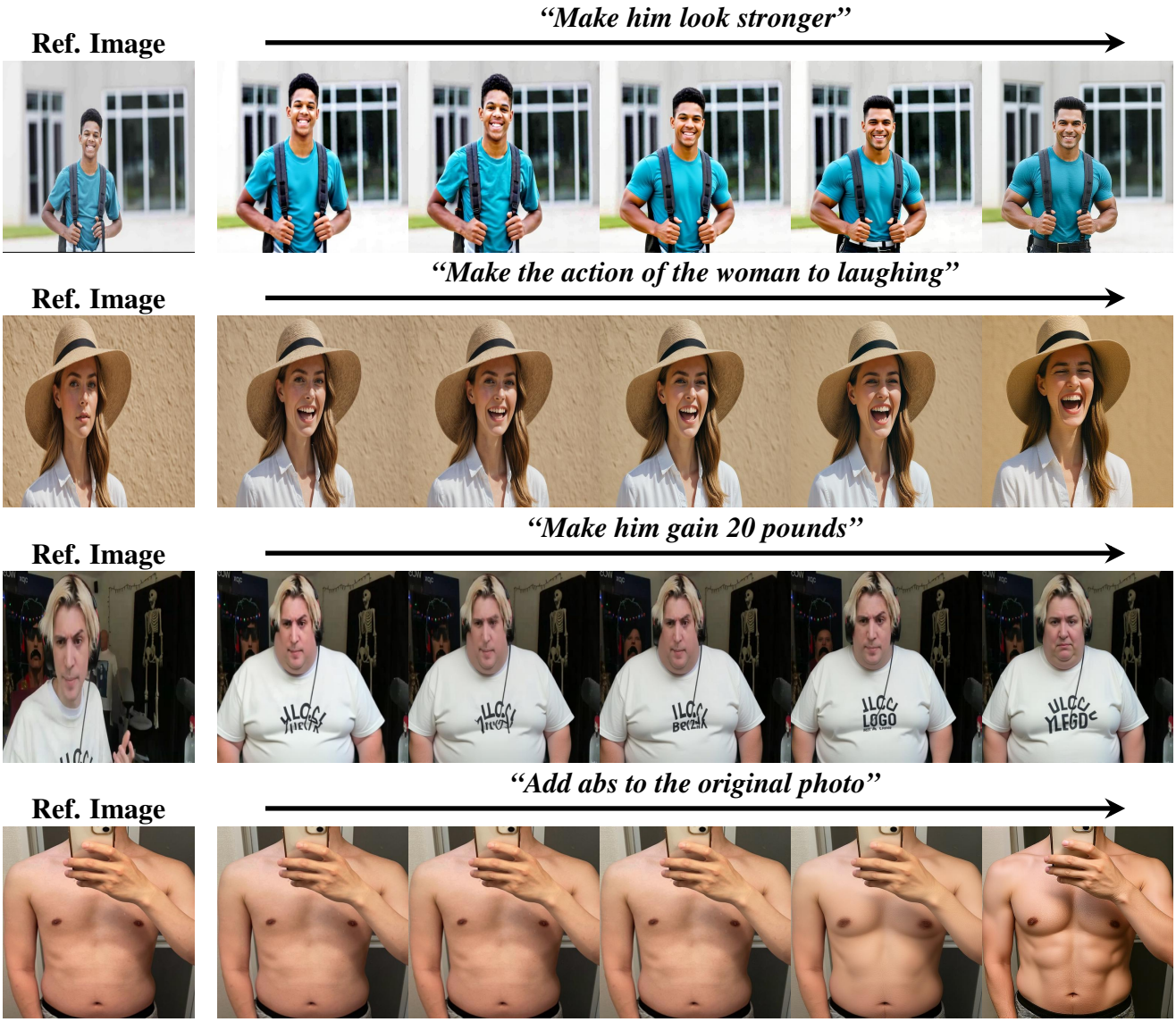


Figure 8. Training progression visualization for *Image Editing* with different prompts.

Prompt: *"Depict the chronological decomposition of a single leaf on a forest floor. All images maintain a realistic style with consistent lighting and environmental elements, focusing on the gradual transformation of the leaf while adhering to natural decay processes. The forest floor setting includes subtle elements like soil texture, scattered debris, and occasional fungi or insects."*

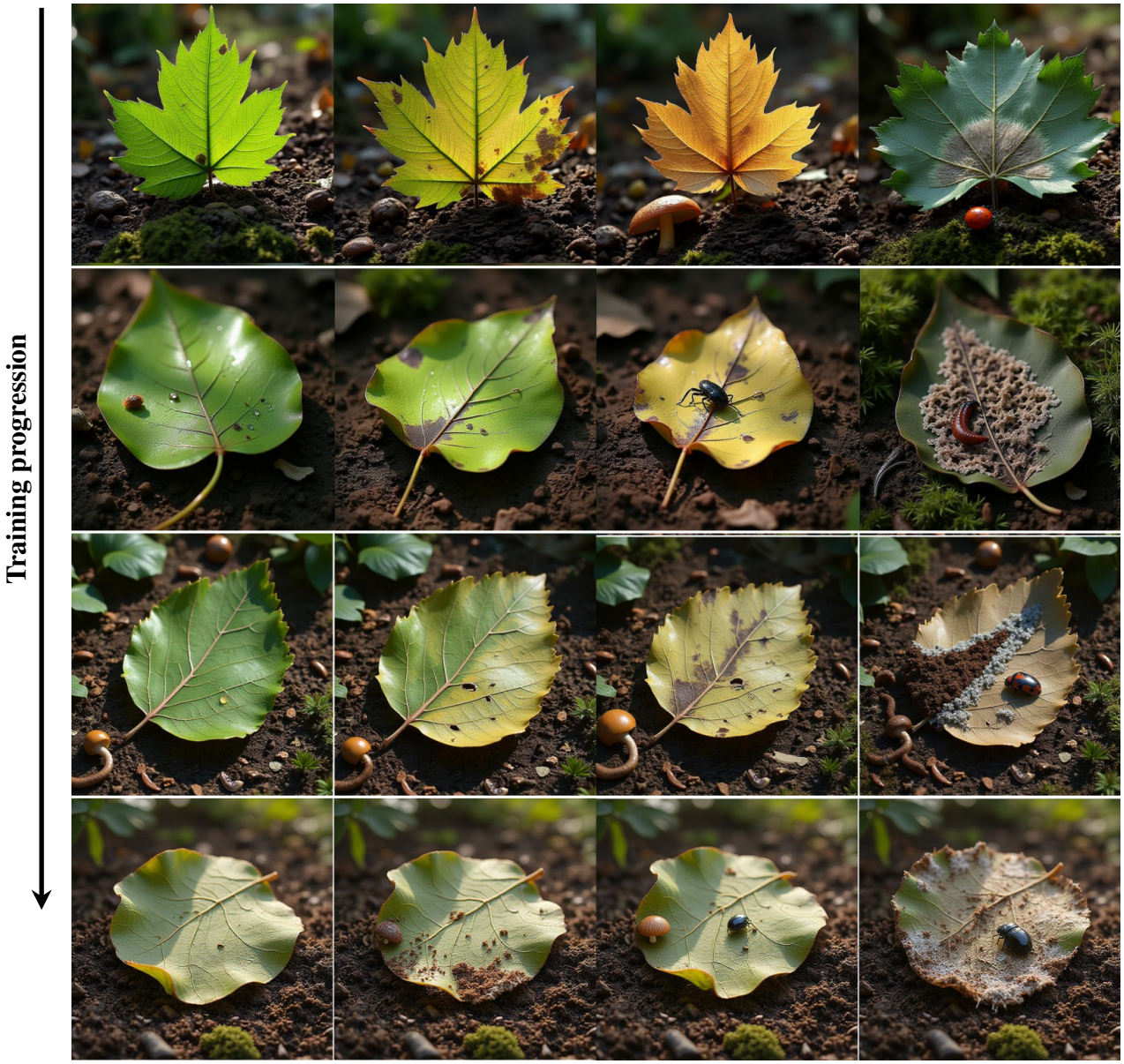


Figure 9. Training progression visualization for *Text-to-ImageSet* generation.

Prompt: *"Health beverage labels featuring honey drip font with viscous liquid texture and hexagonal comb patterns. All labels utilize the honey drip font style, integrating hexagonal comb motifs and natural/organic themes. Consistency in color palette (golden, amber, earthy tones) and texture emphasis ensures visual harmony across the set."*

Figure 10. Training progression visualization for *Text-to-ImageSet* generation.

Prompt: *"Step-by-step progression of creating a cheerful chef emoji. All images use a minimalist, cartoonish style with a clean white background. Bright and cohesive color schemes unify the stages, maintaining continuity in character proportions and playful energy."*

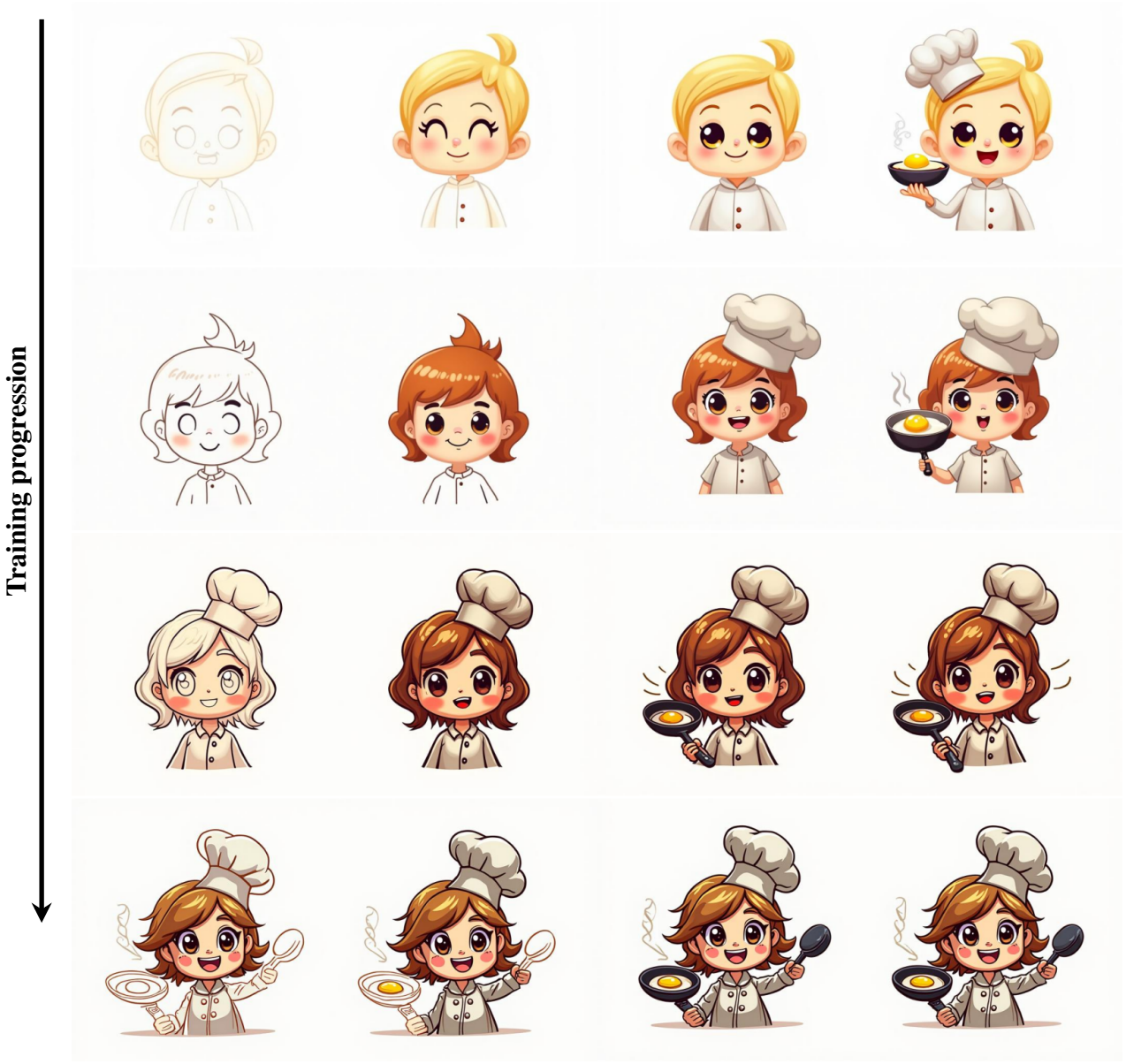


Figure 11. Training progression visualization for *Text-to-ImageSet* generation.

Prompt: *"Please generate four different perspective images of a 3D animated parrot with a vibrant and colorful plumage. The parrot exhibits a stunning array of colors, including shades of red, green, blue, and yellow, with detailed feather textures that reflect light and give a sense of depth..."*

Figure 12. Comparison of different methods for multi-image generation.

Prompt: *"Design product mockups featuring a retro, pixel art logo that reimagines our brand in an 8-bit style paired with a futuristic digital font. Apply the logo on 4 products: a portable gaming console, a vintage-style gaming t-shirt, a pixel art coffee mug, and a limited edition poster, using a monochromatic color scheme."*

Figure 13. Comparison of different methods for multi-image generation.

Prompt: *"This artwork represents the gradual creation of a traditional Chinese ink painting featuring pumpkins and vines. The progression follows these steps:..."*

Figure 14. Comparison of different methods for *Text-to-ImageSet* generation.

PaCo-RL: Advancing Reinforcement Learning for Consistent Image Generation with Pairwise Reward Modeling

Supplementary Material

7. Implementation Details

Infrastructure. All experiments are conducted on a server equipped with eight NVIDIA H100 GPUs, each with 80 GB of memory.

PaCo-Reward Implementation Details. PaCo-Reward is fine-tuned using the LlamaFactory framework [92] with a customized weighted cross-entropy loss defined in Sec. 3.2. Both PaCo-Reward-7B-Fast and PaCo-Reward-7B are fine-tuned from Qwen2.5-VL-7B-Instruct [51], employing a LoRA rank of 32, $\alpha = 64$, a learning rate of 2×10^{-4}, and a batch size of 8. Training on the PaCo-Dataset for two epochs requires approximately 18 GPU hours. For PaCo-Reward-7B, the first-token weight α in Sec. 3.2 is set to 0.1 to emphasize the importance of the initial token. Since PaCo-Reward-7B-Fast is trained solely on binary labels ("Yes."/"No."), it does not apply token weighting.

PaCo-RL Implementation Details. During PaCo-RL training, one GPU is dedicated to running a vLLM [27] server for reward computation, while the remaining seven GPUs are used for reinforcement-learning fine-tuning. Unless otherwise specified, all experiments adopt a learning rate of 3×10^{-4}, a batch size of 1, a group size of $G = 16$, and 42 unique samples per epoch. The clipping parameter ε in Eq. (1) is set to 1×10^{-4}. Following [37], we monitor KL loss throughout training but assign it zero weight ($\beta = 0$) in Eq. (1) to achieve better performance.

Text-to-ImageSet. For the Text-to-ImageSet generation task, we fine-tune FLUX.1-dev [28] using LoRA with rank 64 and $\alpha = 128$, and apply a classifier-free guidance (CFG) scale of 3.5 during both training and inference. The noise scale σ_t in Sec. 4.1 is defined as $\sigma_t = a\sqrt{\frac{t}{1-t}}$ with $a = 0.7$, following FlowGRPO [37]. We use 10 denoising steps and perform SDE sampling only at timestep $t = 1$ (from $0, 1, \ldots, 9$), leveraging MixGRPO [31] and FlowGRPO-Fast [37] strategies for efficient training. In our PaCo-RL setup, the training resolution is 512×512 while the evaluation resolution is 1024×1024. We use $\delta = 0.2$ in Eq. (4) to aggregate the Consistency Score (from PaCo-Reward-7B) and the Text-Image Alignment Score (from CLIP-T [52]).

Image Editing. For Image Editing, we fine-tune both FLUX.1-Kontext-dev [2] and Qwen-Image-Edit [74] using LoRA with rank 64 and $\alpha = 128$. The CFG scales are set to 2.5 and 4.0 for FLUX.1-Kontext-dev and Qwen-Image-Edit, respectively. We apply the MixGRPO [31] strategy for efficient training, setting the noise scale $a = 0.9$ at timestep 1 for FLUX.1-Kontext-dev [2], and $a = 1.0$ for timesteps 1-4 for Qwen-Image-Edit [74]. The training and evaluation resolutions are 384×384 and 1024×1024, respectively. As the Image Editing task relies on a single reward signal, multi-reward aggregation and the log-tame strategy in Eq. (4) are not employed. The prompt template for Image Editing reward computation is provided in Sec. 8.

8. Prompt Templates

We provide the prompt templates used for reward computation in both tasks below.

For the *Text-to-ImageSet generation* task, we design two versions of the prompt template: (1) one incorporating detailed consistency criteria derived from the original dataset for enhanced evaluation reliability, and (2) another containing only the input prompt information, which is used for generalization evaluation.

For the *Image Editing* task, we adopt a modified version of the prompt template from EditScore [40]. This template consists of two components: *Semantic Consistency (SC)* and *Prompt Following (PF)*.

Prompt Template for *Text-to-ImageSet* (v1)

Do images meet the following criteria?
{consistency_criteria}
Please answer "Yes" or "No" first, then provide detailed reasons.

Prompt Template for *Text-to-ImageSet* (v2)

Given two subfigures generated based on the theme:
{main_prompt}
do the two images maintain consistency in terms of style, logic and identity? Answer "Yes" or "No" first, and then provide detailed reasons.

Prompt Template for *Image Editing* (SC)

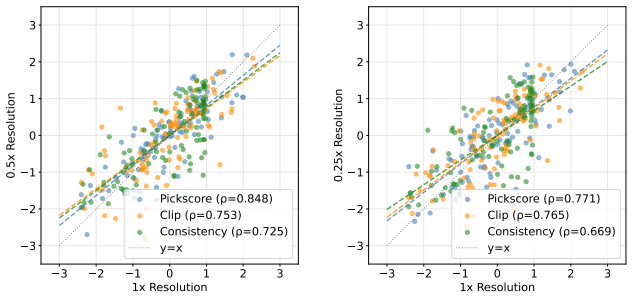
Compare the edited image (second) with the original image (first). Instruction: {prompt}. Except for the parts that are intentionally changed according to the instruction, does the edited image remain consistent with the original in style, logic, and identity? Answer 'Yes' or 'No' first, then provide detailed reasons.

Prompt Template for *Image Editing* (PF)

Compare the edited image (second) with the original image (first). Instruction: {prompt}.
Does the edited image accurately follow this instruction? Answer 'Yes' or 'No' first, then provide detailed reasons.

9. Resolution-Decoupled Training Analysis

To validate the reliability of resolution-decoupled training, we conduct experiments generating image sets at three resolutions: 256×256 (0.25x), 512×512 (0.5x), and 1024×1024 (1x) using FLUX.1-dev, and analyze the Pearson correlation of evaluation metrics across resolutions.



(a) 1x vs. 0.5x (b) 1x vs. 0.25x

Figure 15. Pearson correlation of evaluation metrics across different training-to-inference resolution ratios. Strong correlations at 0.5x confirm that reward signals remain reliable under moderate resolution reduction.

As shown in Fig. 15, all metrics exhibit strong positive correlations (0.725–0.848) when comparing 1x vs. 0.5x resolutions, indicating that the relative quality ordering of image sets is largely preserved and reward signals remain reliable under moderate resolution reduction. However, correlations weaken substantially when comparing 1x vs. 0.25x, especially for Aesthetics and Consistency, suggesting that extremely low resolutions lose detailed visual information crucial for these fine-grained aspects.

We further validated these findings on additional reward models (Text-Rendering, GenEval) and generators (Qwen-Image, FLUX2), confirming that 0.5x reduction consistently preserves reward reliability while 0.25x fails, particularly for fine-grained metrics. Regarding higher-resolution inference, no systematic artifacts were observed when generating at full resolution after 0.5x training, as the reward improvements transfer effectively across resolutions. This supports the use of resolution-decoupled training as a practical strategy for reducing computational cost without sacrificing final generation quality.

10. Ablation on PaCo-GRPO Components

We quantify the individual contributions of two key components in PaCo-GRPO: resolution-decoupled training (Res-Dec.) and log-tamed reward aggregation (Log-Agg.). Results are reported in Tab. 5.

Resolution-Decoupled Training. Removing resolution-decoupled training results in suboptimal performance even with doubled training time (12h vs. 6h), as the model fails to converge fully at higher resolution. This demonstrates that training at a reduced resolution (0.5x) not only halves the computational cost but also leads to better optimization dynamics by enabling more effective exploration within the same time budget.

Table 5. Ablation study on PaCo-GRPO components. Removing resolution-decoupled training leads to suboptimal performance even with doubled training time, while removing log-tamed aggregation causes reward collapse.

Method	Aes.↑	P.F.↑	V.C.↑	Time↓
Full PaCo-GRPO	**0.555**	**0.728**	0.493	**6.0h**
W/O Res-Dec.	0.542	0.698	0.452	12.0h
W/O Log-Agg.	0.471	0.616	**0.557**	6.0h

Log-Tamed Reward Aggregation. Without log-tamed aggregation, the visual consistency reward dominates the training signal, causing the model to collapse into generating near-identical, low-quality image sets. While the Visual Consistency (V.C.) score increases to 0.557, both Aesthetics and Prompt Following degrade significantly, confirming that the log-tamed strategy in Eq. (4) is essential for balancing multiple reward objectives and preventing reward hacking.

11. PaCo-Dataset Details

To ensure the quality and consistency of PaCo-Dataset annotations, all annotators followed detailed guidelines accompanied by illustrative examples to establish a shared understanding of the evaluation criteria across different consistency dimensions (*i.e.*, style, logic, and identity). Each image pair was independently evaluated by multiple annotators, and the final annotations were obtained via majority voting. Ties were discarded to ensure clear binary preferences in the dataset.

The main categories, subcategories, and their corresponding consistency dimensions are summarized in Tab. 6.

Table 6. Main categories, subcategories, and their corresponding consistency dimensions.

Main Category	**Subcategory**	**Consistency Dimensions**
Design Style Generation	Home Decoration	Style
	IP Product	Style, Identity
	Font Design	Style
	Poster Design	Style, Logic
	Creative Style	Style
Story Generation	Children Book	Logic, Identity, Style
	Hist. Narrative	Logic, Identity
	Movie Shot	Logic, Identity, Style
	Comic Story	Logic, Identity, Style
	News Illustration	Logic, Style
progression Generation	Evolution Illustration	Logic
	Draw progression	Logic, Style
	Growth progression	Logic
	Arch. Building	Logic
	Cooking progression	Logic
	Physical Law	Logic
Instruction Generation	Historical Panel	Logic, Style
	Activity Arrange	Logic
	Evolution Illustration	Logic
	Education Illustration	Logic, Style
	Travel Guide	Logic, Style, Identity
	Product Instruction	Logic, Style
Character Generation	Multi-view	Identity, Style
	Multi-pose	Identity
	Portrait Design	Identity, Style
	Multi-Expression	Identity
	Multi-Scenario	Identity, Logic
Editing	Inpainting and replacement	Identity
	Element manipulation	Identity, Style
	Background modification	Identity, Style, Logic
	Attribute and effect manipulation	Style
	Image editing and manipulation	Identity, Style, Logic