PAINT with Words: A Dataset and Cognitive Framework for Evaluating Visual Creativity

Anonymous ACL submission

Abstract

We introduce the task of visual creative description (VCD) and propose three key innovations: 1) the C-CoC framework for structured creative cognition, 2) the PAINT dataset for systematic training and evaluation, and 3) VCD-Bench, the first multidimensional benchmark for visual creativity. Our experiments on 10 models reveal significant limitations—while models excel in spatial reasoning, they struggle with color and plot evaluation, with these gaps remaining across model sizes. These findings suggest the need for architectural innovations beyond simple parameter scaling.

1 Introduction

In recent years, AI-generated content (AIGC) has seen rapid advancements. Text-to-image (T2I) models can generate highly relevant visual compositions based on textual prompts(Ramesh et al., 2022), while large language models (LLMs) have demonstrated remarkable capabilities in creative text generation, including Storys(Yuan et al., 2022), poetry(Belouadi and Eger, 2023), and advertising copy(Mita et al., 2024). However, a critical question remains underexplored: Can LLMs generate visual creative descriptions (VCDs) that drive the generation of more creative images? Despite the significant progress in both text and image generation, research on LLM-based VCD generation remains largely unexplored.

The fundamental challenge lies in the current paradigm, which treats creative generation as a black-box process, relying on "black-box access" for iterative prompt optimization (He et al., 2024). These methods focus solely on the final output while neglecting the cognitive process of creative ideation. This black-box nature gives rise to three key issues that make VCDs difficult to quantify, optimize, and evaluate:

- Lack of Training Data – Human creativity is implicit and rarely recorded in a structured way. Without high-quality datasets, LLMs struggle to generate novel and visually actionable descriptions, leading to unoriginal and impractical outputs.

- Lack of Cognitive Modeling – Current methods do not model LLMs' cognitive mechanisms in VCD generation, preventing effective simulation of human creative thinking, conceptual associations, and iterative refinements, hindering synthetic dataset creation.

- Lack of Evaluation Frameworks – The absence of cognitive modeling results in inadequate evaluation metrics. Existing NLP and T2I metrics fail to capture key attributes of creative descriptions, making quality assessment difficult.

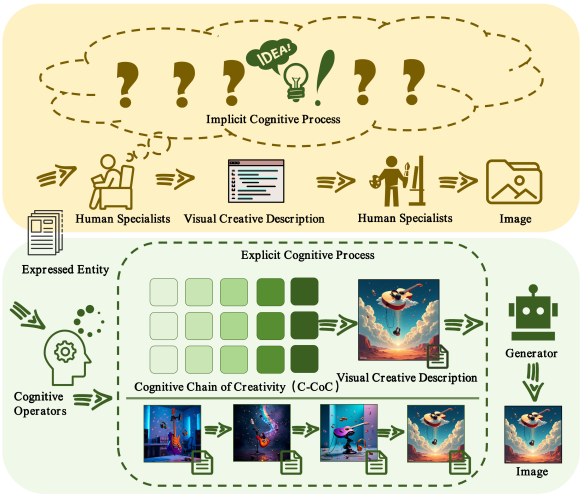

Figure 1: Visual Creative Description Task: Comparison between Human Process and the C-CoC Framework.

To address these challenges, we draw inspiration from cognitive science theories and propose a new generation framework that is shown in Figure1 for visual creative description—**Cognitive Chain of Creativity (C-CoC)**. In this framework, Large

Language Models (LLMs) replace Human Specialists as *Cognitive Operators*, executing *Cognitive Transitions* across multiple reasoning stages to simulate human thought processes in visual creative ideation.

Our framework starts with structured *Expressed Entity* information and progressively generates visually creative textual descriptions that align with communication principles. These descriptions then drive *text-to-image (T2I) generation*, enhancing both the creativity and visual expressiveness of the generated images. This approach not only improves the quality of creative image generation but also introduces an evaluable and optimizable method for creative generation.

To mitigate the issue of missing training data, we construct and annotate the **PAINT (Product Artistic Image Narrative Texts)** dataset based on the C-CoC framework. PAINT is a dedicated dataset for visual creative description tasks, providing structured descriptions that exhibit *novelty, executability, and communicability*. By filling the gap in training data, PAINT enhances the generalization capability of LLMs in visual creative description tasks.

Furthermore, we introduce **VCD-Bench (Visual Creative Description Benchmark)**, an evaluation framework specifically designed to assess the understanding and evaluation capabilities of LLMs in visual creative description. Our focus includes the following core questions:

- Do LLMs' evaluations of creative descriptions align with human judgments?
- How do LLMs differ from human evaluators across various dimensions of visual creative description?
- Are there significant differences among different LLMs in their evaluation of creative descriptions across various dimensions?

The key contributions of this study include:

1. **C-CoC Framework:** We propose C-CoC, an LLM-based generation approach for visual creative description. This approach employs cognitive modeling to establish a multi-stage reasoning mechanism, enhancing the interpretability and controllability of LLMs in creative text generation.
2. **PAINT Dataset:** We construct the PAINT dataset, leveraging C-CoC to generate large-scale, high-quality visual creative descriptions. This dataset addresses the training data gap in LLM-based visual creativity tasks and provides standardized data support for model training and evaluation.
3. **VCD-Bench Benchmark:** We introduce VCD-Bench, the first benchmark dedicated to assessing LLM's visual creative capabilities. By comparing LLM-generated scores with human evaluations, VCD-Bench quantitatively measures performance in visual creative description tasks, offering a systematic evaluation framework for future research.

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, my 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:

- **Generative Phase**: The cognitive system constructs an initial structured representation based on existing knowledge, forming a stable foundation for subsequent transformation.
- **Exploratory Phase**: The system refines, restructures, or blends 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.

To address this, we propose the *Cognitive Chain-based Creativity (C-CoC) framework*, which decomposes creative cognition into four interconnected stages:

1. **Concept Decomposition** ($C_{\textbf{decompose}}$): Constructs symbolic core representations from input data. This aligns with *mental representation theory* (Barsalou, 1999), which describes how cognitive systems extract stable conceptual structures to facilitate downstream reasoning.

2. **Base Expression** ($D_{\textbf{base}}$): Generates initial descriptions that conform to logical and semantic structures, incorporating *cognitive frame theory* (Minsky et al., 1974), which explains how prior knowledge structures influence interpretation and organization.

3. **Cognitive Shift** ($S_{\textbf{creative}}$): Facilitates cross-domain conceptual blending to introduce novelty, inspired by *conceptual blending theory* (Fauconnier and Turner, 2008), which describes creativity as a process of integrating diverse conceptual spaces.

4. **Creative Realization** (D^*): Transforms abstract creative output into actionable, executable descriptions. This process follows principles of *top-down predictive processing* (Friston, 2005; Clark, 2013), ensuring generated content is both interpretable and applicable.

Compared to prior LLM-based creative generation methods, C-CoC provides a structured cognitive pathway that guides the generative process from stable representations to creative transformations. By explicitly modeling cognitive shifts and integrating structured processing, our approach enhances both the controllability and novelty of generated content, moving beyond traditional heuristic-driven methods.

3 Cognitive Chain of Creativity Framework

Visual Creative Description (VCD) is the task of generating textual descriptions that depict creative image content. The goal is to produce descriptions that are not only creative, feasible, and visually expressive but also align with the expected aesthetic and narrative requirements of the generated image. Fundamentally, this process involves a structured cognitive transformation, where the description must balance attribute preservation, conceptual breakthrough, and aesthetic alignment to ensure that the generated text both conveys the key

information of the subject and inspires high-quality visual generation.

In this work, we propose the **Cognitive Chain of Creativity (C-CoC)**, a paradigm that employs large language models (LLMs) as **Cognitive Operators** to model creative image generation as a stepwise *cognitive transition*process. Beginning with the structured information of an **Expressed Entity**, C-CoC systematically performs structured cognitive transformations across multiple reasoning stages, progressing from conceptual association to creative optimization. The final output is a visual creative description that ensures high-quality text-to-image (T2I) generation.

3.1 Task Definition

Given the fundamental information of an Expressed Entity, we define:

$$P = \{p_{\text{name}}, p_{\text{desc}}, p_{\text{attr}}\} \tag{1}$$

where:

- p_{name} represents the entity name,
- p_{desc} provides a brief textual description,
- $p_{\text{attr}} = \{a_1, a_2, ..., a_n\}$ denotes a set of entity attributes (e.g., color, material, functionality).

The system aims to generate a visual creative description D^* that satisfies the criteria of creativity, readability, and executability:

$$D^* = f(P, C) \tag{2}$$

where C denotes the constraints imposed by the **Cognitive Chain of Creativity (C-CoC)** paradigm, integrating knowledge from cognitive psychology, art theory, and visual composition principles. These constraints ensure that the generated descriptions are not only creative but also executable and interpretable by a T2I system. The final description D^* is then used to drive a T2I generator G, producing the expected image I^*:

The process of generating visual creative description is decomposed into four cognitive transition stages(3), where each stage applies structured reasoning and transformation through LLMs as cognitive operators. The overall workflow is summarized in Algorithm 1.

$$P \rightarrow C_{\text{decompose}} \rightarrow D_{\text{base}} \rightarrow S_{\text{creative}} \rightarrow D^* \tag{3}$$

Algorithm 1 Cognitive Chain-of-Thought 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^*

3.1.1 Concept Decomposition ($C_{\text{decompose}}$)

Extracts core concepts by analyzing entity attributes:

$$C_{\text{decompose}} = \Phi(P, \mathcal{M}) \tag{4}$$

where $\Phi(\cdot)$ identifies the key conceptual elements from the entity's attribute set.$\mathcal{M}$ denotes the cognitive operator.

3.1.2 Base Expression (D_{base})

Generates a conventional visual description that follows traditional description paradigms:

$$D_{\text{base}} = \Psi(C_{\text{decompose}}, \mathcal{M}) \tag{5}$$

where $\Psi(\cdot)$ transforms the extracted conceptual elements into a logically sound, attribute-aligned base description. This stage ensures that the generated text remains coherent and factually aligned with the entity but does not yet introduce creativity enhancements.

3.1.3 Cognitive Shift (S_{creative})

Applies constraints from cognitive science, psychology, and artistic principles using cognitive operators to transform the base description into a highly creative visual expression:

$$S_{\text{creative}} = \Omega(D_{\text{base}}, \Theta, \mathcal{M}) \tag{6}$$

where:Θ represents the constraints guiding the cognitive transition,$\Omega(\cdot)$ applies cognitive shifts to ensure the generated description aligns with both visual communication rules and creativity principles.

3.1.4 Creative Realization (D^*)

3.1.5 Creative Realization (D^*)

Produces the final creative description by incorporating **composition aesthetics, narrative logic, and stylistic consistency**:

$$D^* = \Gamma(S_{\text{creative}}) \tag{7}$$

where $\Gamma(\cdot)$ ensures that the transformed creative expression is both actionable and interpretable by the T2I model.

3.2 Multimodal Alignment

The ultimate goal of visual creative description generation is to ensure that D^* effectively drives text-to-image (T2I) generation, producing a visually aligned output I^*. Instead of relying on human specialists for illustration, we leverage T2I models to simulate the drawing process, enabling automated creative visualization.

$$I^* = G(D^*) \tag{8}$$

where G represents the T2I generator.

4 Visual Creative Description Benchmark

4.1 Dataset Construction

VCD Task involves generating textual descriptions that depict visual entities while ensuring *creativity, executability*, and *visual expressiveness*. One concrete application of VCD is **creative advertisement generation**, where 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. PAINT provides high-quality text-image pairs to study and evaluate *cognitive shifts* and *multimodal alignment* within the VCD task. The dataset is systematically constructed following the C-CoC) framework to ensure consistency, control, and quality.

4.1.1 Data Sources

Expressed Entity Datais sourced from publicly available datasets.We utilize the **Amazon Reviews dataset**(Hou et al., 2024) collected by McAuley Lab in 2023. data preprocessing includes:

Figure 2: Dataset Composition: The dataset consists of three parts: text descriptions, model evaluations, human evaluations, and visually aligned images.

- Filtering out samples containing discriminatory or inappropriate content to ensure ethical compliance.

- Extracting product `title` and `description` as input while considering samples with missing or insufficient descriptions.

4.1.2 Data Generation

The dataset is generated under the constraints of the **C-CoC framework**, leveraging large language models (LLMs) for cognitive transitions and state-of-the-art text-to-image (T2I) models for multimodal alignment that is shown in Figure2.

Cognitive Operators We employ ChatGPT-4o-mini (OpenAI, 2024) as the LLM to execute cognitive reasoning within the C-CoC framework and generate visual creative descriptions.

Cognitive Shift Modeling The cognitive shift process in visual creative description generation requires domain-specific adaptation. In the case of advertisement-based VCD tasks, the transformation must align with advertising principles and artistic design conventions to ensure that the generated text both meets creative requirements and optimally guides T2I models.

We introduce four key cognitive shift constraints based on theories from cognitive psychology (Bruner, 1991; Green and Brock, 2000), Gestalt visual principles (Arnheim, 1954; Palmer, 1999), and visual communication (Bar, 2004; Oliva and Torralba, 2007):

$$\Theta = \{\theta_{\text{plot}}, \theta_{\text{color}}, \theta_{\text{volume}}, \theta_{\text{background}}\} \quad (9)$$

- **Plot Creativity** (θ_{plot}): Inspired by narrative psychology (Bruner, 1991; Green and Brock, 2000), this constraint ensures that the advertisement leverages storytelling techniques to enhance audience engagement and information retention.
- **Color Creativity** (θ_{color}): Based on color psychology (Elliot and Maier, 2014; Labrecque and Milne, 2012), this constraint enhances emotion-driven branding by leveraging specific color associations.
- **Volume Creativity** (θ_{volume}): Rooted in Gestalt theory (Arnheim, 1954; Palmer, 1999), this constraint regulates the spatial dominance of visual elements, enhancing focal hierarchy and image composition.
- **Background Creativity** ($\theta_{\text{background}}$): Derived from scene semantics (Bar, 2004; Oliva and Torralba, 2007), this constraint ensures that object-background relationships reinforce visual coherence and symbolic meaning.

Multimodal Alignment To ensure the dataset's cross-modal consistency, we employ the open-source text-to-image model FLUX.1-dev(Black-Forest-Labs, 2024) for alignment.

Evaluation Metrics We set scoring metrics for the four dimensions to evaluate their impact on the overall creative output. These metrics ensure that the advertisement adheres to cognitive shift principles and meets creative requirements. The detailed criteria for these scoring metrics can be found in the AppendixB.

Human Annotation Each cognitive transition is evaluated by three annotators. Scores are given for each dimension, and a majority vote determines the final score. If model and human judgments differ, the transition is marked as discrepant. The voting system reflects general public aesthetic preferences, addressing the subjective nature of aesthetic judgments.

5 Experimental Analysis

We employed VCD-BENCH to evaluate the performance of large language models (LLMs) in assessing visual creative descriptions. This experiment investigates the models' capacity to infer visual concepts based solely on textual descriptions, given their lack of direct visual perception. The evaluation is conducted using three key metrics: (1) consistency with human ratings (Spearman correlation), (2) numerical deviation from human scores (Mean Squared Error, MSE), and (3) classification accuracy in distinguishing high-quality from low-quality descriptions (F1-score and Accuracy). The goal is to assess the ability of LLMs to infer aesthetic and compositional quality from linguistic cues alone.

5.1 Consistency with Human Ratings

We used the Spearman correlation coefficient to assess the alignment between LLM-generated evaluations and human ratings across four dimensions: plot coherence (plot), color composition (color), spatial volume (volume), and background detail (bg). The results, shown in Figure 4, indicate that all models continue to show relatively low correlation scores, suggesting that LLMs still face challenges in fully capturing human evaluative criteria in visual creativity assessment.

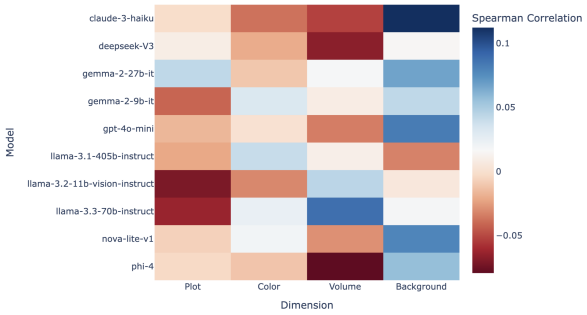

Figure 4: Heatmap of Spearman Correlation

Among the models, **gemma-2-27b-it** achieves the highest average Spearman correlation across all dimensions, with $\rho = 0.031259$ for plot, $\rho = -0.027016$ for color, $\rho = 0.017742$ for volume, and $\rho = 0.071857$ for background. In contrast, **phi-4** exhibits poor correlation in the volume dimension with $\rho = -0.098203$, though it still achieves moderate correlations for plot and background. For the complete data, please refer to Appendix B.2.1.

Breaking down the performance across dimensions:

- Volume dimension: **llama-3.3-70b-instruct** achieves the highest correlation ($\rho = 0.084856$), indicating a relatively better grasp

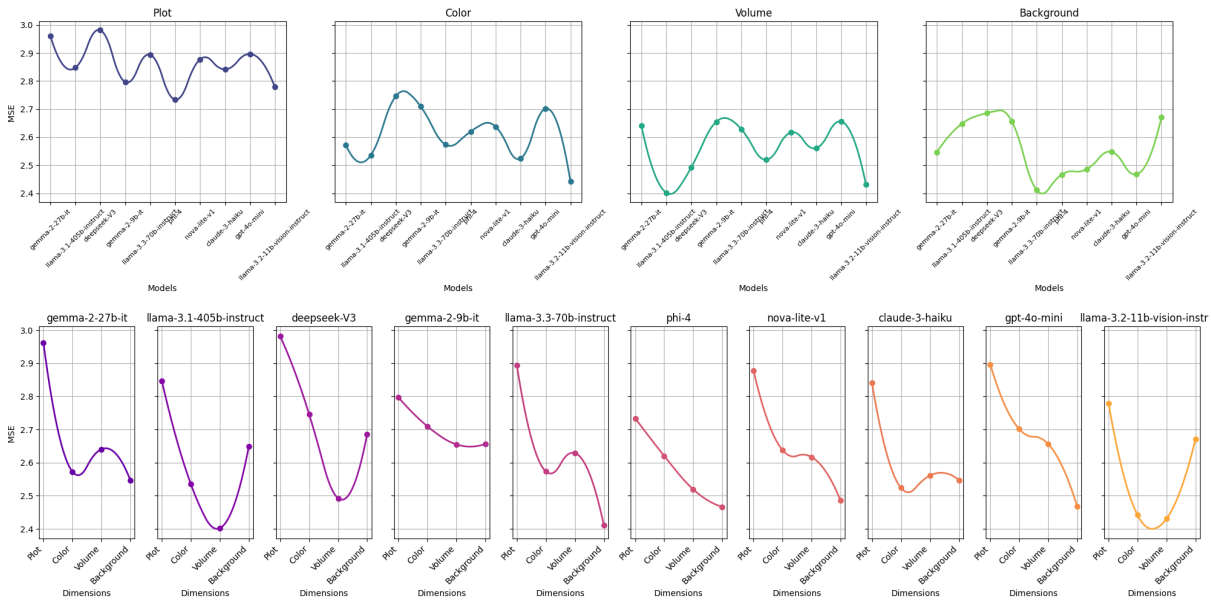

Figure 3: Evaluation scores MSE by Dimension and Model

Dimension	Plot				Color				Volume				Background				Overall			
Model	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
gpt-4o-mini	51.23	60.32	44.86	51.45	51.69	50.00	12.10	19.49	44.00	60.83	35.10	44.51	48.31	58.14	27.47	37.31	48.33	56.23	32.27	41.71
llama-3.3-70b-instruct	55.36	58.72	76.66	66.50	49.84	48.73	64.86	55.65	59.28	65.64	76.02	70.45	50.49	55.56	63.22	59.14	55.53	59.40	73.27	65.61
llama-3.2-11b-vision-instruct	45.76	54.05	32.47	40.57	47.71	35.85	13.10	19.19	39.68	58.54	12.37	20.43	43.93	49.51	30.00	37.36	45.76	54.05	32.47	40.57
deepseek-V3	44.85	55.48	21.63	31.12	48.00	42.31	21.02	28.09	38.77	57.14	17.31	26.57	44.92	53.66	12.09	19.73	44.85	54.48	21.63	31.12
nova-lite-v1	52.38	59.05	56.61	57.81	52.00	50.39	40.76	45.07	46.15	61.38	42.79	50.42	55.38	60.00	60.99	60.49	52.38	59.05	56.61	57.81
gemma-2-27b-it	55.53	59.40	73.27	65.61	49.50	48.02	58.22	52.63	55.37	65.97	63.64	64.78	54.61	58.59	67.44	62.70	55.53	59.40	73.27	65.61
phi-4	53.19	58.08	67.25	62.33	46.91	46.77	74.36	57.43	52.62	64.84	56.73	60.51	52.92	60.58	45.60	52.04	53.19	58.08	67.25	62.33
gemma-2-9b-it	55.05	58.45	75.03	65.71	49.69	48.21	71.05	57.45	57.14	63.82	76.21	69.47	54.61	58.59	67.44	62.70	55.05	58.45	75.03	65.71
claude-3-haiku	44.24	54.69	18.74	27.92	49.35	45.07	21.33	28.96	37.94	58.06	9.09	15.72	48.31	58.14	27.47	37.31	44.24	54.69	18.74	27.92
llama-3.1-405b-instruct	54.41	58.88	72.20	64.86	51.69	50.98	75.73	60.94	53.62	61.38	68.99	64.96	53.61	61.38	68.99	64.96	54.41	58.88	72.20	64.86

Table 1: Overall and Per-Dimension Classification Evaluation Results (%)

of spatial reasoning and object fullness in text descriptions.

- Background dimension: **claude-3-haiku** shows the strongest alignment ($\rho = 0.091889$), suggesting it may be more sensitive to textual cues related to scene depth, environmental detail, or artistic framing.

- Plot and color dimensions: Correlation scores remain generally low, with **llama-3.3-70b-instruct** achieving $\rho = -0.052644$ for plot and $\rho = 0.005578$ for color. This highlights a significant limitation in LLMs' ability to infer narrative structure and aesthetic color harmony from text alone.

A key observation is that some models perform well in specific dimensions but struggle in others. For example, **phi-4** shows moderate correlation in the background dimension ($\rho = 0.052088$), but negative correlations in plot and volume, indicating that LLMs exhibit non-uniform proficiency across different visual attributes.

5.2 Numerical Deviation from Human Scores

We further examined the numerical accuracy of LLM-generated scores by computing the Mean Squared Error. A lower MSE indicates that the model's absolute scoring is closer to human ratings.

gemma-2-27b-it achieves the lowest average MSE (**2.5339**), demonstrating the smallest deviation from human scores, followed by **llama-3.3-70b-instruct** (2.5874). In contrast, **claude-3-haiku** and **deepseek-V3** exhibit significantly higher MSE values, suggesting greater difficulty in predicting scores that align with human evaluations.

5.3 Binary Classification Performance

In the binary classification task of distinguishing high-quality from low-quality descriptions, **llama-3.3-70b-instruct** achieves the highest **F1-score**

(66.50%) and **Accuracy (55.36%)**, making it the most stable classifier. **gemma-2-27b-it** and **llama-3.1-405b-instruct** also perform well (F1 $\approx$ 65.61%), but with slightly higher recall than precision, suggesting a tendency to over-predict the "good" class.

Breaking down performance by dimension:

- **Plot dimension: llama-3.1-405b-instruct** achieves the highest F1-score (0.7469), indicating strong text-based reasoning in assessing narrative structure.
- **Volume dimension: llama-3.3-70b-instruct** performs best (F1 = 0.7045), consistent with its superior correlation in this category.
- **Background dimension: gemma-2-27b-it** achieves the highest F1-score (0.6761), suggesting it is more effective at differentiating high- and low-quality background descriptions.
- **Color dimension:** All models perform significantly worse in this dimension, with the best F1-score reaching only 0.583, indicating that LLMs struggle to evaluate color-related aesthetics using only textual descriptions.

6 Discussion

Our experimental analysis provides critical insights into the capabilities and limitations of LLMs in evaluating visual creative descriptions. The consistently low Spearman correlations across all models (averaging $\rho < 0.1$) indicate a fundamental misalignment between LLM-based assessments and human judgment in creative evaluation tasks. This discrepancy is particularly evident in the color (best $\rho = 0.040$) and plot (worst $\rho = -0.098$) dimensions, suggesting that LLMs face significant challenges in understanding abstract aesthetic principles and narrative coherence from textual descriptions alone.

Dimension-specific performance patterns reveal an intriguing dichotomy: models achieved relatively better mean squared error (MSE) scores in volume (2.402–2.655) and background (2.411–2.686) compared to color (2.441–2.746) and plot (2.734–2.981). We hypothesize that this may stem from spatial attributes being more amenable to linguistic encoding (e.g., the phrase "dominant foreground object" implies control over volume), whereas color harmony and narrative structure rely on implicit visual knowledge that text-based models cannot reliably access.

Furthermore, the classification results expose fundamental limitations in current LLM architectures. While llama-3.3-70b-instruct achieved a 66.50% F1-score in volume classification—indicating moderate competence in judging spatial composition—all models performed near chance levels (45–55% accuracy) in color assessment. This suggests that without explicit visual perception mechanisms, LLMs cannot reliably simulate human aesthetic judgment for certain creative dimensions, despite their strong linguistic capabilities.

Even state-of-the-art models remain far from human-level performance. This gap underscores the need for new training paradigms that incorporate cognitive principles of visual creativity.

7 Conclusion

We introduce visual creative description tasks and present three key innovations to address them:

- 1) The C-CoC framework for structured creative cognition modeling.
- 2) The PAINT dataset enabling systematic training and evaluation.
- 3) VCD-Bench, the first multidimensional benchmark for visual creativity assessment.

Our evaluation of 10 models highlights critical limitations across all dimensions. While the models show relatively better performance in spatial reasoning compared to other aspects, they all fall short in color and plot evaluation. These gaps persist consistently across different model sizes, suggesting that architectural innovations are necessary beyond just scaling parameters. Future work should focus on integrating lightweight visual encoders and cognitive alignment objectives. By bridging linguistic and visual creativity, this work aims to push the boundaries of creative collaboration between models and human-like cognitive processes.

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.

Ethical 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.

LW Barsalou. 1999. Perceptual symbol systems. *The Behavioral and brain sciences/Cambridge University Press*.

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.

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

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.

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

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.

Karl Friston. 2005. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

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*.

Yutong He, Alexander Robey, Naoki Murata, Yiding Jiang, Joshua Nathaniel Williams, George J. Pappas, Hamed Hassani, Yuki Mitsufuji, Ruslan Salakhutdinov, and J. Zico Kolter. 2024. Automated black-box prompt engineering for personalized text-to-image generation. *Preprint*, arXiv:2403.19103.

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.

Marvin Minsky et al. 1974. A framework for representing knowledge.

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

This section provides an overview of the dataset . The dataset consists of 330 input expressed entity, which are derived from 33 different product categories. For each category, 10 samples were selected, totaling 330 input items. These items are used as the starting point for generating creative descriptions through the C-CoC process.

The dataset was used to generate 1980 descriptions through a six-step C-CoC process. The six steps are:

1. **Concept Decomposition (1 Layer)**
2. **Base Expression (1 Layer)**
3. **Cognitive Shift (4 Layers)**

The six-step generation process is illustrated in Figure 5, 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.1 Diversity Analysis

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.

Metric	plot	color	volumn	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 2: Metrics for Text Diversity Analysis

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 Figure6

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:

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

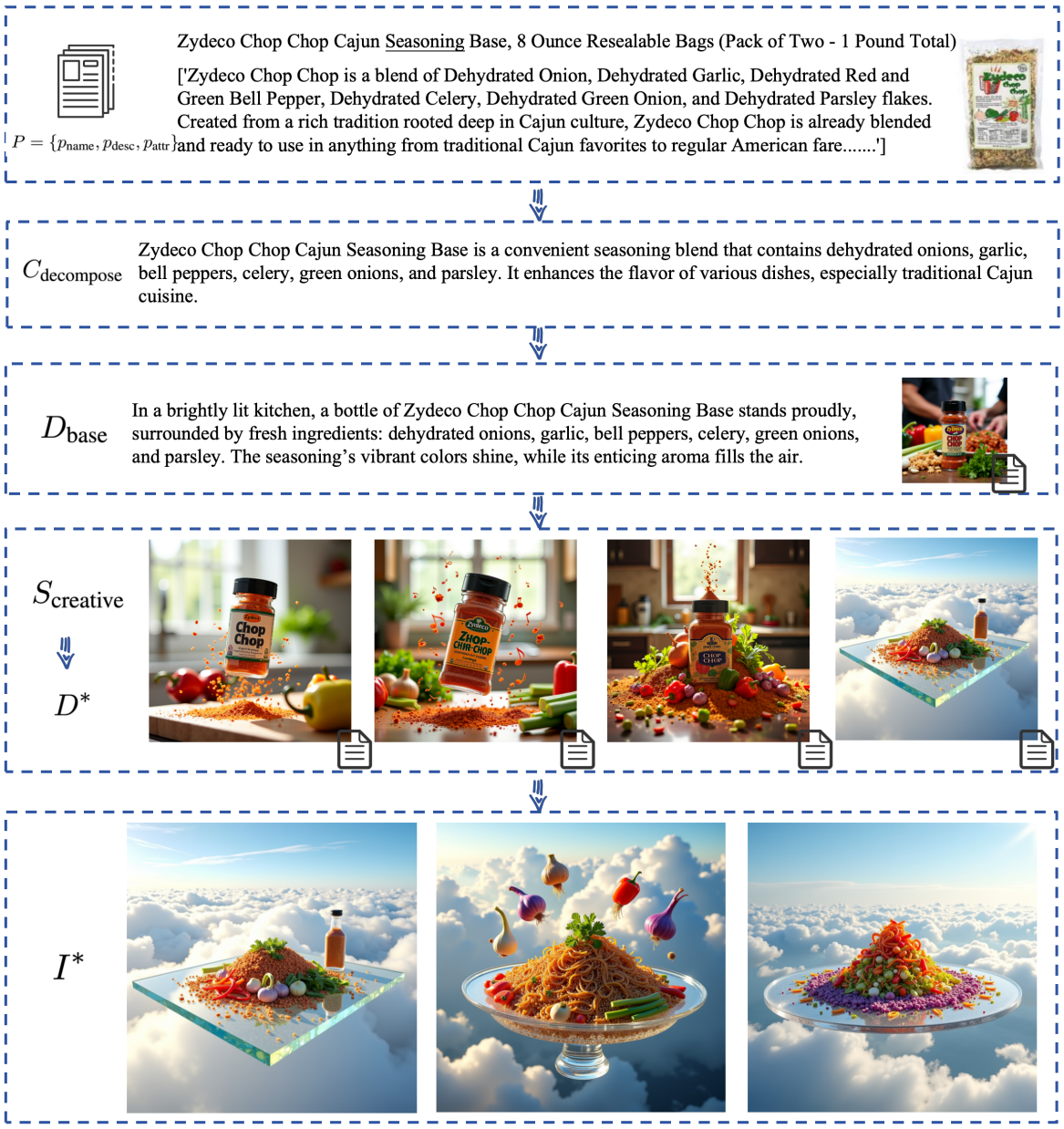

Figure 5: Generation Overview

Plot

1. Does it introduce novel elements or scenarios?
2. Does it vividly depict a unique or strange scene?
3. Are the unique elements aligned with the overall theme?
4. Does it incorporate imaginative elements based on reality?
5. Does the description avoid abstract expressions (e.g., "beautiful melody")?
6. Is it suitable as an image advertisement for the following product: `{title}`?

Color

1. Does it introduce novel elements or scenarios?
2. Does it vividly depict a unique or strange scene?
3. Are the unique elements aligned with the overall theme?
4. Does it incorporate imaginative elements based on reality?
5. Does the description avoid abstract expressions (e.g., "beautiful melody")?
6. Is it suitable as an image advertisement for the following product: `{title}`?

Volume

1. Does it introduce novel elements or scenarios?
2. Does it vividly depict a unique or strange scene?
3. Are the unique elements aligned with the overall theme?
4. Does it incorporate imaginative elements based on reality?
5. Does the description avoid abstract expressions (e.g., "beautiful melody")?
6. Is it suitable as an image advertisement for the following product: `{title}`?

Background

1. Does it introduce novel elements or scenarios?
2. Does it vividly depict a unique or strange scene?
3. Are the unique elements aligned with the overall theme?
4. Does it incorporate imaginative elements based on reality?
5. Does the description avoid abstract expressions (e.g., "beautiful melody")?
6. Is it suitable as an image advertisement for the following product: `{title}`?

Figure 6: Evaluation Prompt

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 Table3.

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 4.

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 7 and 8, while the Label Studio interface setup is shown in Figure 9.

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	-0.016659	0.000775	-0.034206	0.082822
llama-3.3-70b-instruct	-0.063425	0.024775	0.088258	0.018349
llama-3.2-11b-vision-instruct	-0.071466	-0.031215	0.044520	0.004729
deepseek-V3	0.008927	-0.020354	-0.067282	0.015250
nova-lite-v1	-0.006940	0.019765	-0.028856	0.078937
gemma-2-27b-it	0.042933	-0.010640	0.017705	0.067116
phi-4	-0.003633	-0.011429	-0.079310	0.056719
gemma-2-9b-it	-0.041644	0.032320	0.008456	0.042995
claude-3-haiku	-0.001715	-0.037606	-0.051797	0.112270
llama-3.1-405b-instruct	-0.021820	0.040628	0.009954	-0.032808

Table 3: Spearman Correlation by Model and Dimension

Model	Plot MSE	Color MSE	Volume MSE	Background MSE	Average MSE
gpt-4o-mini	2.962	2.573	2.640	2.546	2.680
llama-3.3-70b-instruct	2.847	2.536	2.402	2.648	2.608
llama-3.2-11b-vision-instruct	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	2.411	2.627
gemma-2-27b-it	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-it	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-instruct	2.778	2.441	2.431	2.672	2.581

Table 4: Mean Squared Error (MSE) for Each Model Across Different Dimensions

📌 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 7: 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 8: Annotation instruction2

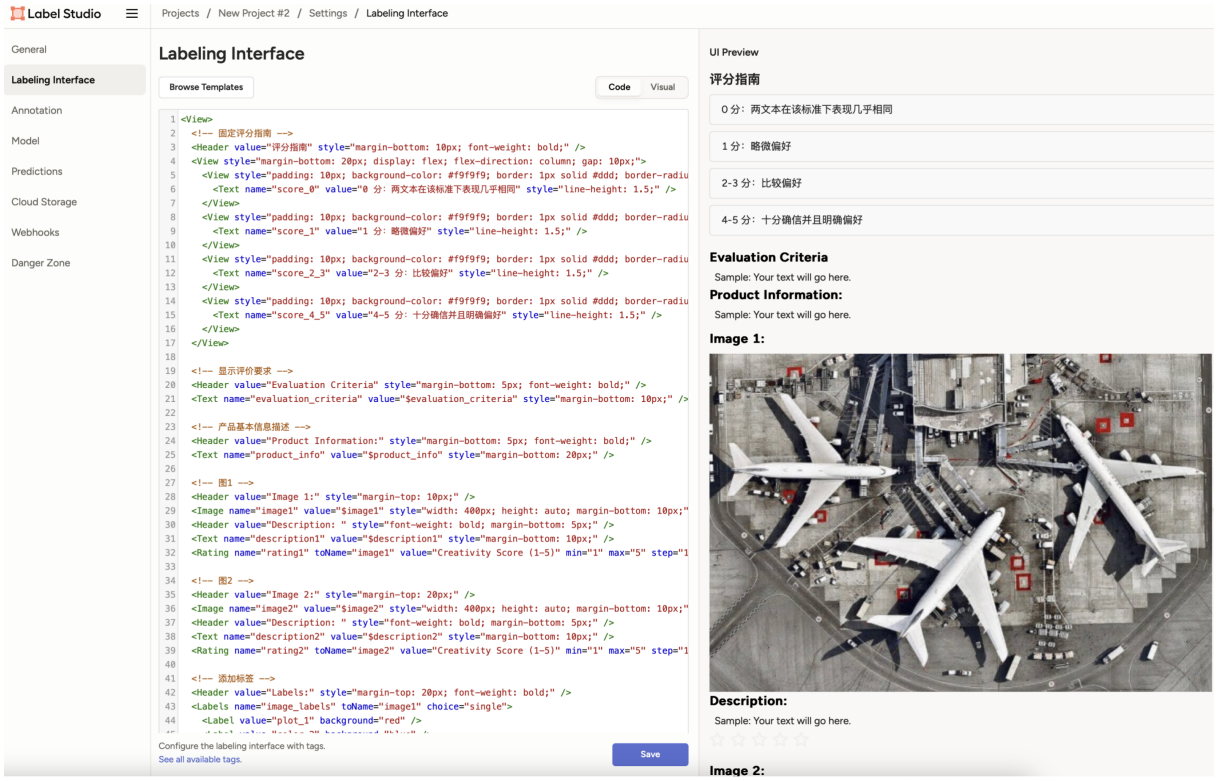

Figure 9: Label Studio Setting