PAINT with words: A Framework and Dataset for Visual Creative Description Generation

Anonymous ACL submission

Abstract

Visual Creative Description (VCD) generation, which involves crafting imaginative and actionable textual prompts for text-to-image models, is a critical yet underexplored task for large language models (LLMs). We propose the Cognitive Chain-of-Creativity (C-CoC), a novel framework that leverages structured cognitive modeling to enhance the novelty and visual expressiveness of generated descriptions. To support this task, we introduce the PAINT dataset, comprising high-quality VCDs across 33 product categories. Experiments demonstrate that C-CoC significantly improves description creativity by 10%-19% compared to baselines. However, our evaluation of LLMs reveals limited alignment with human judgments in assessing VCD quality, highlighting the complexity of creative evaluation. Our contributions lay a foundation for structured creative generation and underscore the need for advancements in LLM-based evaluation.

1 Introduction

In recent years, creative visual content has become a cornerstone of effective communication across digital platforms. Research demonstrates that creative visuals significantly enhance brand perception and consumer engagement (Bostancı and Dursun, 2024), powering a global industry valued at hundreds of billions annually (Hartmann et al., 2025). While Text-to-Image (T2I) models have advanced in generating relevant visual compositions (Ramesh et al., 2022) and Large Language Models (LLMs) have shown remarkable capabilities in creative text generation (Yuan et al., 2022; Belouadi and Eger, 2023; Mita et al., 2024), there remains a critical gap at their intersection. Current T2I systems require specialized prompt engineering expertise that most users lack (Cao et al., 2023), creating a significant barrier to effective visual content creation. This raises an important research question: Can LLMs generate visual creative descriptions (VCDs) that enhance the creativity and effectiveness of generated images? This understudied area presents a valuable opportunity to bridge natural language processing with visual creativity, potentially transforming how visual content is conceptualized and produced.

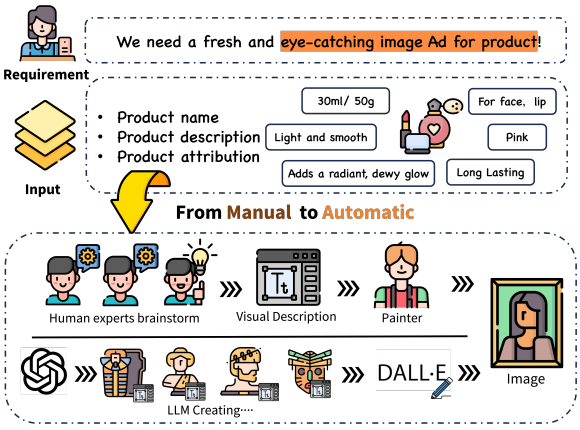

Figure 1: Comparison between traditional human-driven creative workflow and our proposed LLM-based approach for generating product advertising images.

Despite the critical importance of VCDs in enhancing AI-generated image quality, research on LLM-based VCD generation faces three key challenges. First, the lack of high-quality VCD training data impedes LLMs' ability to produce novel and actionable descriptions. When facing data scarcity, an effective strategy is to incorporate inductive bias—specifically, by modeling human creative cognitive processes. These processes, which involve conceptual association, visual ideation, and linguistic expression (Botella et al., 2018), remain inadequately modeled in current approaches, limiting the potential for synthetic data generation. Finally, evaluating VCD quality is hindered by the absence of specialized frameworks, as existing metrics fail to capture the multidimensional nature of creative descriptions. Addressing these challenges

is essential for advancing research at the intersection of natural language processing and computational creativity.

To address these challenges, we draw inspiration from cognitive science and propose **Cognitive Chain of Creativity (C-CoC)**, a framework for generating VCDs (Figure 1). C-CoC leverages LLMs as *Cognitive Operators* that execute structured transitions across reasoning stages, simulating human creative thought processes.

Our framework transforms structured entity information into visually creative descriptions aligned with communication principles, which then drive text-to-image (T2I) generation. This approach enhances both creativity and visual expressiveness while providing an evaluable method for creative generation.

To address data scarcity, we introduce **PAINT (Product Artistic Image Narrative Texts)**, a dataset of structured descriptions exhibiting novelty, executability, and communicability. PAINT fills a critical training data gap and enhances LLMs' generalization in visual creative tasks.

Our contributions are: (1) **C-CoC Framework:** A cognitive-inspired approach using multi-stage reasoning to enhance LLM interpretability and controllability in creative generation; (2) **PAINT Dataset:** A large-scale, high-quality dataset of VCDs, addressing the training data gap in this domain; (3) **Systematic Evaluation:** Comprehensive methodologies for assessing VCD quality across multiple dimensions, demonstrating C-CoC's improvements in both description quality and downstream image generation.

2 Related Work

2.1 Text-to-Image Generation

Recent advances in Text-to-Image (T2I) models have significantly improved the semantic alignment between textual descriptions and generated images. Early approaches relied on Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to map textual inputs to visual outputs. Later, autoregressive models such as DALL·E (Ramesh et al., 2021) and ImageGPT (Chen et al., 2020) leveraged token-based sequence prediction to enhance text-conditioned generation. Currently, Diffusion Models have become the dominant paradigm in T2I tasks, achieving state-of-the-art results in both fidelity and semantic consistency. Models such as Imagen (Saharia et al., 2022), FLUX (Black-Forest-Labs, 2024) and Stable Diffusion (Rombach et al., 2022) leverage latent diffusion processes to generate high-quality images with fine-grained details.

Several recent research endeavors advocate for extensions of T2I models, aiming to increase their fidelity to user prompts(Epstein et al., 2023; Chefer et al., 2023; Wu et al., 2023). However, despite these improvements, the T2I models remain highly dependent on the quality of their textual inputs.

2.2 Prompt Optimization

With the enhancement of text-to-image alignment capabilities in T2I models, the quality of generated images has become increasingly dependent on well-crafted textual inputs. In recent years, many researchers have attempted to optimize prompts to achieve better image generation outcomes, such as refining task instruction prompts using training data (Guo et al., 2023; Fernando et al., 2023).Some studies focus on optimizing individual T2I prompts at the multimodal inference stage (Mañas et al., 2024). For instance, reinforcement learning has been employed to fine-tune large language models to enhance the aesthetic quality of generated images, while (Valerio et al., 2023) have concentrated on filtering out non-visual prompt elements to improve visual consistency.

Current prompt optimization research predominantly refines and modifies pre-existing ideas, essentially functioning as a form of faithful and elegant machine translation(Zhan et al., 2024) of existing concepts. However, our work focuses on creative generation from scratch, emphasizing the ideation process rather than merely optimizing or adapting existing prompts.

2.3 Cognitive Modeling

Creativity in human cognition has been extensively studied in cognitive science and psychology, where it is often conceptualized as a structured process rather than an arbitrary generation of ideas (Boden, 2004; Finke et al., 1996). One of the most influential models, the *Geneplore Model* (Finke et al., 1996), characterizes creativity as a two-phase process.

The first phase, termed the **Generative Phase**, involves the cognitive system constructing an initial structured representation based on existing knowledge, forming a stable foundation for subsequent transformation. Following this, the **Exploratory Phase** sees the system refining, restructuring, or

blending these initial representations to produce novel and creative outputs.

This hierarchical perspective aligns with research on structured cognitive processing, where creativity emerges from a balance between stability and flexibility (Smith et al., 1995). However, how to systematically model this process in computational systems, particularly in large language models (LLMs), remains an open challenge.

Recent work has explored LLMs' potential for creative generation (Yuan et al., 2022; Mita et al., 2024; Belouadi and Eger, 2023). Current methods fail to establish a generation paradigm for creative work, nor do they construct a systematic cognitive framework. The absence of such a structured model limits the controllability and novelty of the generated content.

3 Visual Creative Description Generation

In this section, we formally define the task of VCD generation and introduce the corresponding multi-dimensional evaluation framework.

3.1 Formalization of the VCD Task

The VCD task involves transforming an input prompt P into a creative textual description D^* suitable for image generation. In its simplest form, this can be represented as a direct mapping:

$$f : P \to D^* \tag{1}$$

However, this direct approach often fails to produce descriptions with adequate creativity while maintaining semantic coherence. To address this challenge, we propose the Cognitive Chain-of-Creativity (C-CoC) framework, which structures the VCD generation process through discrete cognitive operations:

$$P \to C_{\text{decompose}} \to D_{\text{base}} \to S_{\text{creative}} \to D^* \tag{2}$$

Algorithm 1 formalizes this process with precise computational steps.

The algorithm takes as input the initial prompt P, cognitive transformation dimensions Θ, language model $\mathcal{M}$, and image generator G. It processes each prompt through four stages: concept decomposition, base description generation, application of cognitive shifts across multiple dimensions, and final creative optimization. This structured approach enables controlled creative transformations while maintaining the coherence necessary for effective image generation.

Algorithm 1 Cognitive Chain-of-Creativity Generation

Require: $P, \Theta, \mathcal{M}, G$
Ensure: D^*, I^*
  for each P **do**
    $C_{\text{decompose}} \leftarrow \text{ExtractConcepts}(P)$
    $D_{\text{base}} \leftarrow \text{GenerateBase}(\mathcal{M}, C_{\text{decompose}})$
    for each $\theta \in \Theta$ **do**
      $S_{\text{creative}} \leftarrow \text{ApplyCognitiveShift}(D_{\text{base}}, \theta)$
    end for
    $D^* \leftarrow \text{OptimizeForCreativity}(S_{\text{creative}})$
    $I^* \leftarrow G(D^*)$
  end for
  return D^*, I^*

Expressed Entity (P) Given the fundamental information of an **Expressed Entity (EE)**, which refers to the subject being described in the creative process, we define $P = \{p_{\text{name}}, p_{\text{desc}}, p_{\text{attr}}\}$, where p_{name} represents the entity name (e.g., "a chair", "a vase"), p_{desc} provides a brief textual description that outlines the entity's overall purpose or essence, and $p_{\text{attr}} = \{a_1, a_2, \ldots, a_n\}$ denotes a set of attributes describing the entity's properties, such as color, material, shape, or functionality.

Concept Decomposition ($C_{\text{decompose}}$) This step extracts core concepts by analyzing entity attributes: $C_{\text{decompose}} = \Phi(P, \mathcal{M})$, where $\Phi(\cdot)$ identifies the key conceptual elements from the entity's attribute set, and $\mathcal{M}$ denotes the cognitive operator applied during this stage.

Base Expression (D_{base}) This step generates a conventional visual description that adheres to traditional paradigms: $D_{\text{base}} = \Psi(C_{\text{decompose}}, \mathcal{M})$, where $\Psi(\cdot)$ transforms the extracted conceptual elements into a logically sound, attribute-aligned base description.

Cognitive Shift (S_{creative}) This step applies constraints from cognitive science, psychology, and artistic principles to transform the base description into a highly creative visual expression: $S_{\text{creative}} = \Omega(D_{\text{base}}, \Theta, \mathcal{M})$, where Θ represents the constraints guiding the cognitive transformation, such as creativity, narrative logic, and stylistic guidelines, and $\Omega(\cdot)$ applies cognitive shifts to ensure the generated description aligns with both visual communication rules and creativity principles.

Creative Realization (D^*) This step produces the

final creative description by incorporating **composition aesthetics, narrative logic, and stylistic consistency**: $D^* = \Gamma(S_{\text{creative}})$, where $\Gamma(\cdot)$ ensures that the transformed creative expression is both actionable and interpretable by the T2I model.

3.2 Evaluation Framework for Visual Creative Descriptions

The evaluation framework for VCDs is grounded in theories of hierarchical visual processing and Gestalt principles. These theories emphasize the importance of holistic perception and layered processing in assessing visual outputs. To ensure the generated descriptions are clear, engaging, and visually coherent, we evaluate them across the following four interrelated dimensions:

$$\Theta = \{\theta_{\text{plot}}, \theta_{\text{color}}, \theta_{\text{volume}}, \theta_{\text{background}}\} \quad (3)$$

Each dimension is designed to measure specific aspects of the description, spanning both high-level semantic construction and low-level sensory optimization.

3.2.1 High-Level Dimensions (Semantic Construction)

Plot Creativity (θ_{plot}): This dimension evaluates the narrative structure embedded in the description. Drawing inspiration from narrative psychology (Bruner, 1991; Green and Brock, 2000), it assesses how storytelling techniques enhance the coherence, engagement, and emotional resonance of the visual representation. A strong narrative not only provides the description with a clear purpose but also ensures it leaves a lasting impression on the audience.

Background Creativity ($\theta_{\text{background}}$): This dimension focuses on how the surrounding scene or environment contributes to the meaning and coherence of the visual description. Based on scene semantics (Bar, 2004; Oliva and Torralba, 2007), an effective background reinforces object-scene relationships and provides crucial contextual information. For instance, a beach background for a chair description might evoke relaxation, while a study room background suggests productivity. This contextual alignment enhances the overall interpretability and relevance of the description.

3.2.2 Low-Level Dimensions (Sensory Optimization)

Color Creativity (θ_{color}): This dimension assesses the use of colors to evoke emotions and enhance visual appeal. Guided by principles of color psychology (Elliot and Maier, 2014; Labrecque and Milne, 2012), specific color schemes are used to convey different moods and tones. For example, warm tones (e.g., red, orange) evoke excitement or energy, while cool tones (e.g., blue, green) suggest calmness or sophistication. A well-aligned color scheme strengthens the descriptive alignment with the intended emotional tone.

Volume Creativity (θ_{volume}): This dimension examines the spatial arrangement and relative dominance of visual elements. Rooted in Gestalt principles of focal hierarchy and compositional balance (Arnheim, 1954; Palmer, 1999), it evaluates whether the description prioritizes attention on key elements while maintaining overall harmony. Proper volume management ensures that the most significant elements stand out while secondary elements support the main focus without overwhelming the composition.

3.2.3 Interaction Between Dimensions

These four dimensions are interdependent and work together to produce a cohesive and visually compelling description. Their interplay can be summarized as follows:

High-level dimensions define purpose: Plot creativity establishes the narrative foundation, while background creativity provides the contextual environment for the visual elements.

Low-level dimensions enhance delivery: Color creativity optimizes the description's emotional tone and visual appeal, while volume creativity ensures compositional harmony and effective focus distribution.

Feedback loops for refinement: Changes in low-level dimensions (e.g., color or volume) can influence high-level dimensions like plot coherence or background alignment. This necessitates iterative refinements to maintain balance across all dimensions.

By integrating these dimensions into a unified framework, we ensure that the evaluation of VCDs is both comprehensive and aligned with cognitive and perceptual principles.

4 Construction of PAINT Dataset

4.1 Dataset Design

Visual Creative Descriptions have a wide range of applications, with one particularly important do-

Table 1: Performance comparison between baseline and our C-CoC method across different product categories.

Method	Category	Evaluation Dimensions			
		Plot	Color	Volume	Background
Baseline	*Beauty*	63.40	60.92	62.35	63.53
	Home	61.22	60.44	56.67	61.78
	Lifestyle	61.80	62.33	59.73	61.80
	Media	64.57	63.62	60.10	65.81
	Electronics	62.19	61.64	57.99	62.06
C-CoC (Ours)	*Beauty*	**75.95** (↑12.55)	**72.94** (↑12.03)	**71.90** (↑9.54)	**75.42** (↑11.90)
	Home	**72.89** (↑11.67)	**72.11** (↑11.67)	**73.89** (↑17.22)	**71.78** (↑10.00)
	Lifestyle	**75.60** (↑13.80)	**70.67** (↑8.33)	**76.40** (↑16.67)	**72.60** (↑10.80)
	Media	**75.52** (↑10.95)	**72.76** (↑9.14)	**77.81** (↑17.71)	**74.48** (↑8.67)
	Electronics	**73.23** (↑11.04)	**73.31** (↑11.67)	**77.51** (↑19.53)	**76.01** (↑13.95)

main being the generation of creative image advertisements. In this domain, textual descriptions must accurately convey product information while maintaining visual appeal, brand identity, and market influence. To support research and benchmarking in this domain, we introduce PAINT (Product Advertisement Image Narrative Texts), a dataset specifically designed for VCD tasks in product advertisements.

The PAINT dataset was developed with the following key objectives: (1) to facilitate research on generating visually expressive and creative textual descriptions for product advertisements, and (2) to provide a benchmarking resource to evaluate the effectiveness of VCDs across diverse advertisement scenarios. These objectives ensure that the dataset is both adaptable to various research tasks and relevant to real-world applications.

To achieve these objectives, two fundamental design principles were established during the construction of PAINT:

Design Principle 1: Emphasizing Creativity and Expressiveness. PAINT focuses on ensuring that textual descriptions are not only informative but also highly creative and visually expressive. This principle reflects the importance of integrating cognitive and artistic elements into the description generation process, allowing for the effective conveyance of product features alongside creative visual imagery. By prioritizing expressiveness, the dataset enables research on exploring how textual descriptions can evoke vivid mental images while maintaining clarity and relevance.

Design Principle 2: Alignment with Visual Contexts. The dataset highlights the importance of aligning textual descriptions with their corresponding visual elements. This principle ensures that descriptions are contextually grounded, accurately reflecting the visual characteristics of the associated product images. For instance, a description of a vase should consider its shape, color, material, and style as depicted in the image, while also incorporating creative elements that enhance its appeal. This alignment is critical for applications such as creative advertisement generation, where textual descriptions must resonate with visual content to maximize user engagement.

4.2 Dataset Development Process

Our data source is the **Amazon Reviews dataset** (Hou et al., 2024), collected by McAuley Lab in 2023. From this dataset, we extracted the two most fundamental product parameters, namely `"title"` and `"description"`, as input for our task.

To ensure the diversity of our dataset, we sampled product information from 33 categories across five major groups: `"Beauty"`, `"Electronics"`, `"Home"`, `"Media"`, and `"Lifestyle"`. The distribution of these categories is illustrated in Figure 2.

In our dataset, we used the basic product parameters as input. To meet the requirements of **Design Principle 1: Emphasizing Creativity and Expressiveness**, we employed the **C-CoC framework** to generate VCDs .

Additionally, to validate **Alignment with Visual Contexts**, we extended our dataset by generating images corresponding to the VCDs, forming a *VCD+image pair* dataset.

Figure 2: Sunburst visualization of PAINT dataset's category distribution. The inner ring represents five major category groups (`Beauty`, `Electronics`, `Home`, `Media`, and `Lifestyle`), while the outer ring displays 33 specific subcategories distributed across these groups.

4.3 Annotation

The annotation process was designed to systematically evaluate VCDs across four cognitive transition dimensions. We recruited five NLP experts with graduate-level qualifications in computational linguistics or related fields to serve as annotators, ensuring professional assessment of the creative and cognitive aspects of the descriptions.

Each sample in our dataset received independent evaluations from three different experts. Following the cognitive transition framework outlined in Section 3.2, annotators scored both the base text ($T_{\text{base},i}$) and the post-transition text ($T_{\text{shift},i,k}$) for each cognitive dimension (θ_{plot}, θ_{color}, θ_{volume}, and $\theta_{\text{background}}$) using a 0-5 scale. Detailed scoring guidelines were provided to all annotators (see Figures 8 and 9 in Appendix B).

For each annotation, the cognitive transition increment was calculated as:

$$\Delta S_{i,k,j}^{\text{human}} = a_{i,k,j}^{\text{shift}} - a_{i,k,j}^{\text{base}}$$

To establish ground truth, we employed a majority voting system across the three annotators, with the final cognitive transition score determined by averaging the scores from the majority group:

$$\Delta S_{i,k}^{\text{human}} = \frac{1}{|J_{\text{majority}}|} \sum_{j \in J_{\text{majority}}} (a_{i,k,j}^{\text{shift}} - a_{i,k,j}^{\text{base}})$$

To ensure consistent evaluation procedures, annotators used a customized Label Studio interface (Figure 10), which standardized the assessment process across all samples and dimensions.

The reliability of our annotation approach was validated through multiple consistency metrics. As shown in Table 2, inter-annotator agreement was substantial across all dimensions, with Spearman correlation coefficients ranging from 0.726 to 0.782 and directional agreement rates between 72.42% and 89.39%. These strong consistency indicators validate our evaluation approach despite the inherently subjective nature of creative assessment (Botella et al., 2018).

Dimension	Spearman's ρ	Direct. Agree (%)
Plot	0.726	80.61
Color	0.774	72.42
Volume	0.746	89.39
Background	0.782	78.18
Mean	**0.757**	**80.15**

Table 2: Inter-annotator agreement on cognitive transition judgments.

5 Experiments

5.1 Experimental Setup

To evaluate our framework, we utilized approximately 20% of the PAINT dataset (1650 description pairs), ensuring balanced representation across all 33 product categories. This sample size was determined to optimize the balance between statistical power and annotation resource constraints, providing sufficient data points for robust analysis while maintaining annotation quality.

For our baseline approach, we employed few-shot prompting with GPT-4o-mini, similar to C-CoC but without targeting specific cognitive dimensions. The key difference is that our C-CoC framework systematically enhances particular cognitive dimensions, while the baseline generates descriptions without dimension-specific guidance.

The evaluation methodology focused on creativity, visual expressiveness, and practical applicability for image generation. As summarized in Section 4.3, five expert annotators evaluated descriptions across the four cognitive dimensions identified in our framework, with scores aggregated through majority voting to ensure reliability.

For comparative LLM evaluation, we selected 10 representative models covering diverse archi-

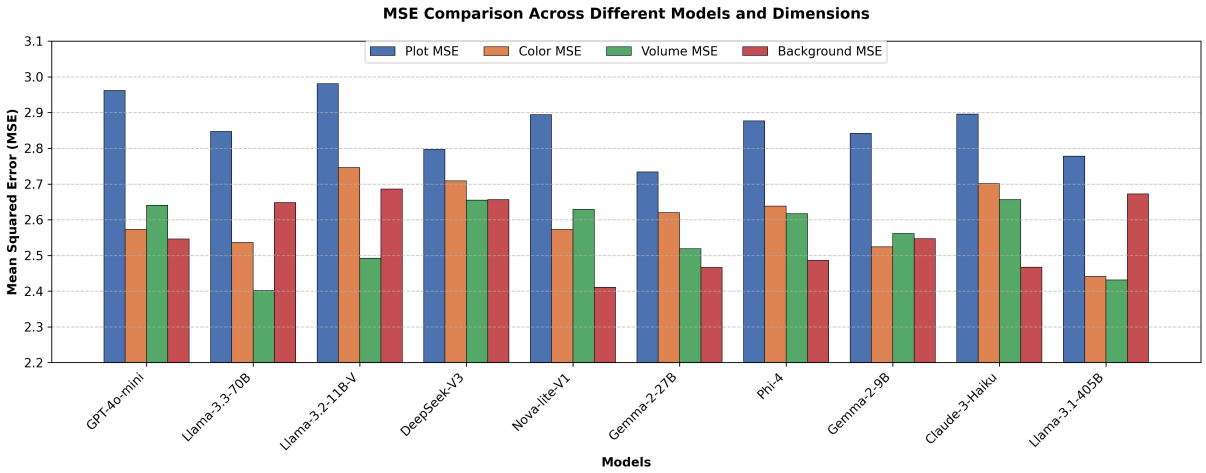

Figure 3: Comparison of Mean Squared Error (MSE) across ten different LLMs on four image generation dimensions: plot, color, volume, and background. Lower MSE values indicate better performance in maintaining fidelity to the intended creative description.

tectures and parameter scales, including Claude, LLaMA, Mistral, and GPT variants (see Appendix C for the complete model list). We measured their alignment with human creativity judgments using Spearman correlation, Mean Squared Error (MSE), classification accuracy, and F1 scores (detailed in Appendix B).

It is important to note that our C-CoC framework differs fundamentally from prompt optimization techniques. While prompt optimization focuses on iteratively refining instructions to improve performance , C-CoC emphasizes structured cognitive pathways for creative generation.

5.2 C-CoC Performance

Table 1 presents the performance comparison between the baseline approach and our C-CoC framework across different product categories. The scores (0-100 scale) represent creativity and practical applicability for image generation based on human expert evaluations.

Our C-CoC framework demonstrates consistent improvements across all evaluation dimensions and product categories. The enhancements range from 8.33 to 19.53 percentage points, with Electronics products showing the most substantial gains in the Volume dimension (+19.53). On average, C-CoC achieves a 12.23 percentage point improvement (approximately 20% relative gain) over the baseline approach.

The qualitative example in Table 3 illustrates how C-CoC transforms a conventional product description into a visually rich narrative. While the baseline provides a straightforward depiction of a facial moisturizer, the C-CoC-generated description incorporates metaphorical storytelling, vivid color symbolism, dynamic spatial composition, and contextual environment integration.

Product	Facial moisturizer with SPF protection
Baseline	A bottle of facial moisturizer with SPF protection placed on a white bathroom counter. The product appears in good lighting, showing its label clearly.
C-CoC	A radiant bottle of facial moisturizer with SPF protection stands like a sentinel at the edge of a sun-drenched desert landscape. The warm golden light of dawn bathes the scene, while the moisturizer casts a protective blue aura around a miniature female figure emerging from the bottle. The background transitions from parched, cracked earth to smooth, hydrated skin, visually narrating the product's transformative journey. Crystal-clear droplets of moisture hover in the air, refracting rainbow prisms of light that symbolize both hydration and sun protection.

Table 3: Comparative example of baseline and C-CoC descriptions for a facial moisturizer.

This transformation clearly demonstrates the effectiveness of our cognitive framework in enhancing creative expression while maintaining visual coherence and product relevance.

5.3 LLM Evaluation Analysis

As a secondary exploration, we examined whether current LLMs could effectively evaluate creative descriptions compared to human judgments. The results, summarized in Tables B.2.1 and 6, indicate significant limitations in this capability.

Overall, we observed weak correlations between

LLM and human evaluations across all models and dimensions. The highest Spearman correlation was just 0.112 (for Background dimension), with most values hovering near zero or slightly negative. MSE values averaged around 2.6 on our 0-5 scale, indicating substantial deviations from human judgments.

Among the dimensions, LLMs showed marginally better alignment with human judgments on Background creativity compared to other dimensions. Interestingly, we found no clear correlation between model size and evaluation capability, suggesting that current scaling approaches may not inherently improve creative assessment abilities.

These findings highlight the complexity of the VCD task and indicate that while our C-CoC framework significantly enhances creative description generation, automated evaluation of such descriptions remains challenging and requires further research.

6 Discussion

Our C-CoC framework's effectiveness stems from several key cognitive principles. By structuring creativity into discrete stages (decomposition, base expression, cognitive shift, and realization), it mirrors hierarchical human creative processes (Finke et al., 1996) while maintaining semantic coherence. The framework implements cross-domain conceptual blending (Fauconnier and Turner, 2008), integrating ideas from disparate domains to create novel yet interpretable metaphorical representations. Additionally, C-CoC's multi-dimensional approach (addressing plot, color, volume, and background) enables targeted enhancement across specific aspects while preserving overall coherence.

This structured approach produces significant improvements across all evaluated dimensions, with particularly notable gains in Volume creativity. The structured cognition allows LLMs to systematically enhance spatial arrangements and composition balance—elements crucial for visual impact but often challenging to articulate. By decomposing the creative process, C-CoC guides LLMs through incremental transformations rather than attempting direct leaps from conventional to creative descriptions, resulting in outputs that maintain coherence while introducing novelty.

The VCD generation task introduces unique challenges as a dual optimization problem balancing creative novelty with practical executability. Descriptions must be both imaginative and realizable by T2I systems, requiring careful consideration of both artistic merit and technical feasibility. Our framework addresses this tension through its staged approach, ensuring that creative transformations remain grounded in the original product attributes while introducing metaphorical and narrative elements that enhance visual appeal.

The evaluation of visual creativity itself presents challenges due to its subjective nature and multi-dimensional character. Traditional metrics for text generation fail to capture the nuanced aspects of creative quality, necessitating expert human evaluation across multiple dimensions. Our evaluation methodology provides a structured approach to this assessment, offering a foundation for future work on automated metrics for creative text evaluation.

7 Conclusion

We introduce Visual Creative Description generation as a formal task, addressing a critical gap in creative content generation. Our work makes two primary contributions to computational creativity research. First, we propose the Cognitive Chain-of-Creativity framework, which structures the creative generation process according to established cognitive science principles. By systematically modeling multiple dimensions of visual expression, C-CoC demonstrates significant improvements over baseline approaches across diverse product categories, highlighting the efficacy of cognitively-informed approaches to creative text generation. Second, we develop the PAINT dataset, featuring creative descriptions across numerous product categories with fine-grained annotations along several cognitive dimensions. This resource provides a standardized benchmark for evaluating creative generation systems and enables a more nuanced assessment of computational creativity. Bridging cognitive science and computation, our work advances theory and practice, crucially instilling structured, cognitive reasoning in AI creative processes. Its broader implications include democratizing content creation, offering novel human-benchmarked evaluation paradigms for computational creativity, and fostering nuanced human-AI collaboration. Grounding generation in cognitive principles enhances AI's practical capabilities, deepens theoretical understanding of creativity, and guides future development of versatile, culturally-aware AI co-creators.

Limitation

The PAINT dataset focuses on product advertisements, which may limit the generalizability of our findings to other creative domains such as artistic illustrations or social media content. The complexity of the annotation task, which involves multiple dimensions and requires a voting mechanism to ensure consistency, also restricted the size of the dataset. Future work should aim to expand the dataset to capture a broader range of creative contexts, which will allow for a more comprehensive evaluation of C-CoC's versatility.

Additionally, our evaluation was based solely on textual inputs, overlooking the potential of multimodal models, which combine vision and language. While this approach isolates linguistic creativity, it misses the opportunity to leverage modern vision-language models for grounded aesthetic reasoning, which could yield richer insights. Moreover, all human evaluations were conducted by annotators from similar cultural backgrounds, potentially introducing bias in color symbolism and narrative preference. Cross-cultural validation is essential for more global applications.

Finally, the C-CoC framework employs a linear creative process, whereas human creativity is often recursive, involving ongoing refinement. The staged approach used here may oversimplify the dynamic interactions between concept generation and critical evaluation in more complex creative workflows.

Ethics Statement

This study adheres to all relevant ethical guidelines. The dataset and model utilized are publicly available and employed in accordance with their respective licenses. Ethical standards were followed throughout the annotation process, with informed consent obtained from all annotators. It should be noted that this text was initially drafted with the assistance of an AI language model to enhance its clarity and accuracy. Moreover, we conducted rigorous internal reviews to further guarantee that every step in this study, from data collection to model deployment, strictly adhered to the highest ethical benchmarks.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Rudolf Arnheim. 1954. *Art and visual perception: A psychology of the creative eye*. Univ of California Press.

Moshe Bar. 2004. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629.

Jonas Belouadi and Steffen Eger. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381, Toronto, Canada. Association for Computational Linguistics.

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

Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Şevin Bostancı and Yunus Dursun. 2024. Role of creative advertising on brand image and its effect on perceived quality, brand loyalty and purchase intention. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*.

Marion Botella, Franck Zenasni, and Todd Lubart. 2018. What are the stages of the creative process? what visual art students are saying. *Frontiers in Psychology*, 9:2266.

Jerome Bruner. 1991. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21.

Yang Cao, Wangchunshu Zhou, Jie Li, Yupeng Hu, Chenghao Lin, and Wei Sun. 2023. BeautifulPrompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–10. Association for Computational Linguistics.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

Andrew J Elliot and Markus A Maier. 2014. Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, 65(1):95–120.

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. 2023. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239.

Gilles Fauconnier and Mark Turner. 2008. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic books.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.

Ronald A Finke, Thomas B Ward, and Steven M Smith. 1996. *Creative cognition: Theory, research, and applications*. MIT press.

GemmaTeam. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*.

Jochen Hartmann, Yannick Exner, and Samuel Domdey. 2025. The power of generative marketing: Can generative ai create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1):13–31.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *Preprint*, arXiv:2403.03952.

Amazon Artificial General Intelligence. 2024. The amazon nova family of models: Technical report and model card. *Amazon Technical Reports*.

Lauren I Labrecque and George R Milne. 2012. Exciting red and competent blue: the importance of color in marketing. *Journal of the Academy of Marketing Science*, 40(5):711–727.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. 2024. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*.

MetaAI. 2024a. Introducing llama 3.1: Our most capable models to date.

MetaAI. 2024b. Introducing meta llama 3: The most capable openly available llm to date.

MetaAI. 2024c. Llama 3.2 11b vision instruct model card.

Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. 2024. Striking gold in advertising: Standardization and exploration of ad text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 955–972, Bangkok, Thailand. Association for Computational Linguistics.

Aude Oliva and Antonio Torralba. 2007. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

Stephen E Palmer. 1999. *Vision science: Photons to phenomenology*. MIT press.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Steven M Smith, Thomas B Ward, and Ronald A Finke. 1995. *The creative cognition approach*. MIT press.

Rodrigo Valerio, Joao Bordalo, Michal Yarom, Yonatan Bitton, Idan Szpektor, and Joao Magalhaes. 2023. Transferring visual attributes from natural language to verified image generation. *arXiv preprint arXiv:2305.15026*.

Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. 2023. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7766–7776.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 841–852, New York, NY, USA. Association for Computing Machinery.

Jingtao Zhan, Qingyao Ai, Yiqun Liu, Yingwei Pan, Ting Yao, Jiaxin Mao, Shaoping Ma, and Tao Mei. 2024. Prompt refinement with image pivot for text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 941–954, Bangkok, Thailand. Association for Computational Linguistics.

,

A PAINT Dataset

A.1 Dataset Composition

The PAINT dataset comprises 330 entity expressions derived from 33 product categories (10 samples per category). These expressions serve as seed inputs for our creative generation pipeline.

A.2 Generation Framework

From this foundation, we synthesized 1,980 descriptions utilizing our three-phase C-CoC process, structured as follows:

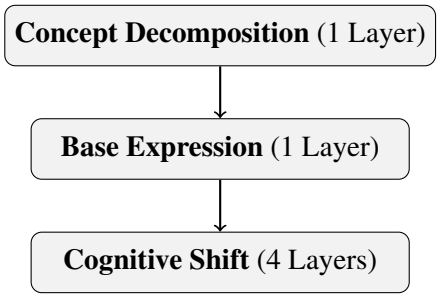

Figure 4: Hierarchical structure of the C-CoC generation process

The six-step generation process is illustrated in Figure 7, showing how input data undergoes decomposition, expression generation, and cognitive shifts.

For each Base Expression and Cognitive Shift, images were generated, resulting in 1650 images. These images primarily served as a reference for the generated descriptions. Human evaluators provided 7920 annotations, where the ratings were mainly focused on the textual descriptions. However, when the image description aligned with the text, the evaluators also considered the aesthetic quality of the images, including factors such as visual appeal, clarity, and relevance to the generated description.

A.3 Diversity Analysis

Metric	Plot	Color	Vol.	Bg.
Distinct-1	**0.804**	0.820	0.814	0.799
Distinct-2	0.978	**0.982**	0.981	0.972
Self-BLEU-1	**0.164**	0.197	0.202	0.225
Self-BLEU-2	0.049	0.061	0.052	**0.096**

Table 4: LEXICAL DIVERSITY BY DIMENSION

The dataset shows high diversity in Distinct-1 and Distinct-2, especially at the phrase level (Distinct-2), with values close to 0.98 for all dimensions, indicating significant variation in the text content and low redundancy.

B Evaluation Metrics

B.1 Cognitive Transition Evaluation Criteria

This study designs four independent evaluation criteria based on four cognitive transition dimensions $\theta_k \in \{\theta_{\text{plot}}, \theta_{\text{color}}, \theta_{\text{volume}}, \theta_{\text{background}}\}$. For each expression object $i \in I$, the scores before and after cognitive transition are calculated for each dimension θ_k.The Prompts are illustrated in Figure5

B.1.1 LLM Scoring Mechanism

For each expression object $i \in I$ and each cognitive transition dimension θ_k, the large language model (LLM) is required to score the text before cognitive transition $T_{\text{base},i}$ and the text after cognitive transition $T_{\text{shift},i,k}$:

$$a_{i,k}^{\text{base}}, \quad a_{i,k}^{\text{shift}}$$

Then, the cognitive transition increment is calculated:

$$\Delta S_{i,k}^{\text{model}} = a_{i,k}^{\text{shift}} - a_{i,k}^{\text{base}}$$

where:

- If $\Delta S_{i,k}^{\text{model}} > 0$, it indicates that the cognitive transition in this dimension has enhanced the creativity of the text.
- If $\Delta S_{i,k}^{\text{model}} < 0$, it indicates that the cognitive transition in this dimension has made the description too abstract or inconsistent with the visual expression logic, lowering the text quality.

B.1.2 Human Annotation and Ground Truth Construction

To ensure the reliability of Ground Truth, each cognitive transition data point is evaluated by three independent annotators $j \in J$. For each expression object $i \in I$, the three individual scores for each cognitive transition dimension θ_k are obtained:

$$(a_{i,k,j}^{\text{base}}, a_{i,k,j}^{\text{shift}}), \quad j \in J$$

The cognitive transition increment for each annotator is then calculated as:

$$\Delta S_{i,k,j}^{\text{human}} = a_{i,k,j}^{\text{shift}} - a_{i,k,j}^{\text{base}}$$

Next, the voting system is applied based on the signs of the increments reported by the annotators:

- If two or more annotators report a positive increment, the cognitive transition is classified as a positive improvement.
- If two or more annotators report a non-positive increment (i.e., negative or zero), the cognitive transition is classified as a negative decrease.

The final cognitive transition score is determined by averaging the scores of the majority vote:

$$\Delta S_{i,k}^{\text{human}} = \frac{1}{|J_{\text{majority}}|} \sum_{j \in J_{\text{majority}}} (a_{i,k,j}^{\text{shift}} - a_{i,k,j}^{\text{base}})$$

where J_{majority} represents the annotators who belong to the majority vote.

For each cognitive transition increment, the classification of positive, negative, or discrepant cases is as follows:

- Positive Increment: If $\Delta S_{i,k}^{\text{model}} > 0$ and $\Delta S_{i,k}^{\text{human}} > 0$, the cognitive transition is classified as a positive improvement.
- Negative Increment: If $\Delta S_{i,k}^{\text{model}} < 0$ and $\Delta S_{i,k}^{\text{human}} < 0$, the cognitive transition is classified as a negative decrease.
- Discrepant Case: If the model and human judgments are in opposite directions, the cognitive transition is classified as discrepant.

B.1.3 Annotator Identity

The annotation process involves five postgraduate students who possess extensive expertise in natural language processing. This ensures the reliability and consistency of the annotated data. Their background in NLP contributes to the rigor and accuracy of the cognitive transition assessments, thereby enhancing the credibility of the annotations.

B.2 Consistency Evaluation Between LLM and Human Ratings

To assess whether the LLM has the ability to judge creative transitions, we calculate the consistency between its scores and human ratings.

B.2.1 Spearman Rank Correlation

For each cognitive transition dimension θ_k, the Spearman Rank Correlation between the LLM score $\Delta S_{i,k}^{\text{model}}$ and human score $\Delta S_{i,k}^{\text{human}}$ is calculated:

$$\rho_k = \frac{\sum_{i \in I}(R_{i,k}^{\text{mod}} - \bar{R}_k^{\text{mod}})(R_{i,k}^{\text{hum}} - \bar{R}_k^{\text{hum}})}{\sqrt{\sum_{i \in I}(R_{i,k}^{\text{mod}} - \bar{R}_k^{\text{mod}})^2} \sqrt{\sum_{i \in I}(R_{i,k}^{\text{hum}} - \bar{R}_k^{\text{hum}})^2}}$$

where:

- $R_{i,k}^{\text{model}}$ and $R_{i,k}^{\text{human}}$ are the rank values of the model and human scores, respectively.
- $\bar{R}_k^{\text{model}}$ and $\bar{R}_k^{\text{human}}$ are the mean values of the model and human scores, respectively.
- A higher ρ_k indicates that the model's scoring is closer to human judgment.

For the complete data, refer to Table**??**.

B.2.2 Mean Squared Error (MSE)

The mean squared error (MSE) between the LLM score and the human score is calculated:

$$\text{MSE}_k = \frac{1}{|I|} \sum_{i \in I} (\Delta S_{i,k}^{\text{model}} - \Delta S_{i,k}^{\text{human}})^2$$

where:

- A lower MSE_k indicates that the LLM's scores are closer to human ratings.

For the complete data, refer to Table 6.

C Testing Model list

We tested the following models in this study:

- **Google**:
  - google/gemma-2-27b-it (GemmaTeam, 2024)
  - google/gemma-2-9b-it (GemmaTeam, 2024)
- **Meta**:
  - meta-llama-3.3-70b-instruct (MetaAI, 2024b)
  - meta-llama-3.2-11b-vision-instruct (MetaAI, 2024c)
  - meta-llama-3.1-405b-instruct (MetaAI, 2024a)
- **Anthropic**:
  - anthropic/claude-3-haiku (Anthropic, 2024)
- **Microsoft**:
  - microsoft/phi-4 (Abdin et al., 2024)
- **Amazon**:
  - amazon/nova-lite-v1 (Intelligence, 2024)

- **Deepseek**:
  - deepseek-V3 (DeepSeek-AI, 2024)
- **OpenAI**:
  - gpt-4o-mini (OpenAI, 2024)

D Annotation Instructions

This section provides a concise overview of the annotation process. Detailed instructions are displayed in Figures 8 and 9, while the Label Studio interface setup is shown in Figure 10.

D.1 Overview

The annotation tasks are divided into four categories: 1. Story Creativity 2. Color Creativity 3. Volume Creativity 4. Background Creativity

Each category has specific evaluation criteria, which include analyzing the text description, comparing it with the corresponding creative picture, and scoring based on creativity and relevance. Annotators are required to follow standardized procedures to ensure consistency and accuracy.

D.2 Scoring Guidelines

The scoring system ranges from 0 to 5 points, as described in the instructions. Higher scores indicate better alignment with task requirements and increased novelty in the described scenes.

Model	Plot	Color	Volume	Background
GPT-4o-mini	-1.7%	0.1%	-3.4%	8.3%
Llama-3.3-70B	-6.3%	2.5%	8.8%	1.8%
Llama-3.2-11B-Vision	-7.1%	-3.1%	4.5%	0.5%
DeepSeek-V3	0.9%	-2.0%	-6.7%	1.5%
Nova-lite-V1	-0.7%	2.0%	-2.9%	7.9%
Gemma-2-27B	4.3%	-1.1%	1.8%	6.7%
Phi-4	-0.4%	-1.1%	-7.9%	5.7%
Gemma-2-9B	-4.2%	3.2%	0.8%	4.3%
Claude-3-Haiku	-0.2%	-3.8%	-5.2%	**11.2%**
Llama-3.1-405B	-2.2%	**4.1%**	1.0%	-3.3%

Table 5: Spearman correlation coefficients (expressed as percentages) across different dimensions. Positive values indicate positive correlation, negative values indicate negative correlation. Highest absolute values are in **bold**.

Model	Plot	Color	Volume	Background	Average
GPT-4o-mini	2.962	2.573	2.640	2.546	2.680
Llama-3.3-70B	<u>2.847</u>	<u>2.536</u>	<u>2.402</u>	2.648	<u>2.608</u>
Llama-3.2-11B-Vision	2.981	2.746	2.492	2.686	2.726
DeepSeek-V3	2.797	2.709	2.655	2.656	2.704
Nova-lite-V1	2.894	2.573	2.629	<u>2.411</u>	2.627
Gemma-2-27B	**2.734**	2.620	2.519	2.466	2.589
Phi-4	2.877	2.638	2.617	2.486	2.654
Gemma-2-9B	2.842	2.524	2.561	2.547	2.618
Claude-3-Haiku	2.896	2.701	2.656	2.467	2.705
Llama-3.1-405B	2.778	**2.441**	**2.431**	2.672	**2.581**

Table 6: Mean Squared Error (MSE) for each model across different creativity dimensions (lower is better). Best results are in **bold**, second best are <u>underlined</u>.

Model	Plot		Color		Volume		Background		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GPT-4o-mini	51.23	51.45	51.69	19.49	44.00	44.51	48.31	37.31	48.33	41.71
Llama-3.3-70B	55.36	<u>66.50</u>	49.84	55.65	<u>59.28</u>	<u>70.45</u>	50.49	59.14	55.53	<u>65.61</u>
Llama-3.2-11B-Vision	45.76	40.57	47.71	19.19	39.68	20.43	43.93	37.36	45.76	40.57
DeepSeek-V3	44.85	31.12	48.00	28.09	38.77	26.57	44.92	19.73	44.85	31.12
Nova-lite-V1	52.38	57.81	52.00	45.07	46.15	50.42	55.38	<u>60.49</u>	52.38	57.81
Gemma-2-27B	<u>55.53</u>	65.61	49.50	52.63	55.37	64.78	<u>54.61</u>	62.70	<u>55.53</u>	65.61
Phi-4	53.19	62.33	46.91	57.43	52.62	60.51	52.92	52.04	53.19	62.33
Gemma-2-9B	55.05	**65.71**	49.69	<u>57.45</u>	57.14	69.47	54.61	62.70	55.05	**65.71**
Claude-3-Haiku	44.24	27.92	49.35	28.96	37.94	15.72	48.31	37.31	44.24	27.92
Llama-3.1-405B	**54.41**	64.86	**51.69**	**60.94**	**53.62**	**64.96**	**53.61**	**64.96**	**54.41**	64.86

Table 7: Classification performance (Accuracy and F1 score, %) across different creative dimensions. Best results are in **bold**, second best are <u>underlined</u>.

LLM Evaluation Prompts

Creative Plot Evaluation:
text_A and text_B have slight differences. Please choose (text_A or text_B) as the more suitable text for image advertisement description, rate it, and briefly explain the reason.
Score_A: [], Score_B: [], Reason: [].
Scoring Guide: - 0 points: Both texts perform almost equally under the given criterion - 1 point: Slight preference - 2-3 points: Moderate preference - 4-5 points: Strong and confident preference
Evaluation Criteria: 1. Does it introduce novel elements or situations? 2. Does it vividly depict surreal scenes? 3. Do surreal elements align with the overall theme? 4. Does it integrate surreal imagination with realistic foundations? 5. Does it avoid using abstract descriptions (e.g., "beautiful melody")? 6. Is it suitable as an image ad for the following product: {title}?

Color Creativity Evaluation:
text_A and text_B have slight differences. Please choose (text_A or text_B) as the more suitable text for image advertisement description, rate it, and briefly explain the reason.
Score_A: [], Score_B: [], Reason: [].
Scoring Guide: - 0 points: Both texts perform almost equally under the given criterion - 1 point: Slight preference - 2-3 points: Moderate preference - 4-5 points: Strong and confident preference
Evaluation Criteria: 1. Does it creatively combine colors with scenes or objects? 2. Does it use unconventional color terminology? 3. Do the colors enhance visual appeal without disrupting image coherence? 4. Does it avoid using abstract descriptions (e.g., "beautiful melody")? 5. Is it suitable as an image ad for the following product: {title}?

Volume Creativity Evaluation:
text_A and text_B have slight differences. Please choose (text_A or text_B) as the more suitable text for image advertisement description, rate it, and briefly explain the reason.
Score_A: [], Score_B: [], Reason: [].
Scoring Guide: - 0 points: Both texts perform almost equally under the given criterion - 1 point: Slight preference - 2-3 points: Moderate preference - 4-5 points: Strong and confident preference
Evaluation Criteria: 1. Does it use specific terms to describe volume changes? 2. Does it effectively convey interactions between the main subject and other elements? 3. Does the volume of the subject enhance visual impact and highlight product features? 4. Does it avoid using abstract descriptions (e.g., "beautiful melody")? 5. Is it suitable as an image ad for the following product: {title}?

Background Creativity Evaluation:
text_A and text_B have slight differences. Please choose (text_A or text_B) as the more suitable text for image advertisement description, rate it, and briefly explain the reason.
Scoring Guide: - 0 points: Both texts perform almost equally under the given criterion - 1 point: Slight preference - 2-3 points: Moderate preference - 4-5 points: Strong and confident preference
Evaluation Criteria: 1. Does the scene break conventional reality frameworks? 2. Does the background align with the ad theme? 3. Does the scene design exhibit artistic appeal? 4. Does it avoid using abstract descriptions (e.g., "beautiful melody")? 5. Is it suitable as an image ad for the following product: {title}?
Score_A: [], Score_B: [], Reason: [].

Figure 5: LLM evaluation prompts for the four dimensions of creative description quality assessment.

Generation Process Prompts

System Prompt: you are a helpful AI agent
User Prompt 1 (Product Explanation): Explain in one sentence what this is. The product is {title}, and its description is as follows: {description}
User Prompt 3 (Original Image Description): You need to construct an advertisement image, 70 words, directly describing this image, about the product: [{title}]
User Prompt 4 (Plot Transformation): Make the story more magical and surreal, 70 words. Adjust the following description: [{previous}]. Requirements: 1. The transformed content should be an image description, without metaphorical expressions. 2. The image description should be about {title}. 3. Reverse the functionality of objects: give common objects features or functions that do not belong to them. Examples (do not use the ideas in these examples): • A chair that is crying. • A lightbulb with a meadow growing inside. 4. Contradiction and contrast: Combine completely different things in the same image to create strong contrast. Examples (do not use the ideas in these examples): • A block of ice that is burning. • A smiling face with mechanical gears flowing out. 5. Visual fusion and deformation: Combine multiple objects or elements to create surreal visual effects. Examples (do not use the ideas in these examples): • A face blended with a landscape, with eyes as the sky and mouth as a river. • A bird's wings turn into book pages, slowly unfolding.
User Prompt 5 (Color Transformation): Change the colors to make the scene more surreal, 70 words. Adjust the following description: [{previous}] Requirements: 1. The transformed content should be an image description, without metaphorical expressions. 2. The image description should be about {title}. 3. Imagine a color that does not match reality based on the product's characteristics: assign unrealistic colors to ordinary objects to enhance surrealism. Examples (do not use the ideas in these examples): • A red sky with a blue sun floating in it. • Purple trees growing in a green ocean.
User Prompt 6 (Volume Transformation): Change the size of the subject to make the scene more surreal, 70 words. Adjust the following description: [{previous}]. Requirements: 1. The transformed content should be an image description, without metaphorical expressions. 2. The image description should be about {title}. 3. Imagine changing the size ratio of objects based on the product's characteristics, making them look both familiar and bizarre. Examples (do not use the ideas in these examples): • A giant hand holding a forest. • A cup large enough to contain an entire lake.
User Prompt 7 (Background Transformation): Change the scene to make it more surreal, 70 words. Adjust the following description: [{previous}]. Requirements: 1. The transformed content should be an image description, without metaphorical expressions. 2. The image description should be about {title}. 3. Imagine a scene that defies logic based on the product's characteristics: place objects in completely unrealistic environments. Examples (do not use the ideas in these examples): • A house floating in the air. • Ocean waves hitting the ceiling instead of the floor.
User Prompt 8 (Base Transformation to 60-word Prompt): Transform into a text-to-image prompt with descriptive expressions. Use exactly 60 English words. The content to be transformed is: {previous}

Figure 6: Step-by-step generation prompts for creating surreal advertisement descriptions across the four creative dimensions: plot, color, size, and background.

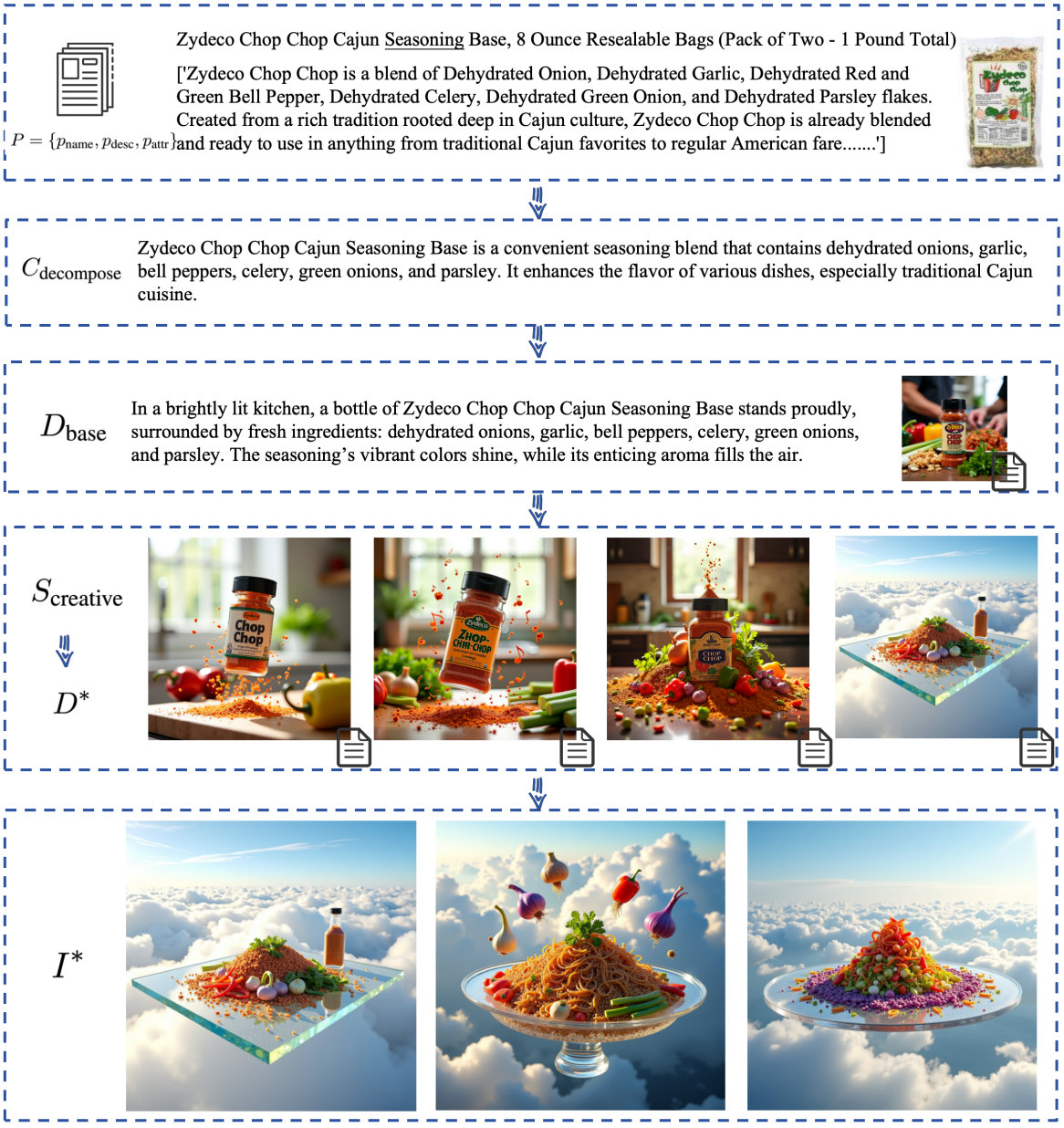

Figure 7: Generation Overview

📌 Annotation Guide

This guide aims to provide a standardized annotation method to ensure the accuracy and consistency of the task. During the annotation process, please follow the following rules and handle them systematically according to the task categories.

🎯 Task Categories

The annotation tasks are divided into **four categories**, each with different evaluation criteria:

1. **Story Creativity**
   - Pay attention to the storytelling, plot development in the picture, and its relevance to the product.
2. **Color Creativity**
   - Focus on color matching, visual impact, and whether the colors enhance the product's attractiveness.
3. **Volume & Shape Creativity**
   - Concentrate on the object shape, product structure, and the expression of volume.
4. **Background & Scene Creativity**
   - Notice the construction of the background environment, the creation of atmosphere, and the interaction between the main body and the background.

📝 Requirements for Annotation Tasks

Each task consists of the following **core parts**:

1. **Introduction to Basic Product Information**
   - Explain the purpose, characteristics, and main functions of the product to ensure that annotators fully understand it.
2. **Description of the Creative Picture Advertisement**
   - Provide the **text description** of the creative picture advertisement for the product.
3. **Schematic Diagram of the Picture Description**
   - Present a schematic diagram to assist in understanding the described scene, ensuring that annotators can accurately evaluate.
4. **Scoring**
   - Score the **quality of the text description** to ensure that the description meets the task requirements.

Figure 8: Annotation instruction1

🏷️ Annotation Methods

1️⃣ Sort by Task Category

- Before formal annotation, it is recommended to **sort by task category (label) first** to improve efficiency and reduce frequent switching between different tasks.

2️⃣ Read Basic Product Information

- Before annotation, carefully read the product introduction to ensure understanding of its functions, features, and market positioning.

3️⃣ Analyze the Picture

- Observe the picture content and compare it with the description to form a preliminary judgment.
- The picture is only used as **auxiliary reference**, and the focus is on evaluating the **quality of the text description**.

4️⃣ Scoring Guidelines

- The scoring is based on the **accuracy of the text description**, not just the aesthetics of the picture.
- **When the picture conforms to the text description**, the aesthetics of the picture can be one of the reference factors for scoring.
- The scores **are comparative in nature**, and the same score should be given when the performance is consistent.

📊 Scoring Criteria

Score	Scoring Criteria
0 points	The described scene has no creativity under this criterion
1 point	The picture description is unremarkable under this criterion
2 - 3 points	Generally meets the requirements but still lacks novelty
4 - 5 points	Meets the requirements and the described scene is refreshing

Figure 9: Annotation instruction2

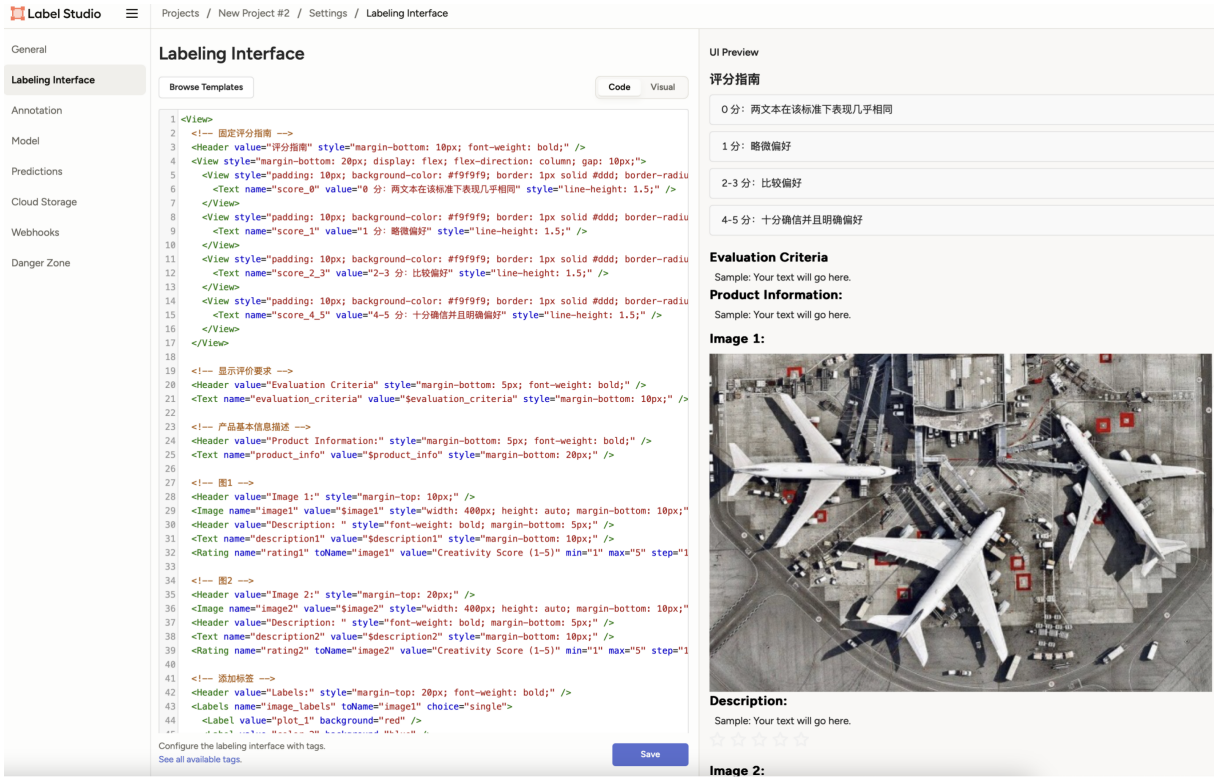

Figure 10: Label Studio Setting