]. Unified generative models and published baselines followed their respective benchmark protocols. We evaluate performance along four different dimensions: *Pass Rate*, the proportion of generated workflows that execute successfully with the correct structure; *Resolve Rate*, the proportion of final outputs that satisfy the semantic requirements of the instruction; *Generation Quality*, the semantic consistency and perceptual quality of the outputs; and *Token Consumption*, the average number of tokens used per task as an indicator of reasoning efficiency [33, 37, 44].

4.2 Qualitative Comparison

We qualitatively compare SymbOmni with state-of-the-art collaborative AI systems [16] and leading closed-source unified models [6, 8, 38]. As shown in Figure 4,

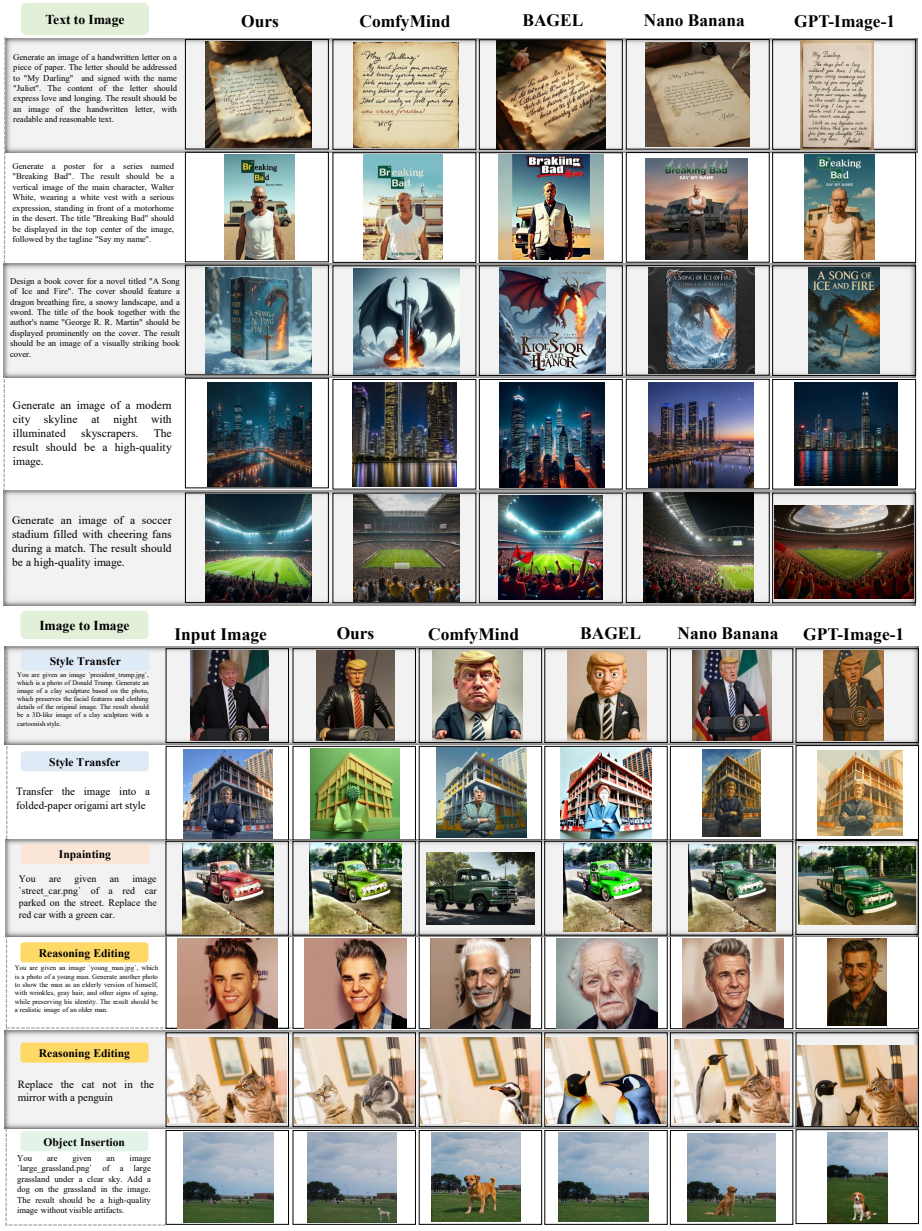


Fig. 4: Qualitative comparison among SymbOmni, ComfyMind, Bagel, Nano Banana and GPT-Image-1 for two representative visual generation tasks. Top five rows: text-to-image generation; Bottom five rows: image-to-image reasoning and generation. We can observe that the quality of SymbOmni's output is on par with the state-of-the-art proprietary image generation systems, and SymbOmni shows the best instruction-following ability among all.

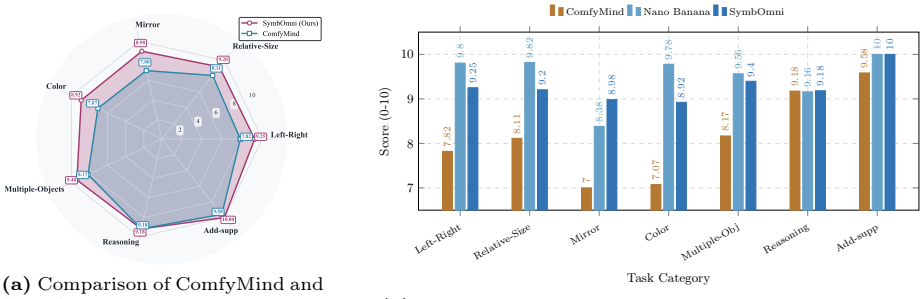
Agent	Vanilla		Complex		Creative		Total	
	%Pass↑	%Resolve↑	%Pass↑	%Resolve↑	%Pass↑	%Resolve↑	%Pass↑	%Resolve↑
GPT-4o + Few-shot [2]	32.0	27.0	16.7	8.3	7.5	0.0	22.5	16.0
GPT-4o + CoT [58]	44.0↑12.0	29.0↑2.0	11.7↓5.0	8.3↑0.0	12.5↑5.0	0.0↑0.0	28.0↑5.5	17.0↑1.0
GPT-4o + CoT-SC [57]	45.0↑13.0	34.0↑7.0	11.7↓5.0	5.0↓3.3	15.0↑7.5	0.0↑0.0	29.0↑6.5	18.5↑2.5
Claude-3.5-Sonnet + RAG [27]	27.0↓5.0	13.0↓14.0	23.0↑6.3	6.7↓1.6	7.5↑0.0	0.0↑0.0	22.0↓0.5	8.5↓7.5
Llama-3.1-70B + RAG	58.0↑26.0	32.0↑5.0	23.0↑6.3	10.0↑1.7	15.0↑7.5	5.0↑5.0	39.0↑16.5	20.0↑4.0
GPT-4o + RAG	62.0↑30.0	41.0↑14.0	45.0↑28.3	21.7↑13.4	40.0↑32.5	7.5↑7.5	52.0↑29.5	23.0↑7.0
o1-mini + RAG	32.0↑0.0	16.0↓11.0	21.7↑5.0	8.3↑0.0	12.5↑5.0	7.5↑7.5	25.0↑2.5	12.0↓4.0
o1-preview + RAG	70.0↑38.0	46.0↑19.0	48.3↑31.6	23.3↑15.0	30.0↑22.5	12.5↑12.5	55.5↑33.0	32.5↑16.5
ComfyAgent [67]	67.0	46.0	48.3	21.7	40.0	15.0	56.0	32.5
ComfyMind [16] (reproduced)	100.0↑33.0	84.0↑38.0	100.0↑51.7	75.0↑53.3	100.0↑60.0	42.5↑27.5	100.0↑44.0	73.0↑40.5
SymbOmni (Ours)	100.0↑33.0	95.0↑49.0	100.0↑51.7	83.3↑61.6	100.0↑60.0	67.5↑52.5	100.0↑44.0	86.0↑53.5
+ Nano Banana [6] (Unified SOTA Model)								
ComfyMind + Nano Banana	100.0↑33.0	97.0↑51.0	100.0↑51.7	80.0↑58.3	100.0↑60.0	57.5↑42.5	100.0↑44.0	84.0↑51.5
SymbOmni + Nano Banana	100.0↑33.0	99.0↑53.0	100.0↑51.7	88.3↑66.6	100.0↑60.0	75.0↑60.0	100.0↑44.0	91.0↑58.5

Table 1: Evaluation of autonomous workflow construction on ComfyBench [67]. We highlight the best results with colored boxes.

our method demonstrates clear advantages in handling complex compositional tasks. In the text-to-image generation tasks which requires precise adherence to complex instructions, SymbOmni achieves stronger semantic alignment when instructions contain multiple fine-grained requirements. For the "Breaking Bad" poster, it correctly depicts Walter White wearing a white vest with a serious expression in a desert setting, while accurately rendering the title and tagline. In contrast, several other baselines fail to preserve the character appearance, scene details, or textual elements. In the complex image editing and image-to-image generation tasks, SymbOmni reliably executes multi-step editing operations. In the "red car to green car" example, it both changes the car color and extends the canvas as instructed, whereas other methods often satisfy only part of the request. Similarly, in transforming Donald Trump into a clay sculpture, SymbOmni better preserves facial identity while applying the intended cartoon-like, three-dimensional clay texture. In contrast, baselines frequently distort the subject or fail to reproduce the desired style. These results suggest that symbolic concept learning enables SymbOmni to interpret and execute complex, multi-faceted instructions more accurately, yielding stronger performance across both image generation and editing tasks compared to existing methods [1,12,45,69,73].

4.3 Quantitative Experiments

ComfyBench evaluates the ability to construct executable workflows from task descriptions [67]. As shown in Table 1, SymbOmni achieves a perfect 100% pass rate and a total resolve rate of 86.0%, significantly outperforming all baseline methods. The advantage is most substantial on complex and creative tasks. For complex tasks, the resolve rate improves from 75.0% to 83.3% (11.1% relative gain). For the challenging creative tasks, it improves from 42.5% to 67.5%, a remarkable 58.8% relative gain. These gains stem from SymbOmni's ability to



(a) Comparison of ComfyMind and SymbOmni on ReasonEdit.

(b) Detailed subcategory comparison with Nano Banana.

Fig. 5: Evaluation of reasoning image-to-image editing on ReasonEdit [23]. (a) Radar chart showing overall performance comparison of SymbOmni and ComfyMind across subcategories. (b) Detailed bar chart comparing SymbOmni, Nano Banana [6], and ComfyMind [16] across seven reasoning-intensive task types.

accumulate and reuse learned symbolic concepts, enabling more flexible workflow composition and stronger generalization to novel task configurations [10, 36, 72].

When further equipped with a unified SOTA model Nano Banana [6], SymbOmni reaches 91.0% total resolve rate. Under identical tool configurations, SymbOmni consistently surpasses ComfyMind (91.0% vs. 84.0% overall; 75.0% vs. 57.5% on Creative), verifying that the gains originate from our symbolic concept learning mechanism rather than the underlying generative model.

Reason-Edit: Complex Visual Reasoning Editing. The ReasonEdit benchmark evaluates multi-step visual reasoning editing capabilities [23]. As shown in Figure 5a, SymbOmni achieves an overall score of 9.190, substantially outperforming ComfyMind (8.135), with the most pronounced gains in complex spatial reasoning (+1.980 on Mirror, +1.229 on Multiple-Objects) [34, 55, 66].

Figure 5b compares SymbOmni with unified multimodal models. SymbOmni performs particularly well on reasoning-intensive tasks, achieving a score of 8.983 on the Mirror task, which is substantially higher than Nano Banana (8.383; +7.2%) and ComfyMind (7.000; +28.3%). This validates the effectiveness of symbolic workflow decomposition for multi-step spatial transformations. Although Nano Banana performs slightly better in some subcategories (*e.g.*, Left-Right and Color), these gains largely reflect the capabilities of its underlying generative model. As an agentic system, SymbOmni is inherently constrained by the quality of its constituent tools; nevertheless, it achieves a perfect score of 10.000 on Add-supp and remains competitive across all evaluation dimensions.

GenEval: Text-to-Image Generation. As shown in Table 2, SymbOmni achieves a state-of-the-art overall score of 0.98, substantially outperforming methods based on frozen text encoders (best: SD3-Medium, 0.74), LLM/MLLM-enhanced approaches (best: GPT-Image-1, 0.84), and collaborative AI systems (best: ComfyMind, 0.90). Its advantage is particularly pronounced on challenging compositional tasks, where it attains near-perfect scores of 0.97 for Position and 0.95 for

Method	Overall↑	Single Obj.↑	Two Obj.↑	Counting↑	Colors↑	Position↑	Attr. Binding↑
Frozen Text Encoder Mapping Methods							
SDv1.5 [47]	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SDv2.1 [47]	0.50$_{\uparrow 0.07}$	0.98$_{\uparrow 0.01}$	0.51$_{\uparrow 0.13}$	0.44$_{\uparrow 0.09}$	0.85$_{\uparrow 0.09}$	0.07$_{\uparrow 0.03}$	0.17$_{\uparrow 0.11}$
SD-XL [42]	0.55$_{\uparrow 0.12}$	0.98$_{\uparrow 0.01}$	0.74$_{\uparrow 0.36}$	0.39$_{\uparrow 0.04}$	0.85$_{\uparrow 0.09}$	0.15$_{\uparrow 0.11}$	0.23$_{\uparrow 0.17}$
DALLE-2 [46]	0.52$_{\uparrow 0.09}$	0.94$_{\downarrow 0.03}$	0.66$_{\uparrow 0.28}$	0.49$_{\uparrow 0.14}$	0.77$_{\uparrow 0.01}$	0.10$_{\uparrow 0.06}$	0.19$_{\uparrow 0.13}$
SD3-Medium [9]	0.74$_{\uparrow 0.31}$	0.99$_{\uparrow 0.02}$	0.94$_{\uparrow 0.56}$	0.72$_{\uparrow 0.37}$	0.89$_{\uparrow 0.13}$	0.33$_{\uparrow 0.29}$	0.60$_{\uparrow 0.54}$
Multimodal Unified Models							
LlamaGen [52]	0.32	0.71	0.34	0.21	0.58	0.07	0.04
LWM [32]	0.47$_{\uparrow 0.15}$	0.93$_{\uparrow 0.22}$	0.41$_{\uparrow 0.07}$	0.46$_{\uparrow 0.25}$	0.79$_{\uparrow 0.21}$	0.09$_{\uparrow 0.02}$	0.15$_{\uparrow 0.11}$
SEED-X [13]	0.49$_{\uparrow 0.17}$	0.97$_{\uparrow 0.26}$	0.58$_{\uparrow 0.24}$	0.26$_{\uparrow 0.05}$	0.80$_{\uparrow 0.22}$	0.19$_{\uparrow 0.12}$	0.14$_{\uparrow 0.10}$
Emu3-Gen [56]	0.54$_{\uparrow 0.22}$	0.98$_{\uparrow 0.27}$	0.71$_{\uparrow 0.37}$	0.34$_{\uparrow 0.13}$	0.81$_{\uparrow 0.23}$	0.17$_{\uparrow 0.10}$	0.21$_{\uparrow 0.17}$
Janus [60]	0.61$_{\uparrow 0.29}$	0.97$_{\uparrow 0.26}$	0.68$_{\uparrow 0.34}$	0.30$_{\uparrow 0.09}$	0.84$_{\uparrow 0.26}$	0.46$_{\uparrow 0.39}$	0.42$_{\uparrow 0.38}$
JanusFlow [35]	0.63$_{\uparrow 0.31}$	0.97$_{\uparrow 0.26}$	0.59$_{\uparrow 0.25}$	0.45$_{\uparrow 0.24}$	0.83$_{\uparrow 0.25}$	0.53$_{\uparrow 0.46}$	0.42$_{\uparrow 0.38}$
Janus-Pro-7B [5]	0.80$_{\uparrow 0.48}$	0.99$_{\uparrow 0.28}$	0.89$_{\uparrow 0.55}$	0.59$_{\uparrow 0.38}$	0.90$_{\uparrow 0.32}$	0.79$_{\uparrow 0.72}$	0.66$_{\uparrow 0.62}$
GoT [11]	0.64$_{\uparrow 0.32}$	0.99$_{\uparrow 0.28}$	0.69$_{\uparrow 0.35}$	0.67$_{\uparrow 0.46}$	0.85$_{\uparrow 0.27}$	0.34$_{\uparrow 0.27}$	0.27$_{\uparrow 0.23}$
Bagel [8]	0.78$_{\uparrow 0.46}$	0.98$_{\uparrow 0.27}$	0.94$_{\uparrow 0.60}$	0.76$_{\uparrow 0.55}$	0.91$_{\uparrow 0.33}$	0.69$_{\uparrow 0.62}$	0.70$_{\uparrow 0.66}$
GPT-Image-1 [38]	0.84$_{\uparrow 0.52}$	0.99$_{\uparrow 0.28}$	0.92$_{\uparrow 0.58}$	0.85$_{\uparrow 0.64}$	0.92$_{\uparrow 0.34}$	0.75$_{\uparrow 0.68}$	0.61$_{\uparrow 0.57}$
Collaborative AI Systems							
ComfyAgent [67]	0.32	0.69	0.30	0.33	0.50	0.04	0.04
ComfyMind [16] (reproduced)	0.90$_{\uparrow 0.58}$	1.00$_{\uparrow 0.31}$	1.00$_{\uparrow 0.70}$	0.96$_{\uparrow 0.63}$	0.97$_{\uparrow 0.47}$	0.63$_{\uparrow 0.59}$	0.81$_{\uparrow 0.77}$
SymbOmni (Ours)	0.98$_{\uparrow 0.66}$	1.00$_{\uparrow 0.31}$	1.00$_{\uparrow 0.70}$	0.99$_{\uparrow 0.66}$	0.98$_{\uparrow 0.48}$	0.97$_{\uparrow 0.93}$	0.95$_{\uparrow 0.91}$

Table 2: Evaluation of T2I generation on GenEval [14]. Obj.: Object. Attr.: Attribute. We highlight the best results with colored boxes.

Method	Avg Tokens/Case			Avg/Case
	Input↓	Output↓	Total↓	Requests↓
Ours (w/o EXP)	18,556.1	3,171.0	21,727.1	9.85
Ours	12,463.4 ↓32.8%	1,898.0 ↓40.1%	14,361.4 ↓33.9%	7.25 ↓26.4%

Table 3: Performance and cost statistics on ReasonEdit benchmark. We highlight the best results with colored boxes.

Attribute Binding. These results underscore the effectiveness of symbolic concept learning in enabling robust and precise compositional reasoning.

Additional Benchmarks and Ablations. We further evaluate our method on GenEval2 [25] and Kris Bench [63], and include additional ablation studies in the supplementary material to assess the generalization of our method across diverse benchmarks and evaluation settings.

4.4 Ablation Study

We ablate the symbolic concept memory by removing the memory retrieval module (without EXP). As shown in Table 3, removing memory retrieval consistently degrades performance across all evaluation metrics [2, 27, 58]. More notably, the memory module reduces total token consumption by 33.9%, including reductions of 32.8% in input tokens and 40.1% in output tokens, while decreasing the number of API requests per case by 26.4%. By reusing accumulated experience, the system can identify effective strategies earlier, thereby avoiding failed attempts

and reducing iterative replanning [15, 24, 59]. These findings establish symbolic concept memory as a key mechanism for experience reuse, enabling the continual accumulation of task-level knowledge beyond what can be achieved through in-context learning or parametric recall alone [3, 21, 49].

4.5 Offline Mode Performance Evaluation

To reduce inference cost and alleviate the context burden from lengthy workflow descriptions, we design an offline evaluation where the agent operates solely on retrieved symbolic concepts without access to external workflow descriptions.

We evaluate SymbOmni on the 200-task ComfyBench benchmark [67], comprising 100 vanilla, 60 complex, and 40 creative tasks. The tasks are partitioned into 10 groups of 20, with even-indexed tasks in each group assigned to the online set and odd-indexed tasks to the offline set, yielding 100 tasks per split while maintaining comparable difficulty. We shuffle the groups to disrupt the original monotonic ordering while preserving task similarity within each group. During the online phase, SymbOmni processes the groups sequentially, executing the 10 online tasks in each group in parallel. Experiences are committed to memory only after all tasks in the group have been completed, preventing information leakage within the group. The system then transitions to the offline phase, where it solves the 100 held-out tasks using only the accumulated symbolic concept memory, without access to workflow descriptions. We compare against ComfyMind [16], which retains full access to workflow descriptions but does not maintain an experience memory. This comparison directly tests whether learned symbolic concepts can substitute for explicit documentation. Our analysis focuses on the more challenging Complex category, with additional protocol details provided in the supplementary material.

As shown in Table 4, our offline approach achieves a resolve rate of 70.0%, surpassing the documentation-dependent baseline. More importantly, it reduces the average number of attempts by 28.4% across all tasks and by 25.6% among successfully resolved tasks. These results show that symbolic concept memory not only compensates for the absence of explicit workflow documentation, but also improves problem-solving efficiency through experience-driven optimization. This highlights its practical value in lowering inference costs and mitigating the context overhead associated with lengthy workflow descriptions [51, 61, 65].

4.6 User Study

Using the generation results from the 38 cases presented in Section 4.3, we conduct a large-scale user preference study to compare the evaluated methods. We employ a pairwise blind evaluation protocol, in which SymbOmni was compared with each of the four other methods across all 38 cases, resulting in 152 comparison questions. In total, 63 participants contributed 1,197 preference judgments, assessing each pair along four dimensions: semantic consistency, visual quality, compositional accuracy, and overall preference. Additional details of the study design and evaluation protocol are provided in the supplementary material.

Method	Resolved↑	Avg Tries↓	Avg Tries (Resolved)↓
ComfyMind (w/ description)	66.7%	2.000	1.600
Ours (Offline)	70.0% ↑3.3%	1.433 ↓0.567	1.190 ↓0.410

Table 4: Offline mode performance on the Complex offline split ($N = 30$).

Method	Van. Pass	Van. Resolve	Cplx. Resolve	Crtv. Resolve	Overall
ComfyMind [16]	100.0%	84.0%	55.0%	42.5%	67.0%
SymbOmni	100.0%	**89.0%**	**66.7%**	**52.5%**	**75.0%**
Improvement	+0.0%	+5.0%	+11.7%	+10.0%	+8.0%

Table 5: Generalization on out-of-domain workflows and ComfyBench.

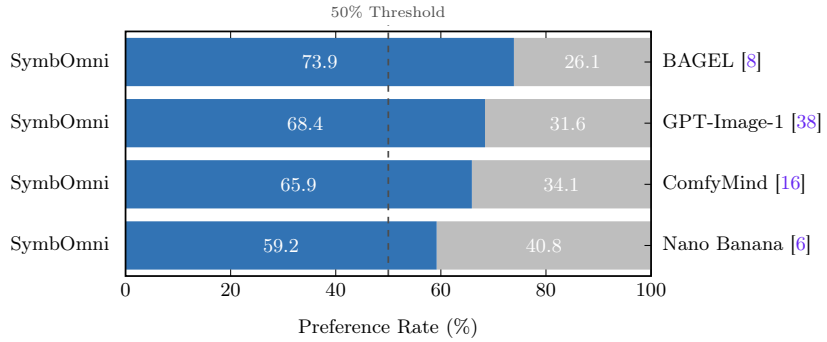


Fig. 6: Pairwise preference rates of SymbOmni vs. other methods.

As shown in Figure 6, SymbOmni is consistently preferred across all pairwise comparisons. It achieves a preference rate of 59.2% over Nano Banana, with the margin increasing to 65.9% against ComfyMind and 68.4% against GPT-Image-1. These results highlight the advantages of symbolic concept learning over heuristic search and purely end-to-end generation paradigms. The largest preference margin is observed against BAGEL, where SymbOmni attains 73.9%, demonstrating the effectiveness of combining agentic workflow decomposition with symbolic memory for compositional reasoning and creative generation.

4.7 Out-of-Domain Generalization

To evaluate whether the learned symbolic concepts generalize beyond the source distribution, we construct an external concept library from 147 in-the-wild workflows spanning diverse domains, following the data collection protocol of ComfyGPT [22]. These workflows encompass a broad range of visual generation and editing paradigms and are entirely disjoint from the original training and evaluation pipeline. In practice, SymbOmni performs symbolic concept learning on this external workflow set and is subsequently evaluated on ComfyBench [67], which can test whether concepts acquired from substantially different workflow distributions can transfer to structured benchmark tasks or not.

As shown in Table 5, SymbOmni achieves consistent improvements across all difficulty levels. The overall resolve rate improves from 67.0% to 75.0% (a 8.0% gain), with the most pronounced gains in harder categories, including Complex +11.7% (55.0% → 66.7%) and Creative +10.0% (42.5% → 52.5%). These results demonstrate that symbolic concepts extracted from diverse, in-the-wild workflows are able to transfer effectively to structured benchmark tasks, confirming the ability of SymbOmni's concept learning mechanism to generalize beyond the workflow distributions encountered during evolution.

We further evaluate SymbOmni across a diverse set of additional benchmarks, and SymbOmni can still consistently achieve state-of-the-art or competitive performance. Our cross-modal generalization analysis also demonstrates that the learned symbolic concepts transfer effectively across different combinations of modalities. Complete results are provided in the supplementary material.

5 Concluding Remarks

We introduce **SymbOmni**, a cognitive architecture for multimodal creation that addresses the perpetual novice problem through Symbolic Concept Learning. By abstracting experience into reusable concepts through an Induction-Transduction cycle guided by verbalized backpropagation, SymbOmni enables continual learning and self-improvement without parameter updates. Extensive experiments demonstrate SymbOmni's strong performance, improved problem-solving efficiency, and continuous gains from accumulated experience, including over 40% reduction in token consumption. Its dual-feedback mechanism further enables the system to learn from both successful and failed executions, providing a principled foundation for adaptive AI systems that continuously refine their behavior. By integrating symbolic reasoning with experiential learning, this work offers a promising path toward scalable and sustainable AI evolution.

References

1. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023) 1, 10
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 10, 12
3. Cemri, M., Pan, M.Z., Yang, S., Agrawal, L.A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., et al.: Why do multi-agent LLM systems fail? Advances in Neural Information Processing Systems **38** (2026) 13
4. Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.M., Yu, H., Lu, Y., Hung, Y.H., Qian, C., et al.: AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In: International Conference on Learning Representations. vol. 2024, pp. 20094–20136 (2024) 2

5. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-Pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025) 12
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 8, 10, 11, 14
7. Comfy-Org: ComfyUI (2026), https://github.com/Comfy-Org/ComfyUI, accessed: 2026-06-26 4, 8
8. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 8, 12, 14
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) 12
10. Fan, S., Cong, X., Fu, Y., Zhang, Z., Zhang, S., Liu, Y., Wu, Y., Lin, Y., Liu, Z., Sun, M.: WorkflowLLM: Enhancing workflow orchestration capability of large language models. In: International Conference on Learning Representations. vol. 2025, pp. 24498–24525 (2025) 2, 3, 11
11. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: GoT: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv preprint arXiv:2503.10639 (2025) 12
12. Fu, T.J., Hu, W., Du, X., Wang, W., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. In: International Conference on Learning Representations. vol. 2024, pp. 54820–54833 (2024) 1, 10
13. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: SEED-X: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024) 12
14. Ghosh, D., Hajishirzi, H., Schmidt, L.: GenEval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023) 8, 12
15. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) 4, 13
16. Guo, L., Xu, X., Wang, L., Lin, J., Zhou, J., Zhang, Z., Su, B., Chen, Y.: ComfyMind: Toward general-purpose generation via tree-based planning and reactive feedback. Advances in Neural Information Processing Systems **38**, 45128–45164 (2026) 8, 10, 11, 12, 13, 14
17. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X.: Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680 (2024) 2
18. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: LTX-Video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024) 1
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 1
20. Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Yau, S., Lin, Z., Zhou, L., et al.: MetaGPT: Meta programming for a multi-agent col-

laborative framework. In: International Conference on Learning Representations. vol. 2024, pp. 23247–23275 (2024) 2, 3
21. Hu, S., Lu, C., Clune, J.: Automated design of agentic systems. In: International Conference on Learning Representations. vol. 2025, pp. 21344–21377 (2025) 13
22. Huang, O., Ma, Y., Zhao, Z., Wu, M., Ji, J., Zhang, R., Hu, Z., Sun, X., Ji, R.: ComfyGPT: A self-optimizing multi-agent system for comprehensive ComfyUI workflow generation. arXiv preprint arXiv:2503.17671 (2025) 2, 3, 8, 14
23. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: SmartEdit: Exploring complex instruction-based image editing with multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8362–8371 (2024) 8, 11
24. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: OpenAI o1 System Card. arXiv preprint arXiv:2412.16720 (2024) 4, 13
25. Kamath, A., Chang, K.W., Krishna, R., Zettlemoyer, L., Hu, Y., Ghazvininejad, M.: GenEval 2: Addressing benchmark drift in text-to-image evaluation. arXiv preprint arXiv:2512.16853 (2025) 12
26. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: HunyuanVideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024) 1
27. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in neural information processing systems **33**, 9459–9474 (2020) 3, 10, 12
28. Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., Zhang, P., Dou, Z.: Search-o1: Agentic search-enhanced large reasoning models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 5420–5438 (2025) 8
29. Li, X., Zou, H., Liu, P.: ToRL: Scaling tool-integrated RL. arXiv preprint arXiv:2503.23383 (2025) 2
30. Li, Y., Liu, Z., Li, Z., Zhang, X., Xu, Z., Chen, X., Shi, H., Jiang, S., Wang, X., Wang, J., et al.: Perception, reason, think, and plan: A survey on large multimodal reasoning models. arXiv preprint arXiv:2505.04921 (2025) 2
31. Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., Liu, H., Zhang, S., Song, K., Zhu, K., et al.: Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. arXiv preprint arXiv:2504.01990 (2025) 8
32. Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with blockwise ringattention. In: International Conference on Learning Representations. vol. 2025, pp. 45953–45977 (2025) 12
33. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al.: AgentBench: Evaluating LLMs as agents. In: International Conference on Learning Representations. vol. 2024, pp. 52989–53046 (2024) 8
34. Liu, Y., Cao, J., Li, Z., He, R., Tan, T.: Breaking mental set to improve reasoning through diverse multi-agent debate. In: The Thirteenth International Conference on Learning Representations (2025) 2, 3, 11
35. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. pp. 7739–7751 (2025) 1, 4, 12
36. Niu, B., Song, Y., Lian, K., Shen, Y., Yao, Y., Zhang, K., Liu, T.: Flow: Modularized agentic workflow automation. arXiv preprint arXiv:2501.07834 (2025) 11
37. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Meng, F., Ning, K., et al.: WISE: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025) 8
38. OpenAI: gpt-image-1 (2025), https://platform.openai.com/docs/models/gpt-image-1, accessed: 2026-06-26 2, 4, 8, 12, 14
39. OpenAI: Introducing Deep Research (2025), https://openai.com/index/introducing-deep-research/, accessed: 2025-02-02 2, 3
40. OpenAI: Introducing OpenAI o3 and o4-mini (2025), https://openai.com/index/introducing-o3-and-o4-mini/, accessed: 2025-04-18 2, 3
41. Patil, S.G., Zhang, T., Wang, X., Gonzalez, J.E.: Gorilla: Large language model connected with massive APIs. Advances in Neural Information Processing Systems **37**, 126544–126565 (2024) 8
42. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024) 12
43. Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., et al.: ChatDev: Communicative agents for software development. In: Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 15174–15186 (2024) 2, 3
44. Qiao, S., Fang, R., Qiu, Z., Wang, X., Zhang, N., Jiang, Y., Xie, P., Huang, F., Chen, H.: Benchmarking agentic workflow generation. In: International Conference on Learning Representations. vol. 2025, pp. 69679–69703 (2025) 8
45. Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., Schölkopf, B.: Controlling text-to-image diffusion by orthogonal finetuning. Advances in Neural Information Processing Systems **36**, 79320–79362 (2023) 10
46. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 12
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1, 4, 12
48. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. Advances in neural information processing systems **36**, 68539–68551 (2023) 8
49. Shang, Y., Li, Y., Zhao, K., Ma, L., Liu, J., Xu, F., Li, Y.: AgentSquare: Automatic LLM agent search in modular design space. In: International Conference on Learning Representations. vol. 2025, pp. 3841–3865 (2025) 13
50. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. Advances in neural information processing systems **36**, 8634–8652 (2023) 2, 3, 8
51. Significant Gravitas: AutoGPT (2023), https://github.com/Significant-Gravitas/AutoGPT, accessed: 2026-06-26 2, 3, 13

52. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024) 12
53. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024) 1, 4
54. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., Zou, J.Y.: Mixture-of-agents enhances large language model capabilities. In: International Conference on Learning Representations. vol. 2025, pp. 33944–33963 (2025) 2, 3, 8
55. Wang, Q., Wang, Z., Su, Y., Tong, H., Song, Y.: Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6106–6131 (2024) 11
56. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024) 1, 12
57. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022) 10
58. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022) 10, 12
59. de Winter, J.C., Dodou, D., Eisma, Y.B.: System 2 thinking in OpenAI's o1-preview model: Near-perfect performance on a mathematics exam. Computers **13**(11), 278 (2024) 2, 13
60. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025) 12
61. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al.: AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155 (2023) 2, 3, 4, 13
62. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: VILA-U: a unified foundation model integrating visual understanding and generation. In: International Conference on Learning Representations. vol. 2025, pp. 93620–93638 (2025) 2
63. Wu, Y., Li, Z., Hu, X., Ye, X., Zeng, X., Yu, G., Zhu, W., Schiele, B., Yang, M.H., Yang, X.: KRIS-Bench: Benchmarking next-level intelligent image editing models. Advances in Neural Information Processing Systems **38** (2026) 12
64. Xiao, T.Z., Bamler, R., Schölkopf, B., Liu, W.: Verbalized machine learning: Revisiting machine learning with language models. Transactions on Machine Learning Research (2025) 3, 7
65. Xie, T., Zhou, F., Cheng, Z., Shi, P., Weng, L., Liu, Y., Hua, T.J., Zhao, J., Liu, Q., Liu, C., et al.: OpenAgents: An open platform for language agents in the wild. arXiv preprint arXiv:2310.10634 (2023) 2, 3, 13
66. Xu, Z., Shi, S., Hu, B., Yu, J., Li, D., Zhang, M., Wu, Y.: Towards reasoning in large language models via multi-agent peer review collaboration. arXiv preprint arXiv:2311.08152 (2023) 2, 11
67. Xue, X., Lu, Z., Huang, D., Wang, Z., Ouyang, W., Bai, L.: ComfyBench: Benchmarking LLM-based agents in ComfyUI for autonomously designing collaborative AI systems. In: Proceedings of the computer vision and pattern recognition conference. pp. 24614–24624 (2025) 2, 3, 8, 10, 12, 13, 14

68. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: ReAct: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022) 2, 4, 8
69. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023) 1, 4, 10
70. Yu, Z., Yuan, Y., Xiao, T.Z., Xia, F.F., Fu, J., Zhang, G., Lin, G., Liu, W.: Generating symbolic world models via test-time scaling of large language models. Transactions on Machine Learning Research (2025) 3
71. Yuksekgonul, M., Bianchi, F., Boen, J., Liu, S., Huang, Z., Guestrin, C., Zou, J.: Textgrad: Automatic "differentiation" via text. arXiv preprint arXiv:2406.07496 (2024) 3
72. Zhang, J., Xiang, J., Yu, Z., Teng, F., Chen, X., Chen, J., Zhuge, M., Cheng, X., Hong, S., Wang, J., et al.: AFlow: Automating agentic workflow generation. In: International Conference on Learning Representations. vol. 2025, pp. 34040–34077 (2025) 3, 11
73. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023) 1, 4, 10
74. Zhao, Y., Yin, H., Zeng, B., Wang, H., Shi, T., Lyu, C., Wang, L., Luo, W., Zhang, K.: Marco-o1: Towards open reasoning models for open-ended solutions. arXiv preprint arXiv:2411.14405 (2024) 3
75. Zheng, Y., Fu, D., Hu, X., Cai, X., Ye, L., Lu, P., Liu, P.: DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 414–431 (2025) 3, 8