Think Bright, Diffuse Nice: Enhancing T2I-ICL via Inductive-Bias Hint Instruction and Query Contrastive Decoding

Anonymous ACL submission

Abstract

Text-to-Image In-Context Learning (T2I-ICL) enables customized image synthesis via interleaved text-image examples but faces two mutually reinforcing bottlenecks, compliance failure and prior-dominated hallucination, that form a vicious cycle degrading generation quality. Existing methods rely on tailored training, which limits flexibility and raises deployment costs. To address these challenges effectively, we propose TBDN, a training-free framework integrating two complementary closed-loop mechanisms: Hint Instruction (HI) and Query Contrastive Decoding (QCD). HI injects task-aware inductive bias via lightweight prompt engineering to anchor models on contextual mapping rules, thereby mitigating compliance failure. QCD adjusts the decoding distributions of language models by contrasting full-input and query-omitted distributions, suppressing prior-dominated hallucination. TBDN achieves State-of-the-Art performance on CoBSAT and Text-to-Image Fast Mini-ImageNet, with robust generalization across model backbones, prompt designs, and hyperparameters. It also maintains promising performance in concept preservation and prompt following on Dreambench++. By breaking the two bottlenecks, TBDN establishes a simple yet effective framework for efficient and reliable T2I-ICL.

1 Introduction

Diffusion models have emerged as the mainstream paradigm in image generation, with numerous related methods proposed and widely deployed in educational and industrial settings (Yang et al., 2023). These methods generate corresponding visual outputs from either a single textual prompt or a text-image pair. Yet certain concepts requiring visualization are difficult to articulate adequately via these inputs. For example, a designer may need to generate a "desert-themed cow" (combining two unrelated concepts) for a marketing campaign, or a teacher may want to visualize a "purple watercolor apple" (modifying attributes and style) to explain color theory. To convey such concepts, humans often rely on a series of interleaved text-image examples (referred to as contexts (Baldassini et al., 2024) or demonstrations (Mi et al., 2025)). Addressing this visualization need requires methods to adapt to such complex input patterns and perform semantic reasoning over the input content to produce coherent results, a research challenge formally defined as the Text-to-Image In-Context Learning (T2I-ICL) task (Zeng et al., 2024).

Existing unified multimodal large language models (unified MLLMs) exhibit in-context learning (ICL) capabilities comparable to those of large language models (LLMs) (Sun et al., 2024b; Ge et al., 2024a), while extending such abilities to multimodal understanding and generation within a unified architectural framework (Sun et al., 2024a). However, recent studies (Zeng et al., 2024; Zong et al., 2025) have revealed that unified MLLMs struggle to effectively leverage their reasoning capabilities in T2I-ICL without tuning. To achieve robust reasoning and higher-fidelity generation results, another line of work (Mi et al., 2025) adopts an intuitively effective paradigm: integrating large vision-language models (LVLMs) with diffusion models. This paradigm aims to harness the strong reasoning capabilities of LVLMs alongside the superior fidelity and semantic controllability of diffusion models to attain desired T2I-ICL performance.

Specifically, LVLMs first process the interleaved contexts to extract semantic rules, and their output y is then fed into diffusion models for image generation. Methods under this paradigm fall into two categories: (1) injecting the latent representations of y into visual decoder for decoding (Sun et al., 2024b; Mi et al., 2025); (2) feeding the textual outputs (or tokens in y) into the image generator to complete encoding and decoding (Liao et al., 2025). The former achieves unified input processing but

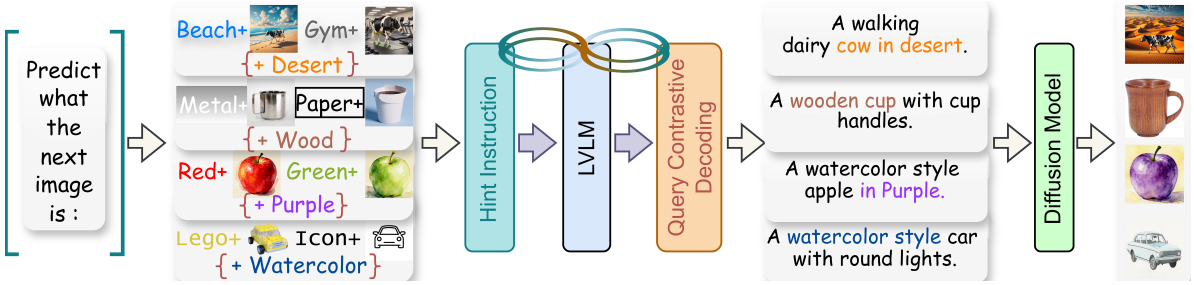

Figure 1: Overview of the TBDN framework for T2I-ICL. Given an instruction (enclosed in '[ ]'), context examples, and queries (enclosed in '( )'), TBDN first injects task-aware inductive bias via Hint Instruction, then refines outputs via Query Contrastive Decoding to suppress priors, and finally generates images with a diffusion model.

often suffers from representation drift, requiring further training and incurs substantial alignment costs. By contrast, the latter is more direct and interpretable, yet relevant explorations remain scarce and lack systematic design principles.

Recalling that the core requirement of T2I-ICL is that, given a few demonstrations, the method should infer underlying mapping rules, generate corresponding descriptive guidance, and perform visualization accordingly. However, both literature and our empirical analysis reveal two key bottlenecks that degrade performance: (†) **Compliance failure**; (††) **Prior-dominated hallucination**. These bottlenecks form a vicious cycle. When a method fails to discern mapping rules, it relies more on prior knowledge to satisfy generation demands. In turn, prior-dominated generation consumes attentional resources (Marinescu et al., 2025), further hindering the method from mining contextual mapping rules and exacerbating compliance failures.

To break this cycle and address two bottlenecks effectively, we propose **TBDN** (Fig. 1), a textual-output-driven framework integrating two complementary, mutually reinforcing mechanisms that form a *Möbius Band*-like closed-loop constraint:

- **Hint Instruction (HI)**: A prompt engineering strategy that injects task-aware inductive bias to resolve compliance failure (Fig. 4).
- **Query Contrastive Decoding (QCD)**: A decoding approach that imposes posterior instruction-following constraints to eliminate prior-dominated hallucination (Fig. 6).

HI appends a priori guidance to the input instruction, directing the LVLM to prioritize the input query. It anchors the model's focus on mapping rules rather than superficial mimicry, addressing compliance failure at the root via targeted inductive bias. Complementing HI, QCD refines generation distributions to suppress prior over-reliance and amplify query-aligned knowledge, neutralizing hallucination by aligning output with contextual rules.

To validate the effectiveness of TBDN, we conduct comprehensive experiments across benchmarks and shot settings, where it achieves state-of-the-art performance in most cases. It further demonstrates strong generalization across LVLM backbones, with pronounced performance gains from the synergistic integration of HI and QCD. Most notably, TBDN is training-free, outperforming existing methods relying on cumbersome training while maintaining superior performance.

Our contributions in T2I-ICL are as follows:

- We identify two mutually exacerbating bottlenecks (compliance failure, prior-dominated hallucination) and their vicious cycle, clarifying a principled method design direction.
- We propose TBDN[1], a training-free framework with two complementary, closed-loop mechanisms (HI / QCD) that resolve these two bottlenecks respectively.
- TBDN achieves strong cross-benchmark performance: it reaches State-of-the-Art on CoBSAT and Text-to-Image Fast Mini-ImageNet, with robust generalization across LVLM backbones, prompt designs and hyperparameters. On Dreambench++, it further demonstrates promising prompt-following capability.

2 Related Work

2.1 Diffused Image Generation

Multimodal image generation aims to generate images conditioned on textual descriptions and, op-

[1] https://github.com/Calendula597/TBDN

tionally, reference images. As promising methods in this field, diffusion models excel at reconstruction fidelity. Representative methods such as DALL-E (Ramesh et al., 2021) and Stable Diffusion (Rombach et al., 2022; Podell et al., 2023) leverage language models to process textual inputs, laying the semantic foundation for downstream generation. To control the structural and stylistic attributes, approaches such as ControlNet (Zhang et al., 2023), T2I-Adapter (Mou et al., 2024), and FLUX (Labs, 2024) incorporate multimodal encoders to capture semantic information from inputs. Beyond text-only or image-text pair inputs, interleaved inputs, which contain multiple images and text, offer richer and more faithful representations of user intention. Recent approaches have attempted to interpret and leverage them for image generation (Yang et al., 2024; Hsiao et al., 2025).

2.2 Text-to-Image In-Context Learning

LLMs (Brown et al., 2020; Wei et al., 2023), LVLMs (Liu et al., 2023a; Alayrac et al., 2022; Wang et al., 2024a), and MLLMs (Koh et al., 2023; Sun et al., 2024b) have exhibited remarkable ICL performance (Dong et al., 2024). To unify image generation with ICL, CoBSAT (Zeng et al., 2024) formalizes the T2I-ICL task. This work not only provides a valuable benchmark but also demonstrates that fine-tuning and prompt engineering are effective strategies for language models in this task. Building on the fine-tuning paradigm, ThinkDiff (Mi et al., 2025) trains an aligner network to transfer multimodal in-context reasoning capabilities from VLMs to diffusion models. It conducts a captioning training task to align the representation space of two foundational modules (VLM and diffusion decoder). Integrating these insights, researchers (Liao et al., 2025) introduce a chain-of-thought dataset (ImageGen-CoT) and fine-tune SEED-X (Ge et al., 2024b) and SEED-LLaMA (Ge et al., 2024a) on it, encouraging language models to generate textual analysis prior to image generation. Despite their effectiveness, these methods typically rely on significant data and computational resources, which restricts their deployment.

3 Two Bottlenecks in T2I-ICL

Our empirical practice shows many T2I-ICL methods fail to recognize mapping rules in contexts and instead capture superficial features, which manifests as context parroting, a mechanical repetition

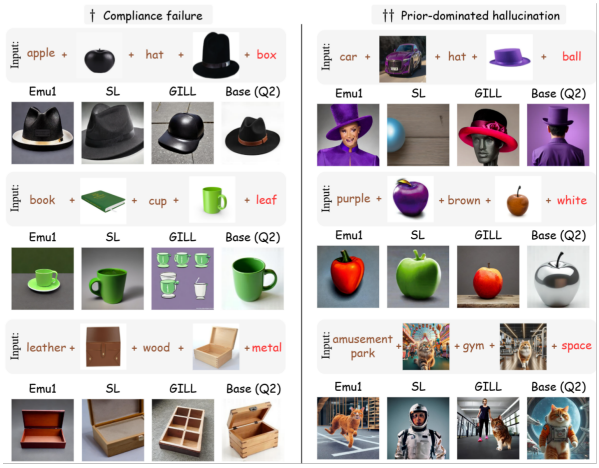

Figure 2: Two critical bottlenecks in T2I-ICL (evaluated on CoBSAT). Compliance failure (left): methods parrot input context (e.g., "hat", "cup") instead of reasoning query semantics. Prior-dominated hallucination (right) methods generate prior-aligned outputs (e.g., "red/green apples") that violate input requirements.

of input context (Zhang and Gilpin, 2025). We term this phenomenon **Compliance failure**. For methods equipped with language models, strong linguistic and visual priors from pre-trained data carry excessive weight during generation (Shi et al., 2024), suppressing input requirements and inducing prior-aligned yet context-violating outputs. We term this **Prior-dominated hallucination**.

Fig. 2 illustrates representative cases of these two bottlenecks. As shown on the left, methods repeat concepts in the input context (e.g., parroting "hat", "cup", and "wood") rather than reasoning according to the input query. On the right, methods exhibit obvious prior-dominated hallucination: concepts like "hat" and "space" are associated with "human", while "apple" (or other objects of similar shape) is associated with the "red" and "green" attributes. Such issues can be further exemplified, and we present more illustrative cases in the appendix.

Prior-dominated hallucination is intuitive and prevalent, while compliance failure is relatively obscure. To clarify this concept and formally measure it, we draw on the framework of CoBSAT, which evaluates model responses along two dimensions (object and attribute) and define an error count metric. This metric counts the number of samples that satisfy at least one of the following conditions: (1) the predicted attribute matches the ground-truth attribute, while the predicted object appears in the input context; (2) the predicted object matches the ground-truth object, while the predicted details ap-

pear in the input context. Using this metric, we quantify the error counts of different methods over 10k samples on CoBSAT, revealing consistent failure patterns (Fig. 3).

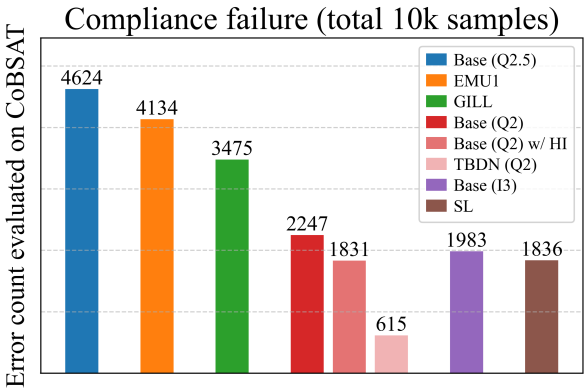

Figure 3: Compliance failure error counts across methods on CoBSAT.

4 Method

4.1 Overview

TBDN is a simple yet effective training-free T2I-ICL framework unifying LVLMs and diffusion models, addressing two core bottlenecks: compliance failure and prior-dominated hallucination. Following the "Think Bright, Diffuse Nice" philosophy, it employs two non-redundant mechanisms, Hint Instruction and Query Contrastive Decoding, to form a closed loop for rule alignment and prior suppression. The workflow of TBDN has five stages: (1) Pre-processing: instruction X_{ins}, interleaved text-image context X_{con}, and query X_{que} are concatenated as a unified multimodal sequence; (2) Injection: hint prompt is added at the end of X_{ins} to anchor contextual mapping rules (Fig. 4); (3) Reasoning: the injected input is fed into the LVLM and produce corresponding distribution; (4) Decoding: two distinct distributions P_{sub} (conditioned on instruction and contexts) and P_{full} (conditioned on full input) are used to perform weighted contrastive adjustment (Fig. 6); (5) Diffusion: the LVLM's rule-aligned textual output is passed to the diffusion model for high-fidelity generation.

4.2 Hint Instruction

Prior works (Zeng et al., 2024; Zong et al., 2025) have observed that LVLMs struggle to reliably comprehend input semantics and might respond irrelevantly in T2I-ICL. In our empirical observations, we further find that LVLMs tend to disregard input

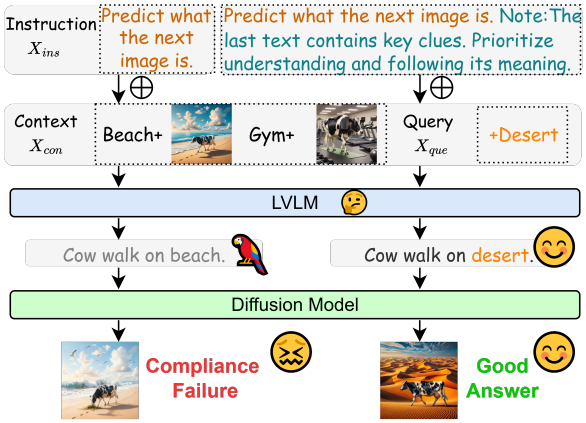

Figure 4: Overview of Hint Instruction, a mechanism guiding LVLMs to prioritize context-aware query reasoning, thereby mitigating compliance failure.

queries and instead recapitulate or even exhibit context parroting (Zhang and Gilpin, 2025). We term such issues collectively as compliance failure. To address compliance failure, tailored training might be effective but incurs prohibitive computational and data costs (Liao et al., 2025). By contrast, prompt engineering (Liu et al., 2023b; Zhou et al., 2022) offers another route by injecting external inductive bias. Previous works (Zeng et al., 2024; Mi et al., 2025) have explored this direction but have failed to achieve satisfactory performance. For instance, Chain-of-Thought (CoT) (Wei et al., 2022) delivers marginal improvements but increases inference cost and risks exceeding length limits.

For effectiveness and efficiency, we propose Hint Instruction (HI), a prompt-based strategy illustrated in Fig. 4. Unlike related works (Zeng et al., 2024; Mi et al., 2025; Liao et al., 2025) that focus on context expansion, HI injects a task-aware inductive bias to directly align the model's reasoning with query semantics. Specifically, based on prior findings (Li et al., 2025) and empirical evaluation, LVLMs tend to ignore or misunderstand the input query, which is typically positioned at the end of the input sequence. This oversight leads to suboptimal performance in T2I-ICL, as the model fails to prioritize the query's guiding role. To counter this issue, we reverse the flaw into a targeted inductive bias: **the query is a vital cue and should take precedence**. Guided by this core bias, we define two design principles for HI's prompt construction: (‡) The query provides key guidance for subsequent generation; (‡‡) The query semantics take precedence even when there is semantic conflict between the query and context. We implement

HI by suffixing an instruction adopted from (Mi et al., 2025) with the following sequence:

> *"The last text I provide contains the most important clue about the next picture. Focus mainly on understanding and following the meaning of the final text when creating your description."*

This specific prompt is selected via extensive ablation experiments, which verify its superiority over alternative designs. Overall, HI augments the LVLM with minimal input overhead by instantiating the aforementioned inductive bias, yielding substantial gains without additional training cost.

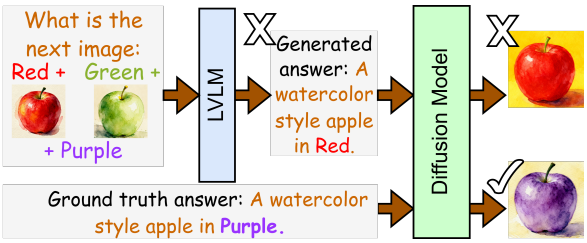

Figure 5: An illustrative example of prior-dominated hallucination in LVLMs for T2I-ICL. Given text-image pairs and the final query, the LVLM exhibits reliance on its prior association (apple ↔ red) to generate an incorrect description, leading the diffusion model to render a red apple that mismatches the semantics of the ground truth answer. This failure motivates our Query Contrastive Decoding (QCD) method, which mitigates such hallucination by adjusting LVM output distributions.

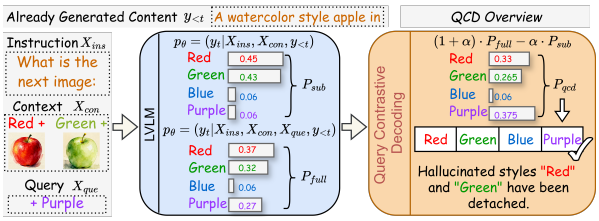

Figure 6: Overview of QCD which adjusts decoding distributions to mitigate prior-dominated hallucinations.

4.3 Query Contrastive Decoding

Apart from the compliance failure, LVLMs are also usually affected by the prior-dominated hallucination (Niu et al., 2021; Li et al., 2023). Fig. 5 presents a concrete case of prior-dominated hallucination, where the LVLM incorrectly associates "apple" with "red" instead of analyzing the target query "purple". To mitigate such problem, we propose Query Contrastive Decoding (QCD) inspired by (Leng et al., 2024; Wang et al., 2024b) that adjust the output distributions of LVLMs (Fig. 6).

Given an input instruction X_{ins}, context X_{con}, and query X_{que}, the input of T2I-ICL can be formally denoted as $X = [X_{\text{ins}}; X_{\text{con}}; X_{\text{que}}]$. The LVLM (parameterized by θ) is expected to autoregressively generate a description Y of the target image, which is a token sequence length with L: $Y = [y_1; y_2; \cdots; y_L]$. Conventionally, each token y_t (where $1 \leq t \leq L$) in Y is sampled from a probability distribution P_{full} during decoding, which is mathematically formulated as:

$$P_{full} = \prod_t p_\theta(y_t \mid X_{ins}, X_{con}, X_{que}, y_{<t}), \quad (1)$$

where $y_{<t}$ denotes tokens generated up to time step $(t-1)$. To implement QCD, we compute a secondary distribution P_{sub} from the same input with the query omitted, which is given by:

$$P_{sub} = \prod_t p_\theta(y_t \mid X_{ins}, X_{con}, y_{<t}). \quad (2)$$

Leveraging P_{sub} and P_{full}, QCD generates the response Y by sampling from the query contrastive distribution P_{qcd}, which amplifies differences between the two distributions via hyperparameter α:

$$P_{qcd} = softmax((1+\alpha) \cdot P_{full} - \alpha \cdot P_{sub}). \quad (3)$$

A larger α strengthens this amplification, and $\alpha = 0$ reverts to regular decoding strategy.

5 Experiments

5.1 Implementation details

Datasets & Evaluation. Three representative T2I-ICL benchmarks serve as our testbeds: CoBSAT (Zeng et al., 2024), Text-to-Image Fast Mini-ImageNet (T2IFMIT) (Tsimpoukelli et al., 2021), and Dreambench++ (Peng et al., 2025). We follow their evaluation protocols, where prediction accuracy is reported for all methods on CoBSAT and T2IFMIT, while the normalized scores for concept preservation and prompt following judged by LLMs are reported for all methods on Dreambench++. Specifically, on CoBSAT, methods receive a task instruction followed by 2 or 4 input image–text pairs (corresponding to the 2-shot and 4-shot settings in our experimental results) and a textual query. Similarly, on T2IFMIT, the 1-shot and 2-shot settings refer to the number of image–text pairs provided per class. Following (Zong et al.,

Method	Object-Inference Task					Attribute-Inference Task					Avg. acc.↑
	Color-I	Bkg.-I	Style-I	Action-I	Texture-I	Color-II	Bkg.-II	Style-II	Action-II	Texture-II	
GILL	.171	.054	.069	.063	.074	.010	.043	.024	.022	.040	.057
Emu1	.065	.051	.057	.052	.078	.062	.109	.081	.092	.074	.072
SL	.616	.216	.272	.592	.112	.088	.168	.192	.220	.056	.254
SL-IGC	.620	.368	.384	.424	.060	.192	.288	.208	.216	.148	.291
SX	.796	.412	.316	.596	.240	.176	.344	.260	.252	.104	.349
ThinkDiff	.622	.349	.237	.459	.290	.511	.534	.340	.534	.292	.417
SX-IGC	.884	.692	**.928**	**.936**	.420	.504	.612	**.660**	.524	.424	.658
Base (Q2)	.766	.719	.299	.498	.426	.501	.769	.338	.686	.366	.536
TBDN (Q2)	.909	**.879**	.500	.760	**.556**	**.724**	.905	.487	.683	.531	**.693**(↑29.2%)
Base (Q2.5)	.354	.407	.127	.392	.223	.231	.442	.213	.522	.217	.312
TBDN (Q2.5)	.843	.594	.347	.624	.459	.534	.794	.409	.589	.441	.563(↑80.1%)
Base (I3)	.826	.760	.317	.711	.432	.385	.866	.442	.701	.429	.586
TBDN (I3)	**.927**	.808	.383	.808	.520	.587	**.956**	.486	**.789**	**.568**	.683(↑16.4%)

Table 1: Comparison of 2-shot accuracy on CoBSAT. "Bkg." denotes background, and "Avg. Acc." is the average accuracy over 10 tasks. **Red** and blue highlight the best and second-best results per task.

Method	Object-Inference Task					Attribute-Inference Task					Avg. acc.↑
	Color-I	Bkg.-I	Style-I	Action-I	Texture-I	Color-II	Bkg.-II	Style-II	Action-II	Texture-II	
GILL	.106	.044	.041	.073	.087	.022	.059	.044	.032	.067	.058
Emu	.063	.018	.045	.048	.097	.037	.122	.109	.077	.088	.070
SL	.482	.211	.141	.053	.122	.252	.076	.268	.207	.105	.192
ThinkDiff	.638	.362	.254	.434	.317	.610	.590	.432	.664	.332	.463
Base (Q2)	.815	.730	.346	.466	.504	.652	.920	.443	.812	.449	.614
TBDN (Q2)	.920	.948	**.582**	.822	**.633**	**.851**	.980	**.550**	.788	.593	.767(↑24.9%)
Base (Q2.5)	.498	.558	.149	.446	.255	.369	.493	.265	.620	.296	.395
TBDN (Q2.5)	.868	.832	.392	.752	.531	.675	.960	.457	.740	.517	.672(↑70.1%)
Base (I3)	.881	.908	.401	.824	.573	.659	.936	.522	.837	.588	.713
TBDN (I3)	**.927**	**.951**	.490	**.875**	.627	.783	**.983**	.547	**.868**	**.636**	**.769**(↑7.8%)

Table 2: Comparison of 4-shot accuracy on CoBSAT. "Bkg." denotes background, and "Avg. Acc." is the average accuracy over 10 tasks. **Orange** and green highlight the best and second-best results per task.

2025), all experiments on T2IFMIT use three independent random seeds, and we report the mean and standard deviation of performance across all methods. For Dreambench++, only a text-image pair is provided. For consistency across experiments, we use the instruction template from (Mi et al., 2025) and prefix HI's prompt token with "Note:".

Baselines. SEED-LLaMA (SL) (Ge et al., 2024a), SEED-X (SX) (Ge et al., 2024b), Emu (Sun et al., 2024b,a), GILL (Koh et al., 2023), Anole (Chern et al., 2024), and ThinkDiff (Mi et al., 2025) are adopted as baselines. SL-IGC and SX-IGC are the abbreviations for SL and SX fine-tuned on the ImageGen-CoT dataset (Liao et al., 2025), respectively. In addition, we define a pipeline (denoted as Base) that combines LVLM with FLUX.1-dev (Labs, 2024), the same visual generator adopted in TBDN. To investigate performance variance introduced by LVLMs, Qwen2-VL-7B-Instruct (Q2) (Wang et al., 2024a), Qwen2.5-VL-7B-Instruct (Q2.5) (Bai et al., 2025), and InternVL3-8B (I3) (Zhu et al., 2025) are evaluated.

Resources & Hyperparameters. Since our proposed methods are **training-free**, they require limited computational resources and are easy to reproduce. Specifically, the peak memory usage of our methods is under 60 GB, which can be supported by either two consumer-grade GPUs (e.g., RTX 5090) or one professional GPU (e.g., A100). We set the sampling temperature in LVLM to 0.7, top-p to 0.9, and the number of inference steps in FLUX to 28. For α in QCD, we set 0.5 as default. Additional results and analysis on the effect of α are provided the Appendix.

5.2 Main results

Results on CoBSAT. We evaluate the performance of TBDN and other baselines on 10 tasks in CoBSAT. Tables 1 and 2 report the results for the 2-shot and 4-shot settings. These results across all settings demonstrate that: (i) the simple pipeline, which is denoted as the Base, surprisingly outperforms most unified MLLMs; (ii) Base (Q2) and Base (I3) outperform ThinkDiff, which shares an analogous architecture with the Base, without further modality alignment; (iii) TBDN achieves the best or second-best performance in most cases across different settings. These results highlight that the paradigm

which incorporates the multimodal reasoning capabilities of LVLMs with diffusion models for T2I-ICL is competitive and efficient. Besides, the effectiveness of TBDN is also demonstrated.

Results on Text-to-Image Fast Mini-ImageNet. The main experimental results on Text-to-Image Fast Mini-ImageNet are shown in Table 3, where the I3 identifier is omitted for the Base and TBDN. Notably, the Base consistently outperforms other baselines across two settings, indicating a strong advantage in the fast binding T2I-ICL task. Besides, augmented by our two proposed mechanisms, the mean performance has shown varying degrees of improvement (34.50 $\rightarrow$ 39.00, 38.17 $\rightarrow$ 39.67) while the standard deviation has decreased accordingly (7.29 $\rightarrow$ 2.25, 5.48 $\rightarrow$ 2.47). Comprehensive results of the Base and the TBDN equipped with various LVLMs are provided in Appendix.

Method	1-shot	2-shot
GILL	16.00 ± 2.27	15.17 ± 2.72
SL-8B	15.00 ± 3.27	12.67 ± 1.18
SL-14B	17.25 ± 2.75	16.75 ± 1.75
Emu1	31.50 ± 1.87	22.83 ± 2.72
Emu2	24.33 ± 3.30	30.67 ± 1.31
Anole-7B	11.00 ± 2.86	7.00 ± 0.71
Base	34.50 ± 7.29	38.17 ± 5.48
+ HI	36.50 ± 1.53	38.00 ± 2.18
TBDN	**39.00 ± 2.25**	**39.67 ± 2.47**

Table 3: Comparison of accuracy on T2IFMIT.

Results on Dreambench++. Table 4 shows the brief comparison results on Dreambench++. Compared to the two aforementioned benchmarks, Dreambench++ tends to measure the ability of methods to capture and processing the visual detail, where TBDN achieve promising scores in prompt following (PF) but underperforms in concept preservation (CP). This result is expected because the ability for CP is inherent in TBDN's visual generator, which is fixed, and the essence of PF aligns with the design of HI and QCD, indicating directions for future improvements. More comprehensive results are provided in the Appendix.

Ablation of HI and QCD. Our ablation results across LVLMs are summarized in Tables 5, 8, and 9. Across settings, HI and QCD consistently enhance the Base method, securing the best and second-best performances in most cases, which demonstrates a direct outcome of their complementary strengths.

Dream bench++	Concept Preservation	Prompt Following	CP·PF ↑
SL	.358	.218	.078
SL-IGC	.325	.310	.101
SX	**.559**	.337	.188
SX-IGC	.458	**.881**	**.403**
TBDN (Q2)	.442	.778	.344

Table 4: Comparison of two LLM-judged scores.

For Qwen2-VL and InternVL3, HI and QCD exhibit distinct non-overlapping advantages. HI excels in Bkg.-I and Action-II tasks while QCD dominates other subtasks. This division of strengths aligns with the complementary concept we introduced earlier. HI and QCD mirror two halves of a *Möbius Band*-like closed-loop constraint in that neither is complete on its own while their integration delivers a continuous and comprehensive solution.

Instruction comparison. As noted in Sec. 4.3, for a given dataset, input differences across methods stem from X_{ins} variations, which induce distinct costs in memory footprint and API call quotas. To evaluate the effectiveness and efficiency of HI, we compare it with 4 instructions: (1) CB-Ins; (2) CoT-Ins; (3) TD-Ins; (4) TD-Ins++. CB-Ins and CoT-Ins are from (Zeng et al., 2024), and TD-Ins from (Mi et al., 2025). TD-Ins++ is TD-Ins suffixed with "Let's think it step by step". We report average accuracy, performance improvement (δ), and average instruction token length across two settings.

Table 6 shows CoT-Ins, TD-Ins, TD-Ins++, and HI all yield performance gains relative to CB-Ins, with HI achieving the most substantial improvement. Notably, CB-Ins, TD-Ins, TD-Ins++, and HI have comparable token lengths, while CoT-Ins exhibits an extremely large token count. This is because CoT-Ins requires inserting in-context samples to construct reasoning chains, drastically raising token usage. As the number of shots increases, this token overhead escalates further and may even exceed the context window limit of some LVLMs (e.g., InternVL3), rendering CoT-Ins impractical for large-shot scenarios. CB-Ins is a prompt variant that explicitly states the object–attribute relationship in the instruction. For instance, for the Color-I task in CoBSAT, CB-Ins must be modified to incorporate the "color signal": "*Please identify the main common object in the images, and generate another image of this object in the requested color.*" While CB-Ins is shorter than HI, it suffers from poor generalization due to the need for task-

Method	Shot	Object-Inference Task					Attribute-Inference Task					Avg. acc.↑
		Color-I	Bkg.-I	Style-I	Action-I	Texture-I	Color-II	Bkg.-II	Style-II	Action-II	Texture-II	
Base (I3)	2	.826	.760	.317	.711	.432	.385	.866	.442	.701	.429	.586
	4	.881	.908	.401	.824	.573	.659	.936	.522	.837	.588	.712
+ HI	2	.813	.739	.225	.627	.385	.285	.878	.334	**.808**	.362	.545(↓7.0%)
	4	.846	.924	.291	.741	.499	.397	.964	.403	**.882**	.497	.644(↓9.6%)
+ QCD	2	.913	.711	**.421**	**.850**	.513	.540	.924	**.522**	.637	.518	.654(↑11.6%)
	4	.908	.921	**.554**	**.907**	**.638**	.726	.981	**.566**	.807	.624	.763(↑7.2%)
+ QCD & HI	2	**.927**	**.808**	.383	.808	**.520**	**.587**	**.956**	.486	.789	**.568**	**.683**(↑16.6%)
	4	**.927**	**.951**	.490	.875	.627	**.783**	**.983**	.547	.868	**.636**	**.768**(↑7.9%)

Table 5: Ablation of HI and QCD with InternVL3 on CoBSAT benchmark. **Red** and blue highlight the best and second-best 2-shot results per task, while **Orange** and green highlight the best and second-best 4-shot results.

Qwen2-VL	Shot	Acc.	δ	Len.
CB-Ins	2	.447	–	42
	4	.353	–	
CoT-Ins	2	.533	.086	2850
	4	.642	.289	5521
TD-Ins	2	.537	.090	48
	4	.614	.261	
TD-Ins++	2	.561	.114	55
	4	–	–	
HI	2	**.601**	**.154**	82
	4	**.673**	**.320**	

Table 6: The average of accuracy and token length about different instructions in CoBSAT.

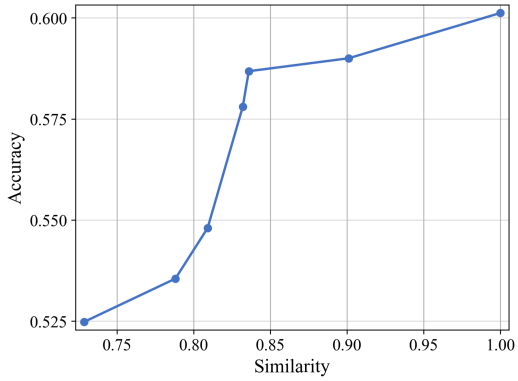

Figure 7: Average accuracy on CoBSAT and sentences similarity between canonical HI prompt and its variants.

specific rewriting. Moreover, LVLMs often fail to correctly interpret CB-Ins and even produce irrelevant responses in our empirical practice. In contrast, HI eliminates the need for task-specific redesign across sub-tasks, simultaneously delivering superior performance efficiently and striking a balance between effectiveness and generalizability. Additionally, as mentioned in Sec. 5.1, HI adopts the instruction template from TD-Ins. Compared with TD-Ins and TD-Ins++, HI achieves substantial improvement with minimal extra token cost.

HI Variant Comparison. A natural concern in HI design is whether prompt variations could lead to performance inconsistency. To address this, we propose and validate a reproducible recipe that incorporates a canonical prompt and a semantic similarity threshold to facilitate reliable prompt adaptation. To instantiate this recipe, we conduct a three-stage implementation as follows. We manually construct an initial prompt following the aforementioned two principles (see Sec. 4.2). We then feed this initial prompt into GPT-4o (OpenAI, 2024) to generate prompt variants, which we subsequently evaluate on CoBSAT to select the candidate with best performing as the canonical prompt of HI. Using sentence-level embeddings generated from BGE-M3 (Chen et al., 2024), we compute dot product similarities between each variant and the canonical prompt. As illustrated in Fig. 7, variants with a similarity ≥ 0.80 consistently achieve an accuracy ≥ 0.54, outperforming the TD-Ins (see Table 6). We report the list of prompt variants and their corresponding accuracies in Table 7 (Appendix).

6 Conclusion

In this paper, we propose TBDN, a training free framework that integrates two complementary enhancement mechanisms in T2I-ICL. By explicitly addressing two recurrent failure patterns identified in our study, compliance failure and prior dominated hallucination, TBDN improves rule following and reduces prior override during in context generation. Across CoBSAT, T2I Fast Mini-ImageNet, and Dreambench++, TBDN consistently achieves strong performance and generalization under diverse settings. Extensive ablation studies further confirm its robustness across model backbones and hyperparameter choices.

7 Limitations

Despite the effectiveness of TBDN, it has several limitations. First, since LVLMs cannot generate images directly, TBDN relies on instructions to guide LVLMs in producing textual descriptions for the T2I generator. This indirect design risks semantic gaps between text and images, unlike MLLMs that enable end-to-end multimodal-to-image generation. Besides, TBDN is less suitable for multimodal image composing tasks (e.g., (Peng et al., 2025)), as our design prioritizes context-aware query reasoning over modeling of fine-grained visual details in reference images. Finally, the generalization of our designs for HI and QCD to MLLMs remains under-explored. Future work will address these gaps by exploring end-to-end multimodal generation paradigms, enhancing fine-grained visual alignment, and extending our designs to MLLMs.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35:23716–23736.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? In *CVPR Workshops*, pages 1539–1550.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024a. Making llama see and draw with seed tokenizer. In *ICLR*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024b. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Teng-Fang Hsiao, Bo-Kai Ruan, Yi-Lun Wu, Tzu-Ling Lin, and Hong-Han Shuai. 2025. Tf-ti2i: Training-free text-and-image-to-image generation via multi-modal implicit-context learning in text-to-image models. In *ICCV*.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2023. Generating images with multimodal language models. *NeurIPS*, 36:21487–21506.

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pages 13872–13882.

Yanshu Li, Jianjiang Yang, Ziteng Yang, Bozheng Li, Ligong Han, Hongyang He, Zhengtao Yao, Yingjie Victor Chen, Songlin Fei, Dongfang Liu, and Ruixiang Tang. 2025. Make lvlms focus: Context-aware attention modulation for better multimodal in-context learning. *Preprint*, arXiv:2505.17097.

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

Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. 2025. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. In *ICCV*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Ioana Marinescu, Kyunghyun Cho, and Eric Karl Oermann. 2025. On the relationship between the choice of representation and in-context learning. *Preprint*, arXiv:2510.08372.

Zhenxing Mi, Kuan-Chieh Wang, Guocheng Qian, Hanrong Ye, Runtao Liu, Sergey Tulyakov, Kfir Aberman, and Dan Xu. 2025. I think, therefore i diffuse: Enabling multimodal in-context reasoning in diffusion models. In *ICLR*.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, volume 38, pages 4296–4304.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2025. Dreambench++: A human-aligned benchmark for personalized image generation. In *ICLR*.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695.

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. 2024. Why larger language models do in-context learning differently? In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024a. Generative multimodal models are in-context learners. In *CVPR*, pages 14398–14409.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024b. Generative pretraining in multimodality. In *ICLR*.

Gemini Team. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *ACL (Findings)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *Preprint*, arXiv:2303.03846.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4).

Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. 2024. Can MLLMs perform text-to-image in-context learning? In *COLM*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models.

Yuanzhao Zhang and William Gilpin. 2025. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. *Preprint*, arXiv:2505.11349.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. 2025. Vl-icl bench: The devil in the details of multimodal in-context learning. In *ICLR*.

A Implementation Details

Fig. 8 illustrates the details of the HI Variant Comparison. The list of prompt variants and their corresponding accuracies is summarized in Table 7.

In Table 6, we compare the performance of different instruction variants on the CoBSAT. CB-Ins and CoT-Ins are from the CoBSAT paper, while TD-Ins is adopted from the ThinkDiff paper. Both HI and TD-Ins++ are derived from TD-Ins: HI uses TD-Ins as the basic instruction, and TD-Ins++ further augments TD-Ins by adding a naive CoT prompt. Below is the list of detailed instructions:

- **CB-Ins (For Color-I task)**: *Please identify the common main object in the images, and describe the next image to be generated based on the sequence below. Your description of image should contain the description of the common main object and the requested color.*

- **CoT-Ins**: *We provide a few examples, each of which is an input-output pair where the output is a description of the image associated with the input.*
  (Multimodal context)
  Based on the examples, the task is to predict the next image description. Before predicting the next image, let's think step by step and analyze what the relationship between the text input and image output in each example is first.
  (Model's response)
  Based on the analysis, please describe what the next image should be look like given the request.

- **TD-Ins**: *I give you several words and pictures. First, please analyse what the next picture is. Then give me a detailed diffusion prompt to describe the next picture. Please only provide me the detailed prompt and start the answer with 'Create an image'.*

- **TD-Ins++**: *I give you several words and pictures. First, please analyse what the next picture is. Then give me a detailed diffusion prompt to describe the next picture. Please only provide me the detailed prompt and start the answer with 'Create an image'.* ***Let's think step by step***.

- **HI**: *I give you several words and pictures. First, please analyse what the next picture is. Then give me a detailed diffusion prompt to describe the next picture. Please only provide me the detailed prompt and start the answer with 'Create an image'.* ***Note: The last text I provide contains the most important clue about the next picture. Focus mainly on understanding and following the meaning of the final text when creating your description.***

B More quantitative results

B.1 More ablation studies of HI and QCD across different LVLMs

Table 8 and 9 present the ablation evaluation of HI and QCD associated with Qwen2-VL and Qwen2.5VL. Taking the result of 2-shot setting as an example, HI increases the performance of Qwen2-VL from 0.537 to 0.601 (+11.9%) and that of Qwen2.5-VL from 0.312 to 0.357 (+14.4%). It suggests that, beyond explicit instructions and prompt-based methods such as chain-of-thought (CoT), incorporating prior knowledge as a hint condition can further enhance the model's adherence to contextual queries, reduce subject-related errors, and thereby improve text-to-image generation accuracy. Generally, compared with HI, QCD yields even larger performance gains. For Qwen2-VL, it achieves 0.638 (+18.8%) and 0.745 (+21.3%) under the 2-shot and 4-shot settings, respectively. For Qwen2.5-VL, the improvements are even more substantial, with scores increasing from 0.312 to 0.554 (+77.6%) in the 2-shot setting and from 0.394 to 0.634 (+60.9%) in the 4-shot setting. These results indicate that QCD can mitigate the language bias of LVLMs induced by in-context information, enabling them to describe the target image more accurately and thus generate more faithful images.

B.2 Effect of α in QCD.

Table 10 summarizes the performance of LVLMs enhanced by QCD with different α values on CoBSAT. Qwen2-VL shows a performance drop when α increases from 0.5 to 1.0, while Qwen2.5-VL achieves peak accuracy at $\alpha = 0.75$. InternVL3 exhibits minimal performance fluctuation over 4 α settings. The optimal performance across three LVLMs is consistently attained at intermediate α values. Based on these observations, we adopt $\alpha = 0.5$ as the default setting for TBDN.

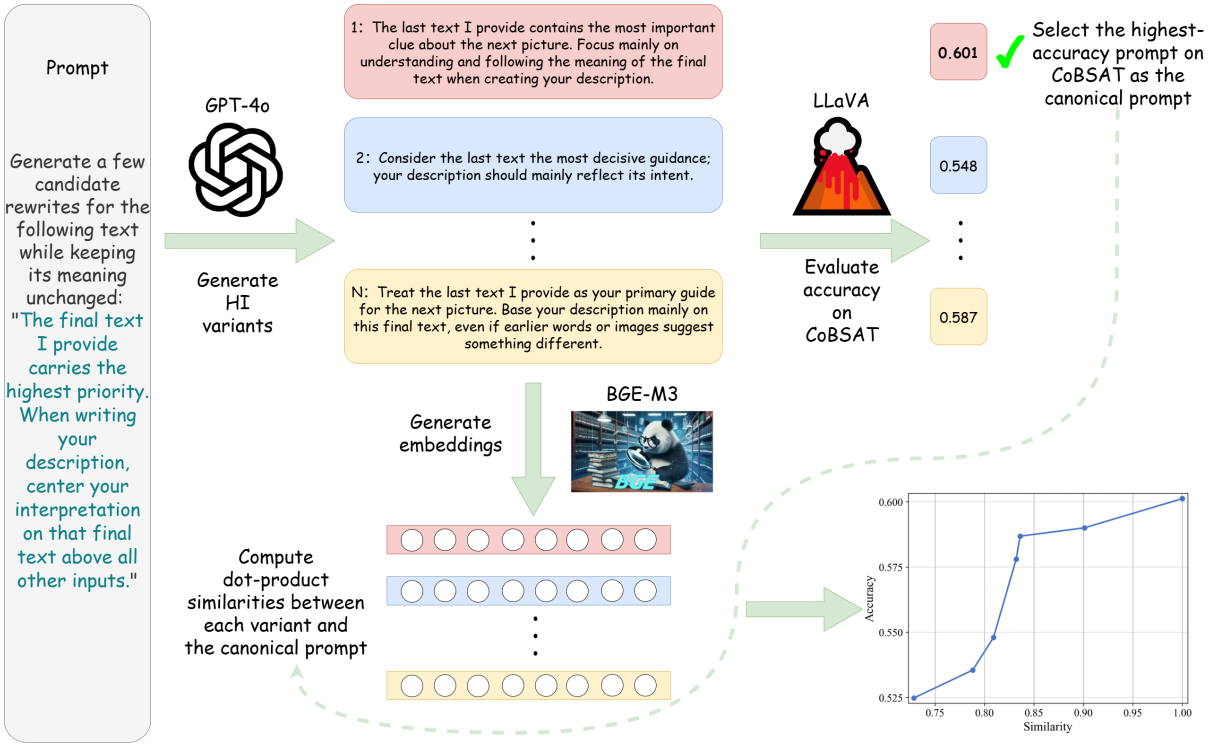

Figure 8: Overview of the pipeline for selecting the canonical Hint Instruction and measuring semantic similarity.

Hint Instructions	Similarity	Avg. Acc.
The last text I provide contains the most important clue about the next picture. Focus mainly on understanding and following the meaning of the final text when creating your description.	1.000	**0.601**
Consider the last text the most decisive guidance; your description should mainly reflect its intent.	0.688	0.548
The concluding text is the key signal—base your description chiefly on it, giving it precedence over the rest of the context.	0.684	0.525
Treat the last text as the decisive guide, and make sure your description is driven mainly by its intent.	0.681	0.536
Treat the last text I provide as your primary guide for the next picture. Base your description mainly on this final text, even if earlier words or images suggest something different.	0.788	0.587

Table 7: Detail hint instructions and corresponding sentences similarity and average accuracy.

Method	Shot	Object-Inference Task					Attribute-Inference Task					Avg. acc.↑
		Color-I	Bkg.-I	Style-I	Action-I	Texture-I	Color-II	Bkg.-II	Style-II	Action-II	Texture-II	
Base (Q2)	2	.766	.719	.299	.498	.426	.501	.769	.338	.686	.366	.537
	4	.815	.730	.346	.466	.504	.652	.920	.443	.812	.449	.614
+ HI	2	.811	.797	.323	.512	.487	.664	.847	.416	**.722**	.433	.601(↑11.9%)
	4	.824	.766	.385	.555	.554	**.832**	.958	.506	**.850**	.497	.673(↑9.6%)
+ QCD	2	.866	.743	.442	.721	.529	.621	.875	.455	.642	.489	.638(↑18.8%)
	4	**.946**	.904	.561	.786	.616	.800	.963	**.552**	.791	.533	.745(↑21.3%)
TBDN	2	**.909**	**.879**	**.500**	**.760**	**.556**	**.724**	**.905**	**.487**	.683	**.531**	**.693**(↑29.1%)
	4	.920	**.948**	**.582**	**.822**	**.633**	**.832**	**.980**	.550	.788	**.593**	**.767**(↑24.9%)

Table 8: Ablation of HI and QCD on Qwen2-VL on CoBSAT benchmark. **Red** and blue highlight the best and second-best 2-shot results per task, while **Orange** and green highlight the best and second-best 4-shot results.

Method	Shot	Object-Inference Task					Attribute-Inference Task					Avg. acc.↑
		Color-I	Bkg.-I	Style-I	Action-I	Texture-I	Color-II	Bkg.-II	Style-II	Action-II	Texture-II	
Base (Q2.5)	2	.354	.407	.127	.392	.223	.231	.442	.213	.522	.217	.312
	4	.498	.558	.149	.446	.255	.369	.493	.265	.620	.296	.394
+ HI	2	.426	.481	.153	.439	.237	.294	.509	.236	.526	.278	.357(↑14.4%)
	4	.619	.590	.201	.521	.327	.536	.628	.345	.686	.393	.484(↑22.8%)
+ QCD	2	.770	.526	.295	**.650**	.424	**.555**	**.840**	.399	**.664**	.425	.554(↑77.6%)
	4	.811	.816	.322	.679	.452	.647	.945	.436	**.745**	.496	.634(↑60.9%)
+ QCD & HI	2	**.843**	**.594**	**.347**	.624	**.459**	.534	.794	**.409**	.589	**.441**	**.563**(↑80.4%)
	4	**.868**	**.832**	**.392**	**.752**	**.531**	**.675**	**.960**	**.457**	.740	**.517**	**.672**(↑70.6%)

Table 9: Ablation of HI and QCD with Qwen2.5-VL on CoBSAT benchmark. **Red** and blue highlight the best and second-best 2-shot results per task, while **Orange** and green highlight the best and second-best 4-shot results.

α	Base (Q2)	Base (Q2.5)	Base (I3)
0.25	.632	.478	.643
0.5	**.638**	.555	**.655**
0.75	.614	**.575**	.653
1.0	.552	.571	.647

Table 10: Average accuracy of LVLMs with different α.

B.3 Text version results on CoBSAT

We present a text-version results on the CoBSAT benchmark, as shown in Table 11. Similar to the image-generation setting defined in CoBSAT, our method delivers consistent improvements across three different LVLMs under both 2-shot and 4-shot configurations in the text-only setting. Meanwhile, our method achieves the best and second-best results, outperforming the strong baseline SX-IGC as well as the proprietary model Gemini (Team, 2025). These results suggest that the high accuracy observed in image-generation mode is grounded in accurate textual descriptions.

In addition, compared with ThinkDiff and several MLLMs, our method offers stronger interpretability. Consistent with the findings in (Liao et al., 2025), however, accuracy in the image-generation mode remains lower than that in text-generation mode, indicating that reliably translating precise textual descriptions into equally accurate images remains a non-trivial challenge.

B.4 Comprehensive results on the Text-to-Image Fast Mini-ImageNet and Dreambench++

Table 12 and Table 13 summarize more comprehensive results on Text-to-Image Fast and Dreambench++ benchmark.

In Text-to-Image Fast Mini-ImageNet, we evaluate the Base and TBDN with three different LVLMs and ablate the HI and QCD. Besides, the results of MLLMs are also compared, demonstrating that our method achieves outstanding performance with different given samples (shot).

In Dreambench++, Animal, Human, Object and Style denote the types of the personalized reference concept for concepts preservation. At the same time, Realistic, Style and Imaginative denote prompt categories with increasing abstraction for prompt following. Each value is the score averaged over prompts in the corresponding subset.

C Qualitative results

C.1 Qualitative ablation on CoBSAT samples

Figure 9, 10, 11, 12 show the qualitative ablation results on four samples in CoBSAT dataset. In these cases, the Base method (see Sec. 5.2) exhibits prior-dominated hallucinations, such as generating "blue sky", "yellow banana", "a cow in the forest", and "a bird in the forest and eating fruits". Such text answers violate the multimodal inputs causing wrong image generation. By explicitly injecting inductive bias, HI encourages the model to attend to the query, thereby producing correct textual outputs. Meanwhile, QCD mitigates the influence of priors by adjusting the output distribution. TBDN combines the strengths of both components and yields correct results. These results justify our observation and further justify the soundness and effectiveness of our method design.

C.2 2-shot evaluation on CoBSAT

Figure 13 and Figure 14 show qualitative results for 2-shot evaluation on the CoBSAT benchmark. For each ground-truth answer, the information that must be inferred from the multimodal inputs is highlighted in blue, while the information explicitly provided is highlighted in brown. As shown in Figure 14, TBDN accurately infers both the target style and subject from the multimodal context across diverse subtasks, thereby generating correct images. We further compare TBDN with GILL (Koh et al.,

Method	Shot	Object-Inference Task					Attribute-Inference Task					Avg. acc.↑
		Color-I	Bkg.-I	Style-I	Action-I	Texture-I	Color-II	Bkg.-II	Style-II	Action-II	Texture-II	
Qwen2-VL												
Base	2	.810	.757	.357	.561	.606	.517	.762	.440	.737	.474	.602
	4	.850	.764	.447	.540	.711	.683	.919	.578	.848	.635	.698
+ HI	2	.848	.851	.430	.618	.668	.697	.834	.541	.778	.550	.682(↑13.3%)
	4	.862	.812	.512	.657	.779	.871	.961	.630	.885	.748	.772(↑10.6%)
+ QCD	2	.910	.764	.583	.829	.739	.655	.872	.587	.699	.637	.728(↑20.9%)
	4	.970	.916	.725	.888	.885	.844	.972	.684	.833	.747	.846(↑21.2%)
+ QCD & HI	2	.933	**.898**	.674	.862	**.789**	**.784**	.896	.640	.725	.706	.791(↑31.4%)
	4	.949	.965	**.754**	.922	**.891**	**.922**	.980	.719	.866	.815	.878(↑25.8%)
Qwen2.5-VL												
Base	2	.363	.439	.215	.483	.316	.239	.440	.270	.543	.275	.358
	4	.517	.586	.222	.532	.346	.386	.494	.332	.650	.396	.446
+ HI	2	.440	.517	.272	.524	.323	.315	.502	.294	.555	.365	.411(↑14.8%)
	4	.647	.628	.289	.626	.473	.558	.630	.431	.702	.508	.549(↑23.1%)
+ QCD	2	.801	.531	.430	.785	.624	.550	.841	.496	.684	.558	.630(↑76.0%)
	4	.834	.832	.452	.812	.666	.670	.951	.559	.813	.650	.724(↑62.3%)
+ QCD & HI	2	.878	.615	.530	.774	.628	.577	.806	.507	.620	.603	.654(↑82.7%)
	4	.915	.862	.605	.892	.752	.703	.960	.576	.785	.688	.774(↑73.5%)
InternVL3												
Base	2	.867	.775	.418	.832	.601	.401	.869	.543	.750	.536	.659
	4	.917	.930	.510	.944	.810	.682	.947	.644	.866	.770	.802
+ HI	2	.857	.780	.343	.751	.544	.309	.884	.449	**.848**	.473	.624(↓5.3%)
	4	.870	.945	.408	.859	.720	.427	.966	.525	.923	.679	.732(↓8.7%)
+ QCD	2	.949	.723	.617	.956	.709	.587	.927	.662	.665	.674	.747(↑13.4%)
	4	.947	.940	.748	**.985**	.875	.757	.984	**.726**	.849	.814	.863(↑7.6%)
+ QCD & HI	2	.968	.814	.571	.938	.730	.705	.960	**.665**	.843	**.729**	**.792**(↑20.2%)
	4	**.967**	**.977**	.664	.981	.859	.852	**.988**	.715	**.932**	**.863**	**.880**(↑9.7%)
Others												
Gemini	2	.865	.794	.315	.517	.704	.555	.583	.360	.725	.340	.576
	4	.904	.908	.540	.737	.861	.709	.773	.484	.818	.553	.729
SX	2	.440	.388	.096	.080	.060	.116	.080	.180	.164	.132	.174
	4	-	-	-	-	-	-	-	-	-	-	-
SX-IGC	2	**.984**	.568	**.968**	**1.00**	.640	.516	**.984**	.592	.712	.628	.760
	4	-	-	-	-	-	-	-	-	-	-	-

Table 11: Text version results on CoBSAT.

Method	1-shot	2-shot
MLLMs		
GILL	16.00 ± 2.27	15.17 ± 2.72
SL-8B	15.00 ± 3.27	12.67 ± 1.18
SL-14B	17.25 ± 2.75	16.75 ± 1.75
Emu1	31.50 ± 1.87	22.83 ± 2.72
Emu2	24.33 ± 3.30	30.67 ± 1.31
Anole-7B	11.00 ± 2.86	7.00 ± 0.71
LVLM-FLUX		
Base (Q2)	34.67 ± 4.48	44.83 ± 5.25
+ HI	34.33 ± 7.29	48.17 ± 3.21
+ QCD	35.83 ± 5.77	45.00 ± 3.12
+ HI & QCD	33.17 ± 2.29	47.50 ± 9.53
Base (Q2.5)	30.50 ± 1.80	34.67 ± 2.47
+ HI	31.17 ± 1.04	34.33 ± 5.75
+ QCD	33.00 ± 3.06	35.83 ± 1.15
+ HI & QCD	32.33 ± 2.78	35.67 ± 2.57
Base (I3)	34.50 ± 7.29	38.17 ± 5.48
+ HI	36.50 ± 1.53	38.00 ± 2.18
+ QCD	35.83 ± 0.76	34.00 ± 2.50
+ HI & QCD	39.00 ± 2.25	39.67 ± 2.47

Table 12: Results of different models on Text-to-Image Fast Mini-ImageNet (Accuracy %)

2023), SL (Ge et al., 2024a), and Emu1 (Sun et al., 2024b) in Figure 14. These MLLMs, to varying degrees, exhibit compliance failures (e.g., the paper box case) and prior-dominated hallucinations (e.g., the denim leaf case), whereas TBDN mitigates these issues and produces images consistent with the ground-truth answers. Moreover, leveraging FLUX's (Labs, 2024) strong T2I capability, FLUX can faithfully render the text produced by TBDN and generate higher-quality images than the compared MLLMs.

C.3 Qualitative results on Text-to-Image Fast Mini-ImageNet

Figure 15 present the qualitative results of TBDN on 1-shot evaluation in T2IFMI benchmark. TBDN can well capture the mapping between different image categories and the synthetic category name. TBDN can then infer and generate images that closely match the queried category, yielding outputs with highly consistent subjects and styles.

Method	Concept Preservation					Prompt Following				CP·PF↑
	Animal	Human	Object	Style	Overall	Photorealistic	Style	Imaginative	Overall	
SL	.436	.315	.288	.381	.358	.306	.202	.154	.218	.078
SL-IGC	.399	.290	.271	.318	.325	.348	.355	.210	.310	.101
SX	**.647**	**.420**	**.526**	**.571**	**.559**	.346	.342	.303	.337	.188
SX-IGC	.549	.410	.403	.432	.458	**.922**	**.851**	**.846**	**.881**	**.403**
TBDN(Q2)	.550	.221	.447	.407	.442	.801	.799	.698	.778	.344

Table 13: Comprehensive results on the Dreambench++ benchmark.

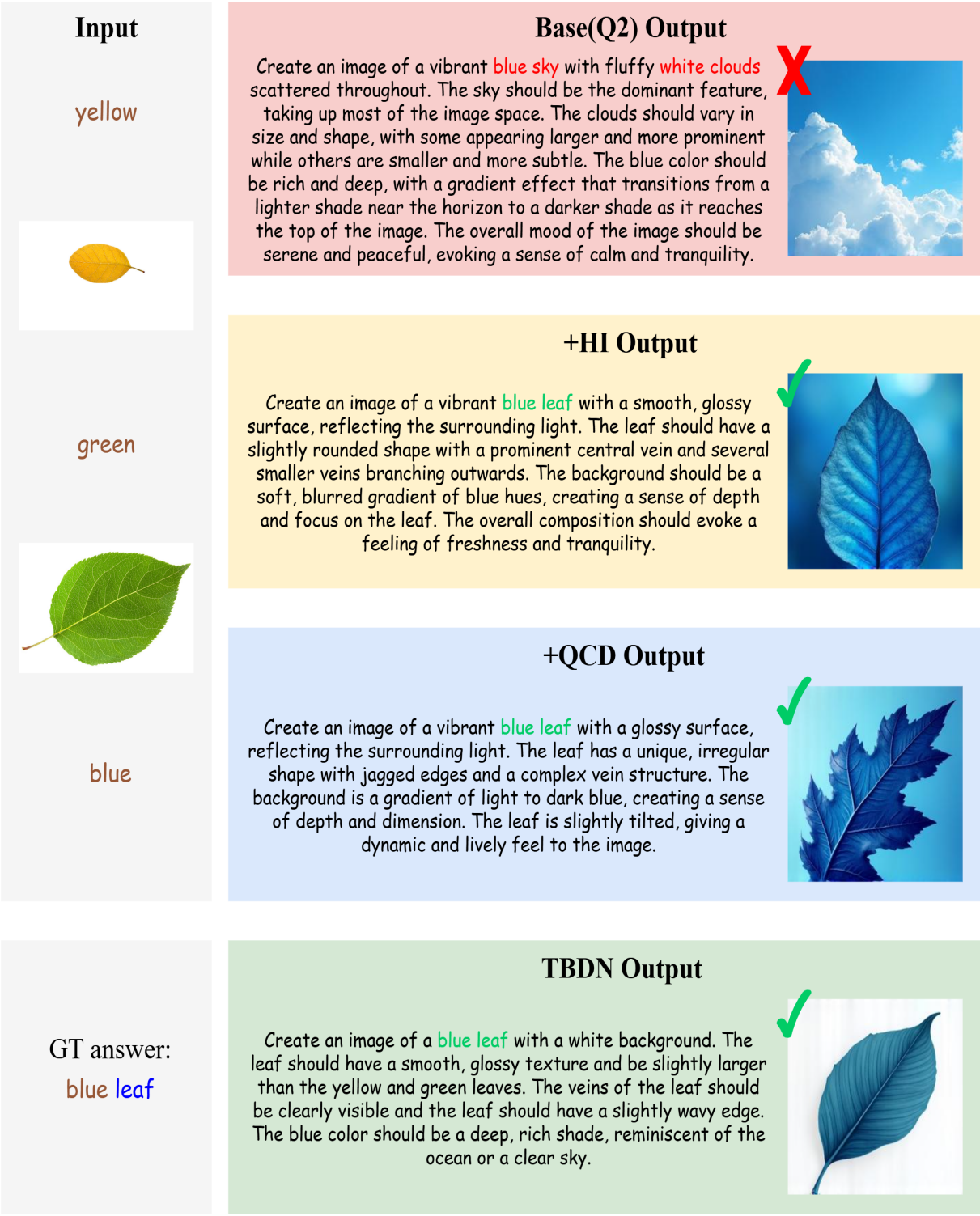

Figure 9: Qualitative results for Base (Q2), HI, QCD, and TBDN in CoBSAT (Sample 1).

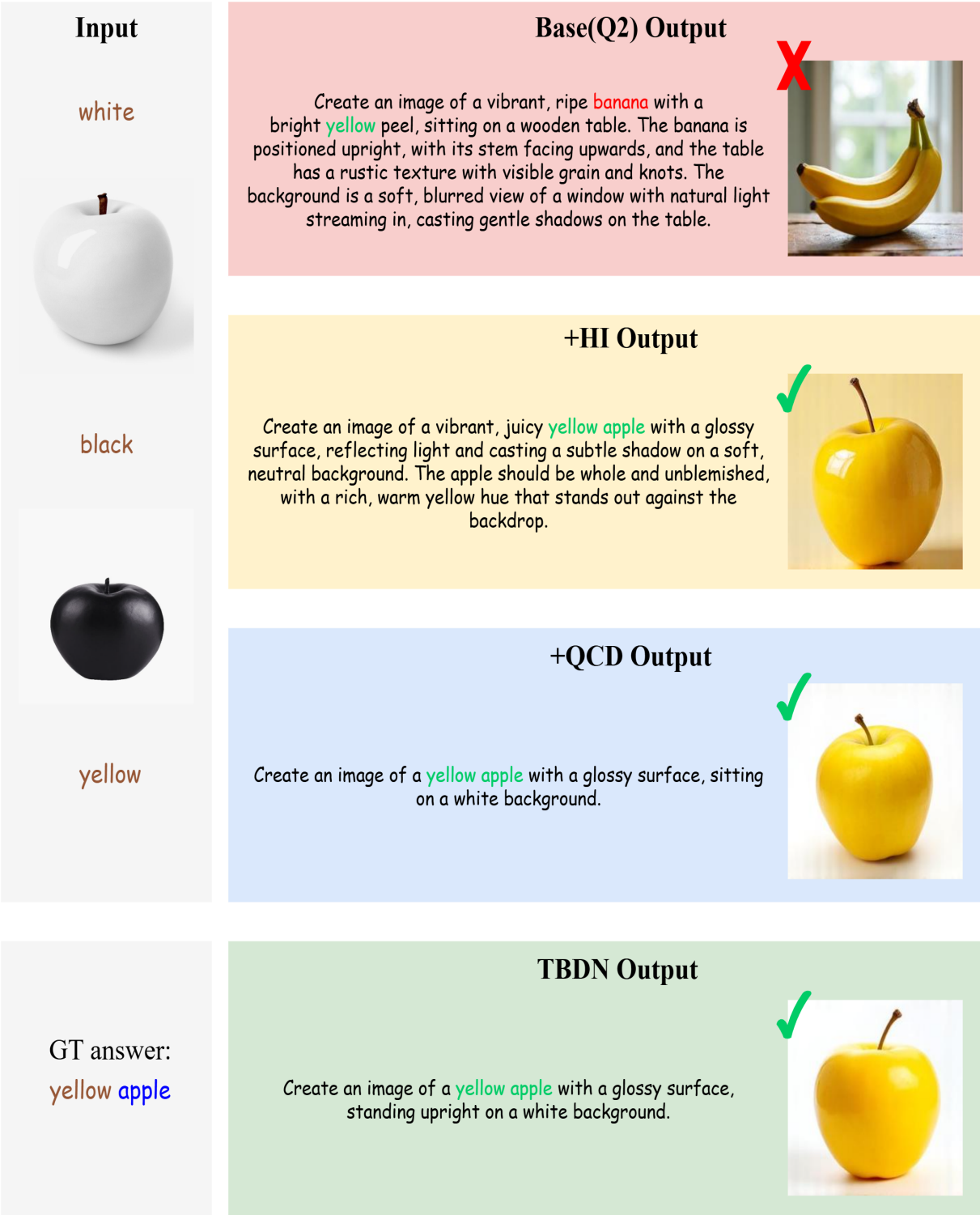

Figure 10: Qualitative results for Base (Q2), HI, QCD, and TBDN in CoBSAT (Sample 2).

Figure 11: Qualitative results for Base (Q2), HI, QCD, and TBDN in CoBSAT (Sample 3).

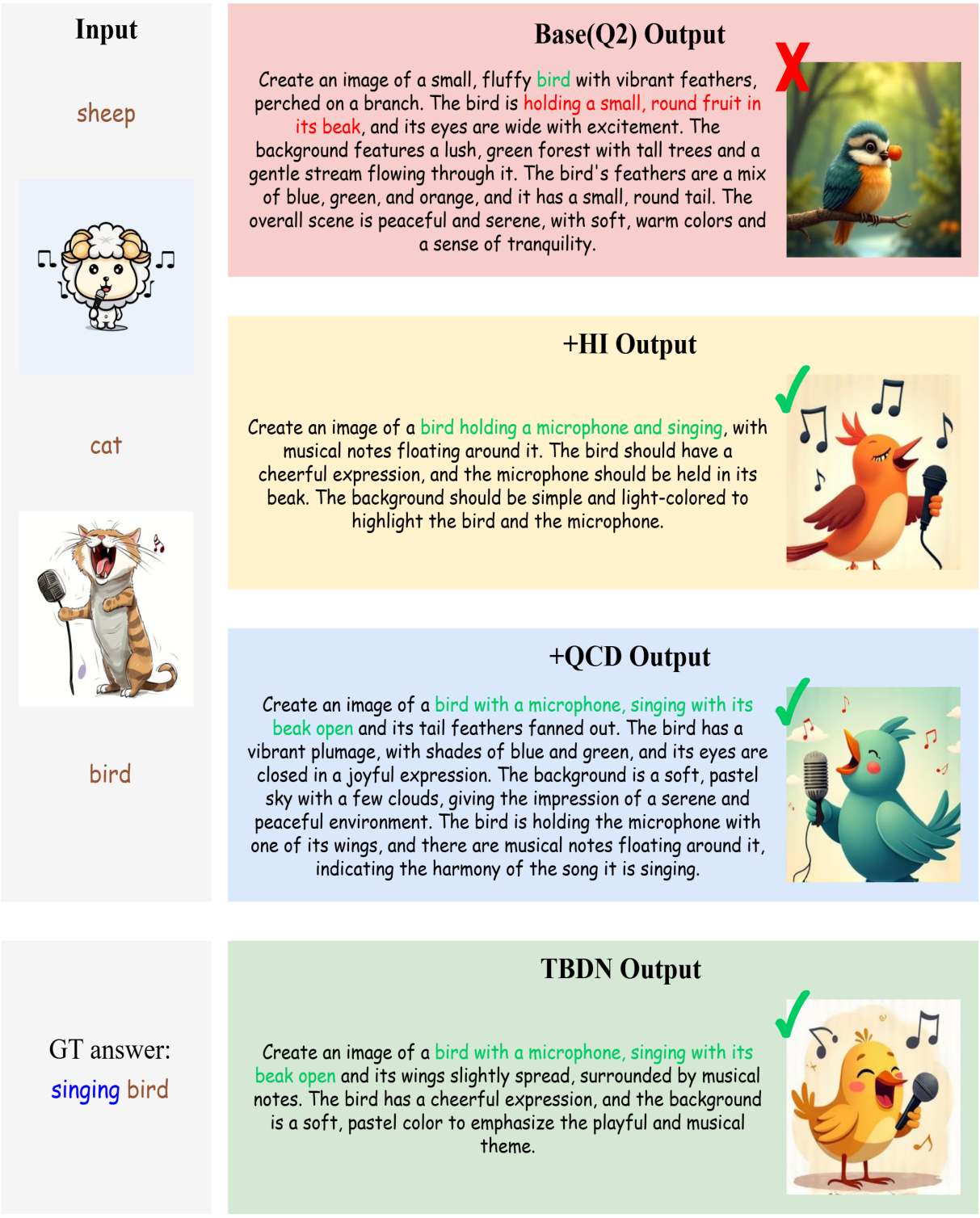

Figure 12: Qualitative results for Base (Q2), HI, QCD, and TBDN in CoBSAT (Sample 4).

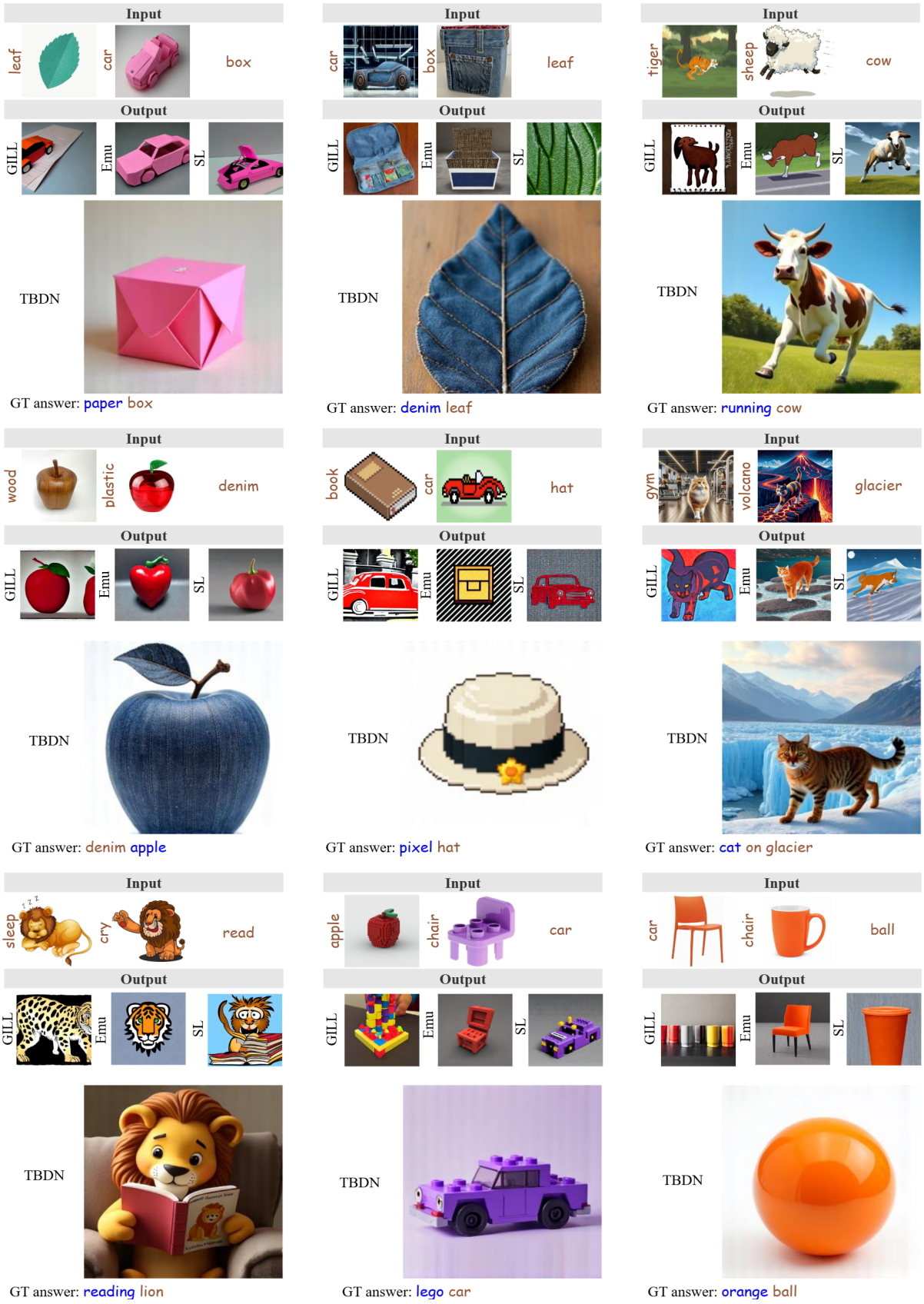

Figure 13: 2-shot image generation results of TBDN and other methods on the CoBSAT benchmark.

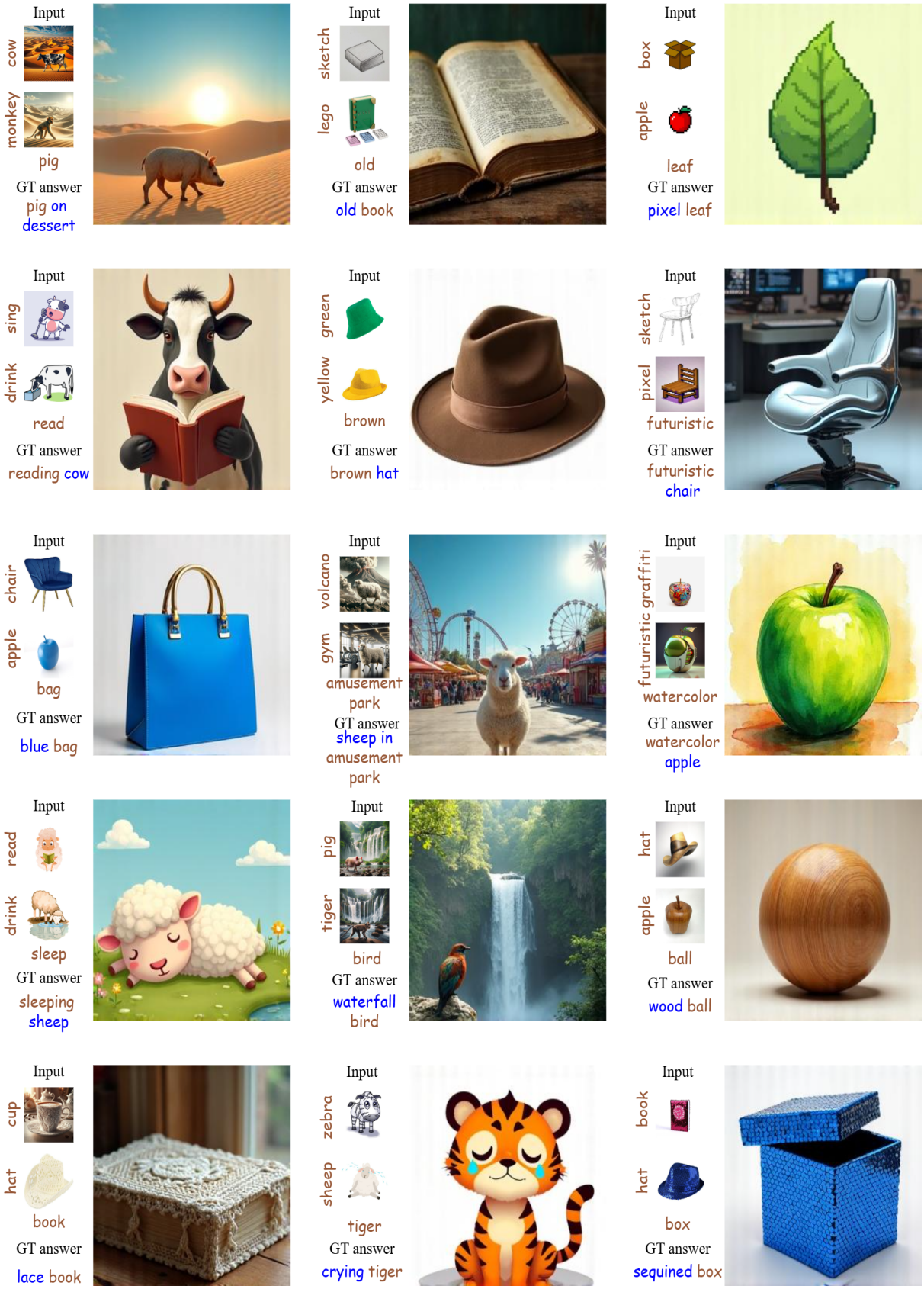

Figure 14: More 2-shot image generation results of TBDN on the CoBSAT benchmark.

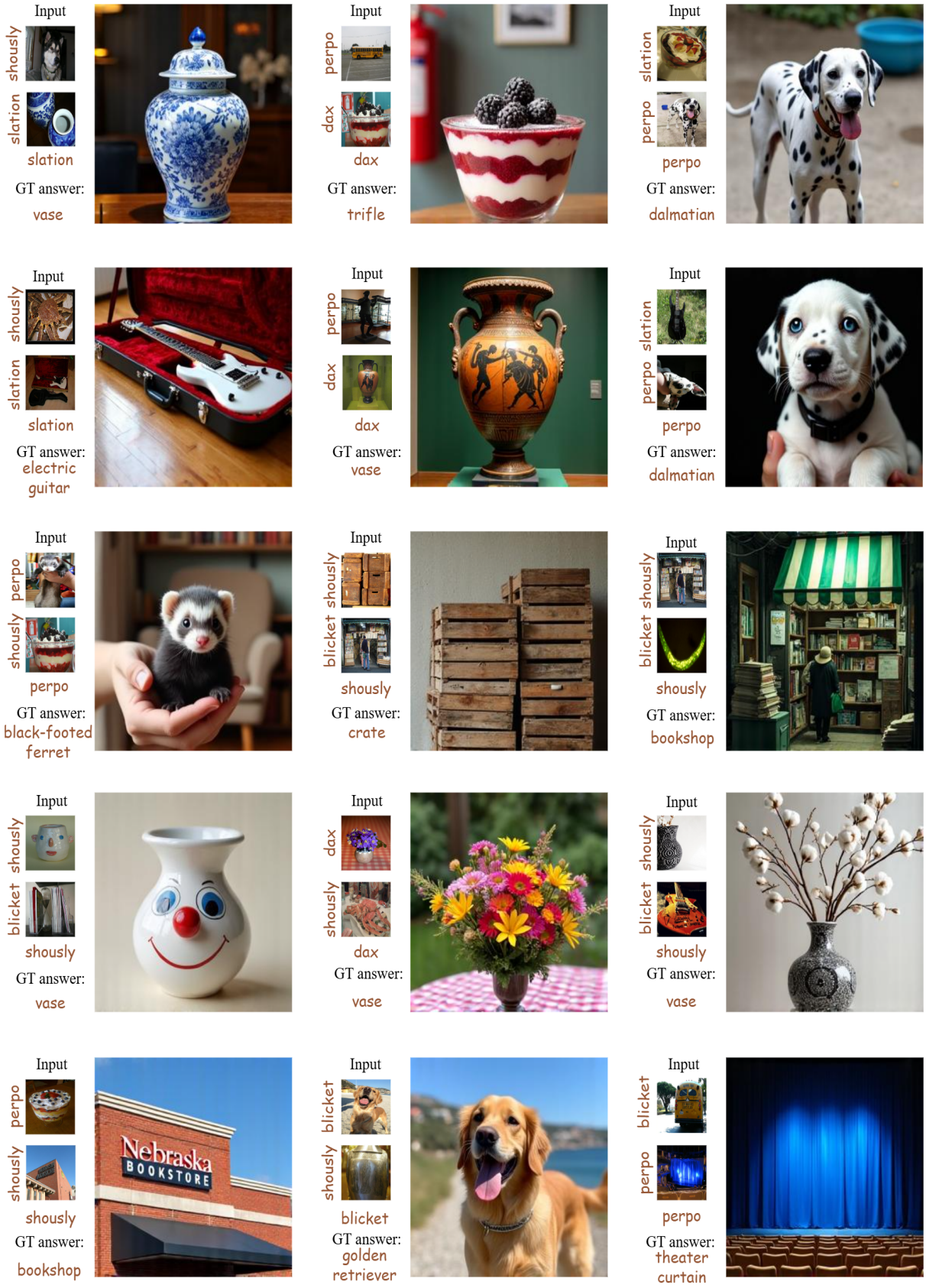

Figure 15: Image generation results on Fast Mini-ImageNet dataset.