FROM PIXELS TO FEELINGS: ALIGNING MLLMS WITH HUMAN COGNITIVE PERCEPTION OF IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

While Multimodal Large Language Models (MLLMs) are adept at answering "what" is in an image—identifying objects and describing scenes—they often lack the ability to understand "how" an image feels to a human observer. This gap is most evident when considering subjective cognitive properties, such as what makes an image memorable, funny, aesthetically pleasing, or emotionally evocative. To systematically address this challenge, we introduce CogIP-Bench, a comprehensive benchmark for evaluating MLLMs on such image cognitive properties. Our evaluation reveals a significant gap: current models are poorly aligned with human perception of these nuanced properties. We then demonstrate that a post-training phase can effectively bridge this gap, significantly enhancing the model's alignment with human judgments. Furthermore, we show that this learned cognitive alignment is not merely predictive but also transferable to downstream creative tasks. By integrating our cognitively-aligned MLLM into an image generation pipeline, we can guide the synthesis process to produce images that better embody desired traits, such as being more memorable or visually appealing. Our work provides a benchmark to measure this human-like perception, a post-training pipeline to enhance it, and a demonstration that this alignment unlocks more human-centric AI.

The code and data will be made publicly available.

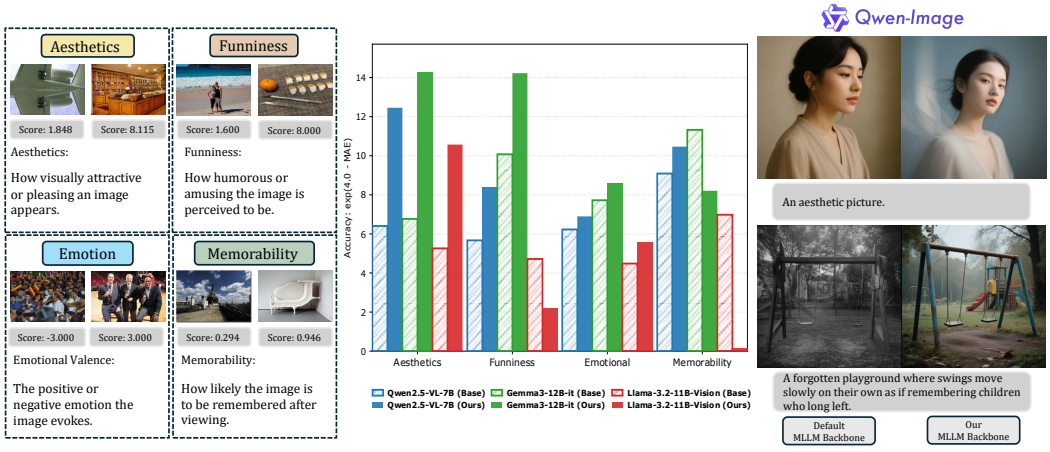


Figure 1: We present CogIP-Bench, a comprehensive cognition benchmark that evaluates the alignment of cognition score prediction between MLLM and humans. **Left:** example datapoints for each dimension: aesthetics, funniness, emotion and memorability. **Middle:** post-training results of three popular MLLMs across different dimensions. **Right:** results of swapping the MLLM backbone, comparing the effect of cognition-related image generation with the Qwen-Image pipeline.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Advances in MLLMs, such as GPT OpenAI et al. (2024) and Gemini Team & Anil (2025), have led to significant breakthroughs in visual understanding. Their success in tasks like visual question answering and detailed image captioning demonstrates a strong ability to perform objective recognition and description Zhou et al. (2020); Jayachandran et al. (2024); Alayrac et al. (2022). Yet, despite this progress, a crucial dimension of visual intelligence remains largely unexplored: the perception of subjective, human-centered cognitive properties Goetschalckx et al. (2019); Isola et al. (2014); Valenzise et al. (2022). These properties—encompassing aesthetics, humor, emotion, and memorability—are fundamental to the human visual experience but pose a challenge for current models, which are primarily trained on factual, descriptive data.

Previous research has explored individual cognitive image properties, developing models to predict the memorability of an image Han et al. (2023); Khosla et al. (2015); Isola et al. (2014), quantify its aesthetic appeal LAION-AI (2024); Liu et al. (2025); Ren et al. (2017), or even assess its potential for humor Li et al. (2023b); Jain et al. (2025); Hessel et al. (2023). Although valuable, these efforts often focus on specialized, vision-only models or statistical analysis, leaving a significant gap: **there is no systematic framework to evaluate and enhance these comprehensive cognitive abilities within powerful, general-purpose MLLMs**. It remains unclear how well MLLMs align with human judgments on aesthetics, funniness, emotion, and memorability; whether they can be taught to acquire this sophisticated understanding; and what the downstream effects would be once they are more aligned with humans.

To bridge this gap, we introduce the **CogIP-Bench**, the first comprehensive benchmark designed to systematically measure the cognitive alignment between MLLMs and humans in four key dimensions. Our extensive evaluation of leading MLLMs in CogIP-Bench reveals a significant discrepancy, confirming that these models, despite their advanced capabilities, are poorly aligned with human perception of these subjective traits. For example, most open-source MLLMs showed nearly 0 correlation with humans on the memorability level of images, and the average Spearman correlation in funniness remains below 0.5. (Table 1)

Beyond merely identifying this problem and benchmarking, we demonstrate a practical solution. We show that a targeted post-training pipeline with training data in CogIP-Bench can effectively instill this cognitive knowledge into an MLLM, significantly improving its ability to predict human cognitive scores. One more compelling finding is that this newly acquired cognitive alignment is not just a passive predictive skill, but a transferable capability. By integrating our fine-tuned MLLM as the MLLM backbone in a state-of-the-art image generation pipeline (Qwen-Image) Wu et al. (2025), we show that it can guide the synthesis process to create images that are intentionally more memorable, aesthetically pleasing, or emotionally resonant.

Our main contributions can be summarized as follows:

- **Benchmark:** We introduce CogIP-Bench for evaluating MLLMs on four cognitive image properties: aesthetics, funniness, emotional valence, and memorability.
- **Alignment:** We present a post-training pipeline and dataset that improve MLLM alignment with human cognitive judgments.
- **Transferability:** We show that improved cognitive alignment transfers to more controllable, human-centric downstream generation, such as text-to-image synthesis.

Ultimately, our work takes a critical step towards developing MLLMs that not only see the world but also understand it with human-centric nuance, paving the way for more human aligned AI systems.

2 RELATED WORK

2.1 MLLMS AND THEIR EVALUATIONS

Multimodal large language models (MLLM), Wang et al. (2024); Qwen et al. (2025); Yang et al. (2025); Team & Anil (2025); Li et al. (2024a); OpenAI et al. (2024); Liu et al. (2023; 2024a); Tong et al. (2024a); Yin et al. (2024); Caffagni et al. (2024); Zhang et al. (2024; 2025); Chen et al. (2024) combine visual encoders with pretrained language models, allowing them to reason jointly over text

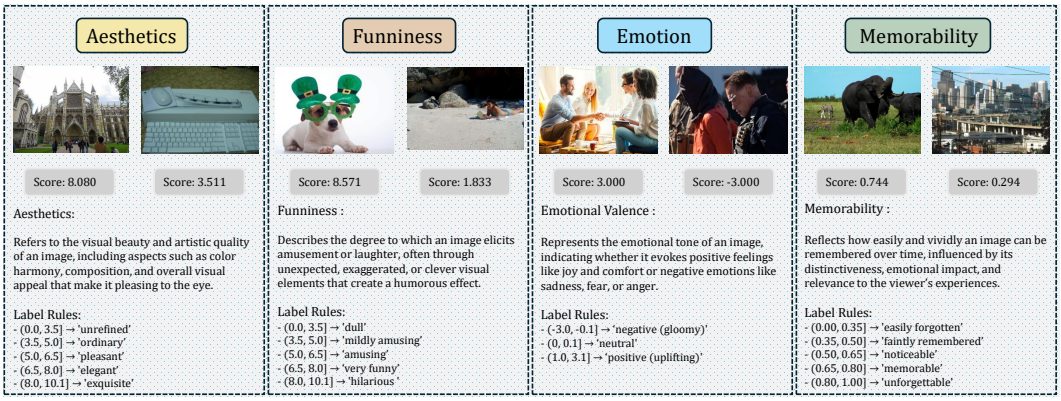


Figure 2: Examples of the CogIP-Bench, for each cognition dimension, we show two images along with their cognition scores and the interpretation of that cognitive dimension.

and images. These systems often use visual encoders such as CLIP Radford et al. (2021), along with adapter mechanisms—such as MLPs, query-based transformers, or attention layers—to bridge the two modalities. They demonstrate promising application values in a broad fields Bai et al. (2024); Chu et al. (2025); Tong et al. (2024b); Wang et al. (2025a); Xu et al. (2025); Zhai et al. (2024); Zhou et al. (2024); Driess et al. (2023) such as CAD drawing, robotics agentic system.

MLLM Benchmarks. With the rapid advancement of MLLM, all-round benchmarks have been proposed to assess diverse capabilities of MLLMs Liu et al. (2024b); Yue et al. (2024); Fu et al. (2023); Li et al. (2023a); Fu et al. (2024); Lu et al. (2024); Mathew et al. (2021); Kembhavi et al. (2016); Bai et al. (2025); Li et al. (2024b). They tested a wide range of capabilities that require both visual and textual modalities as input, ranging from text-visual grounding, perception, optical character recognition and compositional reasoning. However, there lacks a benchmark that assess the cognitive ability and the degree of alignment, in subjective properties like aesthetics, funniness, emotional valence and memorability, between MLLMs and human. Hence, our CogIP-Bench provides a pipeline for assessing how similar MLLMs perceive an image with humans from a cognitive perspective.

2.2 Human Cognition in AI

Research on the cognitive abilities of AI models has explored both the alignment of decision-making between humans and models Xie et al. (2024); Binz et al. (2025); Gui & Toubia (2023) and the interpretation of mechanisms of neural network models from neuroscientific and psychological perspectives Song et al. (2023); Ji-An et al. (2025); Singh et al. (2023). Before LLMs, some studies also demonstrated the abilities of AI models in cognitive properties of images; for example, research Han et al. (2023); Isola et al. (2014) focus on the memorability of images trying many different AI models, from linear classifiers to ViTs in order to quantify and understand it. GANalyze Goetschalckx et al. (2019) use Generative Adversarial Networks (GANs) to study the aesthetics, memorability, and emotional valence of images.

However, most existing studies primarily focus on text-only large language models Binz et al. (2025); Xie et al. (2024), vision-only models Song et al. (2023), or simpler recurrent neural networks Ji-An et al. (2025). Recent work has begun to assess the visual cognition abilities of MLLMs Buschoff et al. (2024), though such analysis are largely behavioral—centering on spatial reasoning and causal inference tasks. In contrast, the cognition abilities examined by our CogIP-Bench emphasize the subjective experience elicited by visual stimuli, assessing the alignment between MLLMs and human in terms of aesthetics, funniness, emotion and memorability properties of images. While prior studies have examined individual cognitive properties, some of which adopt statistical approaches Jang et al. (2022), others concentrating on vision-only models Li et al. (2023b); Jain et al. (2025); Jing et al. (2025); Han et al. (2023), and still others evaluating the cognitive abilities of MLLMs Mertens et al. (2024); Liu et al. (2025), these efforts typically address only a single dimension of cognition rather than providing a comprehensive perspective. Our work is the first systematic attempt to construct a benchmark that focuses on comprehensive subjective cognition (covering 4 dimensions) on images, offering new insights into MLLM cognitive evaluation.

3 COGIP-BENCH

Our benchmark is designed to test the cognition alignment of MLLM with human, in terms of score prediction on image. So each data-point contains: (1) an image for perceiving, (2) a prompt to provide specific cognition context and instruct the model to predict cognition score and (3) a human preference score as ground truth.

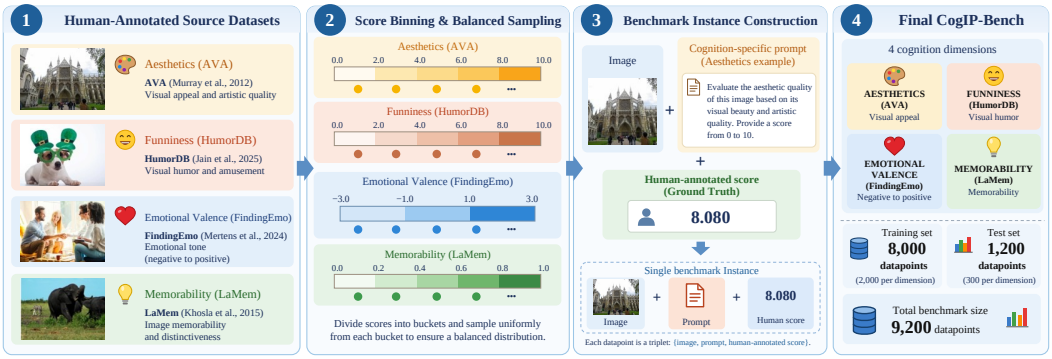


Figure 3: **Pipeline for CogIP-Bench data collection and construction.** We first identify four source datasets containing images with human preference annotations corresponding to four cognitive dimensions. To obtain a balanced score distribution, we partition emotional valence into three score buckets and each of the other three dimensions into five buckets, and sample examples across the resulting buckets. We then construct cognition-specific prompts to explicitly query the corresponding property. Each benchmark instance is represented as a triplet {image, cognition-specific prompt, human preference score}. The resulting CogIP-Bench contains 9,200 instances across aesthetics, funniness, emotional valence, and memorability.

3.1 BENCHMARK CONSTRUCTION

Our benchmark contains 4 dimensions: Aesthetics, Funniness, Emotional Valence and Memorability. The collection pipeline is listed as follows and also shown on figure 3:

Aesthetics: In the context of images, aesthetics capture the perceptual and affective qualities that evoke visual appeal, harmony, and artistic value in human observers. For this dimension, we draw image data from Aesthetic Visual Analysis (AVA) dataset which include human-annotated scores et al. (2012).

Funniness: Humor is widely accepted by "incongruity theory", which suggests that it arises from unexpected subversions of contextual expectations Attardo (2020). However, we are doing an unimodal analysis (only assess image without caption) which is inherently a "stateless" process, that is, no presummed expectation of the context, hence we call this cognition dimension funniness. We use the HumorDB dataset Jain et al. (2025), which is a carefully curated dataset designed to evaluate and advance visual humor understanding by AI systems. It comprises a diverse range of images from photos, cartoons, sketches, and AI-generated content all annotated with a humorous score, which is collected by crowdsourcing in three annotation tasks: binary classification, range rating and pairwise comparison.

Emotional Valence: We use the FindingEmo dataset Mertens et al. (2024), which is a dataset focusing on emotional understanding of MLLM and containing diverse image-text pairs annotated with human judgements of emotional valence and intensity via online crowdsourcing, where emotional valence is a bi-directional metric that emphasizes on the positive or negative quality of an emotion rather than the arousal level of it. It aims to test a model's ability to perceive, interpret, and align visual cues with emotional semantics.

Memorability: People tend to memorize images with central objects, faces or salient actions better than natural landscapes which suggests that there are certain visual intrinsics or pattern that make some images more memorable than others Han et al. (2023); Khosla et al. (2015) despite personal

experiences. To investigate this property and alignment with human perception in MLLM, we used the LaMem dataset along with their memorability scores provided Khosla et al. (2015) as the source dataset for the memorability dimension to construct our benchmark. The LaMem dataset is a large-scale memorability dataset that contains 60,000 images from diverse sources and the scores are derived from reliable and large-scale human experiments with a carefully designed visual memory game.

3.2 Benchmark Details

For each cognition property, we collect 2000 datapoints for training and 300 for test. Each datapoint is composed of an image, a cognition-specific prompt and the ground truth answer which includes the numeric score. The prompt is carefully curated to help the MLLM understand the meaning of each particular cognitive property. More details are provided in the *Appendix G*. Some examples are shown in Figure 2.

4 Benchmarking Results

We evaluate the effectiveness of our CogIP-Bench in aligning subjective cognition on images between humans and MLLMs. In Section 4.1, we outlined our experimental settings. Section 4.2 presents the results from multiple models on our CogIP-Bench and Section 4.3 introduces our post-training pipeline for injecting cognitive image knowledge into MLLMs. In Section 4.4, we show the effect of post-training and the impact on other benchmarks.

4.1 Experimental Settings

We test open-source as well as API-based MLLMs on our benchmark to evaluate their alignment with human cognition. These open-source models include: Qwen2-VL-7B Wang et al. (2024), Qwen2.5-VL-7B Qwen et al. (2025), Qwen3-VL-8B Yang et al. (2025), Gemma3-12B-it Team & Anil (2025), llava-v1.6-vicuna-13b-hf Liu et al. (2024a), llava-onevision-qwen2-7b-ov-hf Li et al. (2024a), and Llama-3.2-11B-Vision-Instruct Grattafiori & et al (2024). The API-based models are: GPT-4o OpenAI et al. (2024), GPT-5 OpenAI (2025), Claude-Haiku-4.5 Anthropic (2025) and Gemini-2.5-Pro Team & Anil (2025).

We also report the results of cognition-specific models for each cognition dimension for comparisons:
Aesthetics: We used LAION aesthetic predictor LAION-AI (2024) to get the score.
Funniness: We trained a simple score-prediction regressor model with a CLIP vision encoder and a linear projector as the output layer. Using our CogIP bench training set.
Emotional Valence: We use the same CLIP-based regression architecture as for funniness.
Memorability: We used the off-the-shelf HumanMem predictor model Han et al. (2023).

4.2 Benchmark Results on MLLMs

We tested the cognition alignment capability of MLLMs using our own CogIP-Bench (Table 1). We observed that all models fail to capture the memorability of images with near-zero Spearman correlation with human preferences except GPT-5 which is slightly better than others (correlation of 0.319). We attribute this to the abstract nature of memorability training data for MLLMs and the fundamental mismatch between human and machine cues: humans tend to remember images with simple, salient content, whereas neural networks may favor images with greater visual complexity. Han et al. (2023).

For open-source models, alignment with human judgments appears stronger in aesthetics and emotional valence, whereas alignment in the funniness dimension shows greater variability. A similar pattern is observed for API-based models, which also achieve better alignment in aesthetics and emotional valence than in funniness. This suggests that funniness may be a more challenging cognitive attribute for MLLMs to capture compared to aesthetic and emotional perception.

Moreover, we observe that MLLMs demonstrating stronger performance on other benchmarks, such as the Qwen series and Gemma3, also perform relatively well on our CogIP-Bench. Among API-based models, GPT-5 achieves comparatively better performance than the others. Overall, API-

based models exhibit stronger alignment with human judgments than open-source models. These findings suggest a positive correlation between subjective human cognitive properties and general vision–language capabilities.

Model	Aesthetics		Funniness		Emotional		Memorability		SFT Gains
	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	
Specific Model	0.819	0.800	1.206	0.569	1.732	0.720	0.987	0.790	–
Open-source Models									
Qwen2-VL-7B-Instruct	1.932	0.617	2.474	0.208	2.057	0.687	1.686	0.126	–
Qwen3-VL-8B-Instruct	2.496	0.646	2.463	0.481	2.245	0.699	1.764	0.163	–
llava-v1.6-vicuna-13b-hf	2.136	0.151	2.437	0.244	2.287	0.531	2.344	0.073	–
llava-onevision-qwen2-7b-ov-hf	2.114	0.505	2.068	0.030	2.472	0.633	1.688	0.128	–
Base Models and SFT Variants									
Qwen2.5-VL-7B-Instruct	2.143	0.569	2.265	0.435	2.171	0.655	1.793	0.108	–
↪ + SFT	1.478 ↓	0.680 ↑	1.871 ↓	0.418 ↓	2.069 ↓	0.655 =	1.652 ↓	0.077 ↓	**5/8**
Gemma3-12B-it	2.088	0.669	1.690	0.516	1.956	0.726	1.573	0.094	–
↪ + SFT	1.341 ↓	0.701 ↑	1.345 ↓	0.521 ↑	1.848 ↓	0.719 ↓	1.895 ↑	0.123 ↑	**6/8**
Llama-3.2-11B-Vision-Instruct	2.340	0.363	2.449	0.159	2.500	0.510	2.057	0.033	–
↪ + SFT	1.642 ↓	0.438 ↑	3.209 ↑	0.268 ↑	2.282 ↓	0.520 ↑	5.949 ↑	-0.060 ↓	**5/8**
API-based Models									
GPT-4o	1.730	0.758	2.178	0.456	1.907	0.721	1.673	0.168	–
GPT-5	2.020	0.838	1.440	0.591	1.633	0.759	1.463	0.319	–
Claude-Haiku-4.5	2.184	0.698	1.653	0.433	2.253	0.674	1.609	0.141	–
Gemini-2.5-Pro	1.858	0.798	2.197	0.507	1.889	0.744	1.691	0.138	–

Table 1: Performance comparison on **CogIP-Bench** across four cognitive dimensions: Aesthetics, Funniness, Emotional, and Memorability. We report MAE (↓) (normalize to the same scale when computing MAE) to measure absolute prediction error and Spearman correlation ρ (↑) to measure rank consistency with human cognition scores. The dimension-specific vision models are included as reference models. For the three open-source backbones further aligned using supervised fine-tuning (SFT), the corresponding SFT variant is shown directly below its base model. Blue arrows indicate improvements relative to the corresponding base model, brown arrows indicate degradation, and "=" denotes unchanged performance. **SFT Gains** reports the number of improved metrics out of the eight MAE/Spearman comparisons across the four cognitive dimensions. Within the open-source and API-based model groups, green highlights the best result and red highlights the worst result for each metric.

4.3 Post-training Method

We employ standard supervised fine-tuning (SFT) to inject cognitive knowledge into the model. A key challenge in this numerical score prediction task lies in the inherent insensitivity of language models to numerical values. Since numbers are tokenized into discrete tokens in the same manner as words, the model lacks an explicit understanding of the ordinal or distance relationships between them. To address this limitation, we explore two complementary approaches:

First, the prompts are designed to instruct the model to predict a categorical label as a classification task. For each cognition dimension except emotional valence, the labels are: [very low, low, medium, high, very high], while for emotional valence the labels are: [negative, neutral, positive]. Based on the provided label-to-score mapping rule, the model predicts a corresponding numeric score. This two-step formulation enables the model to first provide a qualitative judgment, leveraging its strength in natural language reasoning, and subsequently produce a quantitative score grounded in that classification.

Next, we experimented with the soft-label loss Wang et al. (2025b). As mentioned above, the standard cross-entropy loss (L_{CE}) uses one-hot encoding, which means that $\mathbf{q}$ is a Dirac distribution ($q(i) = 1$ only when $i = t$), where t is the index of the target token (1), V is the vocabulary size and p is the probability distribution of the vocabulary ($\mathbf{p} \in \mathbb{R}^V$ such that $\sum_{i=1}^{V} p_i = 1.$). Hence, the numerical distance information between digit tokens is lost when computing the loss. For example, the numbers 1 and 9 would incur the same loss when the target token is 2, even though 1 is much closer to 2.

$$L_{\mathrm{CE}}(\mathbf{p}, \mathbf{q}(t)) = \sum_{i=1}^{V} q_i(t)\,(-\log p_i), \tag{1}$$

To address this, we switch to a soft-label distribution ($\mathbf{q}^{SL}$) for numerical tokens. Instead of a Dirac function, we construct $\mathbf{q}$ (only for numeric target tokens) as shown in (2), where δ is the old Dirac distribution and η is a mixture coefficient balancing the degree of "softness" of the new loss function. This encourages the model to output digit token that are "closer" to the target value in a continuous manner. We use $\psi(t)$ as a triangular function, as suggested by the original paper Wang et al. (2025b).

$$\mathbf{q}^{SL}(t) = (1-\eta)\,\delta(t) + \eta\,\psi(t) \tag{2}$$

We use LoRA training instead of full fine-tuning to prevent overfitting. More explorations on using reinforcement learning (RL) for post-training are in the *Appendices*, Sec B.

4.4 Post-training Results

In this section, we analyze the effect of SFT on our CogIP-Bench, comparing the alignment before and after SFT for three models: Qwen2.5-VL-7B, Gemma3-12B-it, and Llama-3.2-11B, as shown in Table 1. We then examine how this alignment influences other aspects of each model's capability by evaluating them across a diverse set of benchmarks. Following the common task categorization used in Cambrian-1, Web-SSL, and LSBS Tong et al. (2024a); Fan et al. (2025); Han et al. (2025), we group the tasks into four categories: Vision-Centric, OCR, General, and Knowledge (Table 2), each comprising two benchmarks except knowledge category Tong et al. (2024c); xAI (2024); Liu et al. (2024c); Mathew et al. (2021); Fu et al. (2023); Li et al. (2023a); Lu et al. (2024). More details are provided in the *Appendices*, Table 8.

From Table 1, SFT part, we observe that, aside from memorability-where all models exhibit difficulty in prediction—there is a significant alignment improvement across most cognitive dimensions and metrics (14 out of 18 metrics show improvements). This suggests that SFT effectively incorporates human subjective preferences in image perception into MLLMs except memorability, thereby enhancing their alignment with human judgments.

Regarding performance on other benchmarks (Table 2), we find that Qwen2.5 maintain comparable performance with acceptable variance, while Gemma3 and Llama3.2 demonstrates significant improvements across most categories which further indicate a positive correlation between subjective cognitive alignment and overall model capability. We demonstrate our ablation study in *Appendices*, Sec A.

Model	Vision-Centric	OCR	General	Knowledge
Qwen2.5-VL-7B-Base	0.733	0.917	0.811	0.681
Qwen2.5-VL-7B-SFT	0.733	0.916	0.808	0.681
Gemma3-12B-it-Base	0.663	0.712	0.729	0.551
Gemma3-12B-it-SFT	0.664	0.782	0.714	0.597
Llama-3.2-11B-VI-Base	0.674	0.790	0.718	0.482
Llama-3.2-11B-VI-SFT	0.676	0.813	0.709	0.489

Table 2: Benchmark comparison across multiple models with per-section averages (Vision-Centric, OCR, General, Knowledge) with blue highlight indicating improvement.

5 Image Generation Results

In this section, we demonstrate how cognition-aligned models differ from their base counterparts through an image generation pipeline using a consistent random seed for fair comparison. Specifically, we employ Qwen-Image Wu et al. (2025), a recently released image generation framework that adopts Qwen2.5-VL-7B-Instruct as its MLLM backbone for semantic encoding. Since this model also

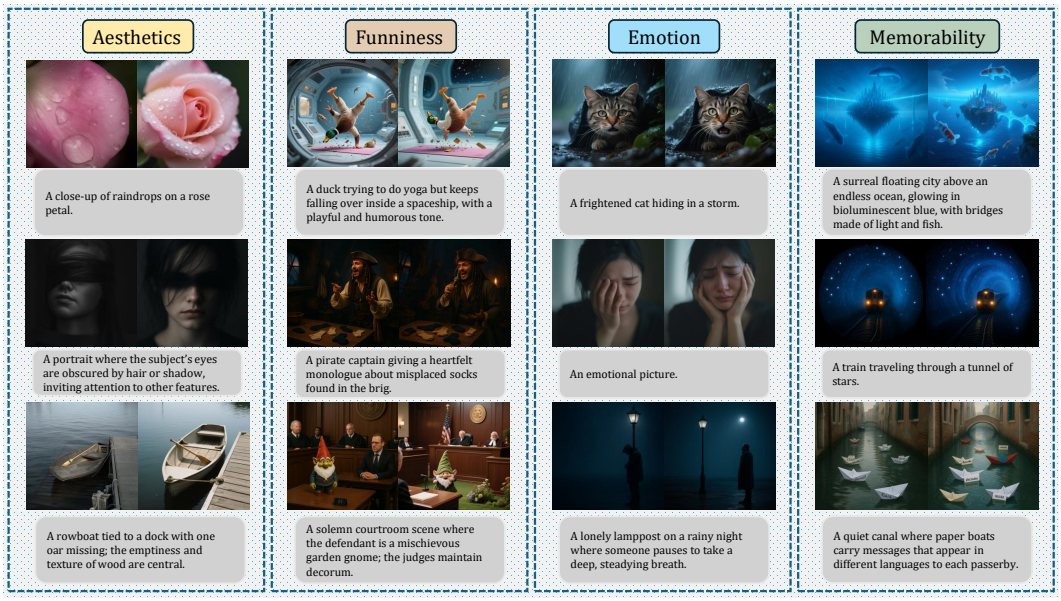


Figure 4: Qualitative comparison of images generated by the Qwen-Image pipeline using different LLM backbones (with the same prompt). The figure shows the effect of base versus supervised fine-tuning (SFT) on image cognition properties. For each image pair, Left: Base model; right: SFT model. Generation prompts are shown under each image pair. We can see that images generated with our SFT MLLM backbone better demonstrate the cognitive cues embedded in the prompts.

serves as the backbone for one of our SFT variants, we can seamlessly integrate it to compare image generations before and after fine-tuning. Uncurated visual examples are shown in Figure 4, highlighting the differences produced by the two LLM backbones.

5.1 Prompt Preparation

We split the prompts for image generation into five sets: **Aesthetics, Funniness, Emotional, Memorability** and **General**. For each set, we employed GPT-5 to generate **100** text-to-image (T2I) prompts that demonstrate the properties of that type.

5.2 Image Metrics Results

For the **General-Prompting** images, we use some general human preferences and the text-image alignment metrics to analyze the impact of the cognition alignment by SFT. We used ClipScore Hessel et al. (2022), Human Preference Score (HPS-v2) Wu et al. (2023), LAION score LAION-AI (2024), and ImageReward score Xu et al. (2023). This is illustrated in Table 3. Among these, ImageReward is a human-trained reward model designed to assess prompt-to-image consistency and aesthetic preference. As shown in Table 3, all evaluation metrics improve after replacing the MLLM backbone with its SFT-aligned counterpart. Notably, ImageReward shows a substantial gain of +0.200 points, indicating that cognition alignment significantly enhances generation quality and strengthens alignment with human visual and aesthetic preferences.

For the **Cognition-Prompting** image generation, we use cognition-specific models to evaluate the cognition scores of each image set (Table 4). Models specifications can be found in 4.1, these cognition-specific models outperform most MLLMs across all cognitive dimensions (Table 1).

We observe consistent improvements in cognition scores predicted by the post-trained models across all cognitive dimensions, confirming the effectiveness of post-training in enhancing MLLMs' alignment with human cognition on visual inputs. Notably, the emotional cognition dimension exhibits the most significant gain (+0.270 points), suggesting that Qwen-Image has a stronger ability to perceive and interpret emotional cues present in the prompts.

MLLM Backbone	ClipScore	HPS	PickScore	LAION	ImageReward
Qwen2.5-VL-7B (base)	0.330	0.267	21.973	6.765	0.878
Qwen2.5-VL-7B (ours)	0.332	0.277	22.134	6.796	1.078

Table 3: Image scores (General) for Qwen-Image–generated images with different MLLM backbones across various image-quality and human preference metrics.

MLLM Backbone	Aesthetics	Funniness	Emotional	Memorability
Qwen2.5-VL-7B (base)	6.364	3.627	1.420	7.760
Qwen2.5-VL-7B (ours)	6.462	3.705	1.690	7.770

Table 4: Image scores (Cognition) predicted by specific cognition-predicting models for Qwen-Image–generated images with different MLLM backbones.

5.3 User Study

Here, we present a user study evaluating how our cognition alignment pipeline improves human preference towards images generated by the Qwen-Image pipeline. We sampled 30 image pairs for each cognitive dimension, one generated by the baseline MLLM backbone and the other by our aligned version. 32 volunteers were then asked to select their preferred image or choose "Hard to Tell" if undecided. As shown in Figure 5, the images from our aligned model consistently achieved higher preference rates. On average, our model was preferred 1.2 times more frequently than the baseline. Check *Appendix E* for user study specifications including statistical significance analysis.

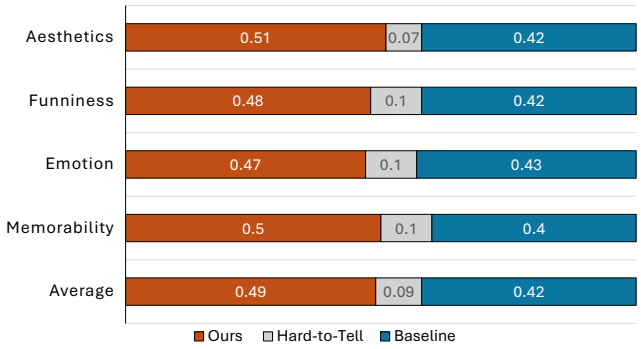


Figure 5: Preference percentages of images generated by Qwen-Image using the baseline MLLM backbone and our fine-tuned version.

6 Conclusion

Our work demonstrates that the subjective, cognitive dimension of visual understanding is no longer beyond the reach of MLLMs. By systematically measuring and then bridging the gap between machine and human perception, we have shown that these models can be taught a semblance of human-like taste and intuition. Furthermore, this learned alignment is not only an end in itself but a transferable capability that unlocks more intuitive and human-centric creative applications–a step towards more deeply human-aligned AI. We hope this work serves as a step toward creating AI systems that can not only see but also feel the world as humans do.

AI USE STATEMENT

We used generative AI tools, including `ChatGPT-5`, for the following tasks:

- Rewriting task instructions and augmenting numeric target responses with categorical qualifiers during dataset construction.
- Generating candidate text-to-image prompts for the Qwen-Image evaluation.
- Assisting with language editing, formatting, and feedback on the manuscript's clarity and internal consistency.

All AI-generated prompts and text were reviewed by the authors. The original human-provided numerical labels were preserved, and the authors independently verified the experimental results, statistical analyses, citations, and scientific claims. The authors made all final research and writing decisions and take full responsibility for the content of this paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Anthropic. Claude haiku 4.5 system card. Technical report, Anthropic, 2025. URL https://www-cdn.anthropic.com/7aad69bf12627d42234e01ee7c36305dc2f6a970/Claude%20Haiku%204.5%20System%20Card.pdf. Accessed: 2025-10-08.

Salvatore Attardo. Theories of humor and their levels. In *The Linguistics of Humor: An Introduction*, pp. 57–. Oxford University Press, 2020.

Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning, June 2024. URL http://arxiv.org/abs/2406.11896. arXiv:2406.11896 [cs].

Tianyi Bai, Yuxuan Fan, Jiantao Qiu, Fupeng Sun, Jiayi Song, Junlin Han, Zichen Liu, Conghui He, Wentao Zhang, and Binhang Yuan. Hallucination at a glance: Controlled visual edits and fine-grained multimodal learning. *arXiv preprint arXiv:2506.07227*, 2025.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, pp. 1–8, July 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09215-4. URL https://www.nature.com/articles/s41586-025-09215-4. Publisher: Nature Publishing Group.

Luca M. Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models, August 2024. URL http://arxiv.org/abs/2311.16093. arXiv:2311.16093 [cs].

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey, 2024. URL https://arxiv.org/abs/2402.12451.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. Can multimodal large language models truly perform multimodal in-context learning?, 2024. URL https://arxiv.org/abs/2311.18021.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training, May 2025. URL http://arxiv.org/abs/2501.17161. arXiv:2501.17161 [cs].

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. URL https://arxiv.org/abs/2303.03378.

Murray et al. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *CVPR*, pp. 2408–2415. IEEE, 2012. doi: 10.1109/CVPR.2012.6247954. URL https://doi.org/10.1109/CVPR.2012.6247954.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, and Saining Xie. Scaling language-free visual representation learning. *arXiv preprint arXiv:2504.01017*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2024. URL https://arxiv.org/abs/2501.00321.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties, 2019. URL https://arxiv.org/abs/1906.10112.

Aaron Grattafiori and et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

George Gui and Olivier Toubia. The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4650172. URL http://arxiv.org/abs/2312.15524. arXiv:2312.15524 [cs].

Junlin Han, Huangying Zhan, Jie Hong, Pengfei Fang, Hongdong Li, Lars Petersson, and Ian Reid. What images are more memorable to machines?, 2023. URL https://arxiv.org/abs/2211.07625.

Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying llm visual priors from language pre-training. *arXiv preprint arXiv:2509.26625*, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL https://arxiv.org/abs/2104.08718.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 688–714, 2023.

Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (7):1469–1482, July 2014. ISSN 1939-3539. doi: 10.1109/TPAMI.2013.200. URL https://ieeexplore.ieee.org/abstract/document/6629991. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Vedaant Jain, Felipe dos Santos Alves Feitosa, and Gabriel Kreiman. Humordb: Can ai understand graphical humor?, 2025. URL https://arxiv.org/abs/2406.13564.

Hyeongnam Jang, Yeejin Lee, and Jong-Seok Lee. Modeling, quantifying, and predicting subjectivity of image aesthetics, 2022. URL https://arxiv.org/abs/2208.09666.

P. Jayachandran, R. Rajendran, et al. Deep learning-based visual question answering: Survey. *AIP Conference Proceedings*, 3264(1):030011, 2024. doi: 10.1063/5.0226390. URL https://pubs.aip.org/aip/acp/article/3264/1/030011/3338475.

Li Ji-An, Marcus K. Benna, and Marcelo G. Mattar. Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, pp. 1–9, July 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09142-4. URL https://www.nature.com/articles/s41586-025-09142-4. Publisher: Nature Publishing Group.

Jie Jing, Qing Lin, Shuangpeng Han, Lucia Schiatti, Yen-Ling Kuo, and Mengmi Zhang. Unforgettable Lessons from Forgettable Images: Intra-Class Memorability Matters in Computer Vision Tasks, February 2025. URL http://arxiv.org/abs/2412.20761. arXiv:2412.20761 [cs].

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *International Conference on Computer Vision (ICCV)*, 2015.

LAION-AI. aesthetic-predictor: A linear estimator on top of clip to predict the aesthetic quality of pictures. GitHub repository, https://github.com/LAION-AI/aesthetic-predictor, 2024. MIT License; accessed 23 Oct 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer, October 2024a. URL http://arxiv.org/abs/2408.03326. arXiv:2408.03326 [cs].

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. A survey on benchmarks of multimodal large language models, 2024b. URL https://arxiv.org/abs/2408.08632.

Runjia Li, Shuyang Sun, Mohamed Elhoseiny, and Philip Torr. OxfordTVG-HIC: Can Machine Make Humorous Captions from Images?, July 2023b. URL http://arxiv.org/abs/2307.11636. arXiv:2307.11636 [cs].

Boyang Liu, Yifan Hu, Senjie Jin, Shihan Dou, Gonglei Shi, Jie Shao, Tao Gui, and Xuanjing Huang. Unlocking the Essence of Beauty: Advanced Aesthetic Reasoning with Relative-Absolute Policy Optimization, September 2025. URL http://arxiv.org/abs/2509.21871. arXiv:2509.21871 [cs] version: 1.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, December 2023. URL http://arxiv.org/abs/2304.08485. arXiv:2304.08485 [cs].

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, May 2024a. URL http://arxiv.org/abs/2310.03744. arXiv:2310.03744 [cs].

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024c. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225. URL https://arxiv.org/abs/2007.00398.

Laurent Mertens, Elahe' Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. Findingemo: An image dataset for emotion recognition in the wild, 2024. URL https://arxiv.org/abs/2402.01355.

OpenAI. Gpt-5 system card. Technical report, OpenAI, Aug 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf. Accessed: 2025-11-08.

OpenAI, Aaron Hurst, Adam Lerer, and et al Goucher. GPT-4o System Card, October 2024. URL http://arxiv.org/abs/2410.21276. arXiv:2410.21276 [cs].

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *Proceedings of the IEEE international conference on computer vision*, pp. 638–647, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. EEG2IMAGE: Image Reconstruction from EEG Brain Signals, March 2023. URL http://arxiv.org/abs/2302.10121. arXiv:2302.10121 [cs].

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding Natural Images from EEG for Object Recognition. October 2023. URL https://openreview.net/forum?id=dhLIno8FmH.

Gemini Team and et al Anil. Gemini: A Family of Highly Capable Multimodal Models, May 2025. URL http://arxiv.org/abs/2312.11805. arXiv:2312.11805 [cs].

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, December 2024a. URL http://arxiv.org/abs/2406.16860. arXiv:2406.16860 [cs].

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. MetaMorph: Multimodal Understanding and Generation via Instruction Tuning, December 2024b. URL http://arxiv.org/abs/2412.14164. arXiv:2412.14164 [cs].

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024c. URL https://arxiv.org/abs/2401.06209.

Giuseppe Valenzise, Chen Kang, and Frédéric Dufaux. *Advances and Challenges in Computational Image Aesthetics*, pp. 133–181. Springer International Publishing, Cham, 2022. ISBN 978-3-030-81465-6. doi: 10.1007/978-3-030-81465-6_6. URL https://doi.org/10.1007/978-3-030-81465-6_6.

Chenyu Wang, Weixin Luo, Sixun Dong, Xiaohua Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. MLLM-Tool: A Multimodal Large Language Model For Tool Agent Learning, April 2025a. URL http://arxiv.org/abs/2401.10727. arXiv:2401.10727 [cs].

Pei Wang, Zhaowei Cai, Hao Yang, Davide Modolo, and Ashwin Swaminathan. Enhancing numerical prediction of mllms with soft labeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2025*, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

xAI. Realworldqa: A benchmark for real-world visual question answering. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-10-28.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can Large Language Model Agents Simulate Human Trust Behavior?, November 2024. URL http://arxiv.org/abs/2402.04559. arXiv:2402.04559 [cs].

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

Jingwei Xu, Chenyu Wang, Zibo Zhao, Wen Liu, Yi Ma, and Shenghua Gao. CAD-MLLM: Unifying Multimodality-Conditioned CAD Generation With MLLM, August 2025. URL http://arxiv.org/abs/2411.04954. arXiv:2411.04954 [cs].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL http://dx.doi.org/10.1093/nsr/nwae403.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning, October 2024. URL http://arxiv.org/abs/2405.10292. arXiv:2405.10292 [cs].

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models, 2024. URL https://arxiv.org/abs/2401.13601.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models, 2025. URL https://arxiv.org/abs/2411.18620.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model, August 2024. URL http://arxiv.org/abs/2408.11039. arXiv:2408.11039 [cs].

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13041–13049, 2020. URL https://arxiv.org/abs/1909.11059.

APPENDICES

A ABLATION STUDY

To examine the contribution of each component in our alignment strategy, we conduct an ablation study in four settings: (1) fine-tuning without the soft-label loss, (2) using a simplified prompting scheme that omits the "describe-then-predict" instruction format and directly asks the model to predict the score, (3) fine-tuning only the visual encoder while freezing the language model (LLM) and (4) fine-tuning only the LLM with the visual encoder frozen. We conduct this ablation study using Qwen2.5-VL-7B. We also include the average regression percentages of these fine-tuned models across four benchmarks: MMVP Tong et al. (2024c), OCRBench Liu et al. (2024c), MME Fu et al. (2023) and MathVista Lu et al. (2024) (each representing one task category Tong et al. (2024a)). Comparing the freeze-vision and freeze-LLM configurations (Table 5), we find that the freeze-llm setting yields the weakest alignment across all cognitive dimensions except memorability, whereas the freeze-vision setting achieves slightly better performance, particularly on the aesthetics and emotion dimensions and it also achieve less regression on other benchmarks. This observation agrees with common fine-tuning practices in downstream multimodal tasks, where freezing the visual encoder is often preferred for stability and efficiency.

Model	Aesthetics	Funniness	Emotional	Memorability	Regression
SFT	0.680	0.418	0.655	0.08	-0.25%
Freeze-Vision	0.659	0.417	0.605	0.097	-1.0%
Freeze-LLM	0.607	0.416	0.577	0.362	-7.5%

Table 5: Ablation results (Spearman) across four cognitive dimensions, comparing three frozen-encoder configurations of Qwen2.5-VL-7B. The SFT setting is our baseline alignment method. Regression represents the average performance degradation percentage on other multimodal reasoning benchmarks.

For the ablation of prompting techniques and soft-label loss (Table 6), we observe that incorporating the soft-label loss generally enhances alignment across half of cognitive dimensions, demonstrating its effectiveness in numeric prediction tasks. Finally, the proposed prompting strategy yields the most substantial gains, improving cognitive alignment notably across all dimensions except memorability.

Model	Aesthetics	Funniness	Emotional	Memorability	Regression
SFT	0.680	0.418	0.655	0.08	-0.25%
Simple-Prompt	0.550	0.333	0.652	0.140	-0.25%
No Soft-Label	0.675	0.333	0.665	0.094	-0%

Table 6: Ablation results (Spearman) demonstrating the effect of our prompting techniques and soft-label loss on our baseline alignment method using Qwen2.5-VL-7B.

B COGNITION-ALIGNMENT WITH RL

Can RL methods further align MLLMs with human cognition? To investigate this, we apply Group Relative Policy Optimization (GRPO) Shao et al. (2024) on our benchmark, fine-tuning the MLLM with a reward based on the numerical proximity between its predicted score and the ground-truth (GT). Unlike SFT, which relies on static label supervision, GRPO provides an adaptive signal that encourages the model to iteratively calibrate its predictions toward human evaluations. We nevertheless use SFT as our primary post-training method because it provides a common, stable baseline across diverse model families and reliably teaches both score correlation and output-format discipline.

Table 7 reports complete checkpoint evaluations for GRPO on Qwen2.5-VL-7B and Gemma3-12B-it, together with the Base and SFT results from Table 1. The results show that GRPO does not show the same effectiveness for both models. For Qwen2.5-VL-7B, the GRPO result exceeding base and SFT comparament models on all metrics.

Model	Aesthetics		Funniness		Emotional		Memorability	
	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
Qwen2.5-VL-7B	2.143	0.569	2.265	0.435	2.171	0.655	1.793	0.108
↪ + SFT	1.478 ↓	0.680 ↑	1.871 ↓	0.418 ↓	2.069 ↓	0.655 =	1.652 ↓	0.077 ↓
↪ + GRPO	0.691 ↓	0.780 ↑	1.345 ↓	0.483 ↑	1.038 ↓	0.689 ↑	1.392 ↓	0.389 ↑
Gemma3-12B-it	2.088	0.669	1.690	0.516	1.956	0.726	1.573	0.094
↪ + SFT	1.341 ↓	0.701 ↑	1.345 ↓	0.521 ↑	1.848 ↓	0.719 ↓	1.895 ↑	0.123 ↑
↪ + GRPO	1.282 ↓	0.075 ↓	1.497 ↓	0.053 ↓	1.055 ↓	0.722 ↓	1.507 ↓	0.050 ↓

Table 7: Comparison of Base, SFT, and GRPO models on **CogIP-Bench** across four cognitive dimensions. We report MAE (↓) and Spearman correlation ρ (↑). For each backbone, SFT and GRPO variants are shown directly below the corresponding base model. Blue arrows indicate improvements relative to the base model, red arrows indicate degradation, and "=" denotes approximately unchanged performance.

The Gemma3-12B-it results are consistent with **prediction collapse toward central scores**. Subjective score prediction is a regression problem, and GRPO can reduce a distance-based reward by concentrating predictions near the dataset mean without preserving instance-level ordering. This explains why memorability obtains low MAE but a Spearman correlation of only 0.0504.

We also find that GRPO is highly sensitive to rollout configurations, including group size, temperature, and top-p, as well as reward scaling. Together, these complete checkpoint results do not support a uniform GRPO improvement over SFT and reinforce our use of SFT as the reliable cross-model baseline. GRPO remains a backbone- and dimension-sensitive exploration, and hybrid SFT-RL tuning is left for future work.

C Details of Results on Other Benchmarks

This section presents the detailed performance of the three MLLMs on 7 benchmarks spanning four task categories, evaluated both before and after SFT on our CogIP-Bench:

Model	Vision-Centric			OCR			General			Knowledge
	MMVP	RealWorldQA	Avg	OCRBench	DocVQA	Avg	MME	SeedBench	Avg	MathVista
Qwen2.5-Base	0.780	0.686	**0.733**	0.884	0.949	**0.917**	0.849	0.772	**0.811**	**0.681**
Qwen2.5B-SFT	0.780	0.685	**0.733**	0.881	0.951	**0.916**	0.846	0.771	**0.808**	**0.681**
Gemma3-Base	0.723	0.603	**0.663**	0.702	0.722	**0.712**	0.765	0.692	**0.729**	**0.551**
Gemma3-SFT	0.720	0.607	**0.664**	0.753	0.811	**0.782**	0.777	0.651	**0.714**	**0.597**
Llama-3.2-Base	0.717	0.630	**0.674**	0.755	0.825	**0.790**	0.705	0.731	**0.718**	**0.482**
Llama-3.2-SFT	0.717	0.635	**0.676**	0.751	0.875	**0.813**	0.683	0.734	**0.709**	**0.489**

Table 8: Benchmark comparison across multiple models with section averages (Vision-Centric, OCR, General, Knowledge). Green entry color indicates an improvement on the average benchmark result on a certain task category of a particular MLLM, while the red entry color indicates a regression.

D TRAINING DETAILS

In this section, we report all hyperparameters used in our experiments—including the soft-label parameters—for training the three model variants, as summarized in Table 9.

Hyperparameter	**Qwen2.5**	**Gemma3**	**Llama3.2**
LoRA Rank	64	64	64
LoRA α	64	64	64
LoRA Dropout	0.05	0.05	0.05
LoRA Target	all	all	all
GPU	4 $\times$ A40	4 $\times$ A40	4 $\times$ A40
Batch Size	4	4	4
Gradient Accumulation Steps	8	8	8
Number of Trained Epochs	1	1	1
Warmup Ratio	0.03	0.03	0.03
Weight Decay	0.1	0.1	0.1
Learning Rate	1e-5	7e-6	3e-5
Vision Learning Rate	2e-6	2e-6	1e-6
Projector Learning Rate	1e-5	1e-5	1e-5
Learning Rate Scheduler	Cosine	Cosine	Cosine
Freeze LLM	True	True	True
Unfreeze Vision Tower	True	True	True
Soft Label			
η	0.15	0.15	0.10
λ	2.0	2.0	2.0

Table 9: Hyperparameters specification for training Qwen2.5-VL-7B, Gemma3-12B-it, and Llama-3.2-11B-VI models.

E User Study Specifications

For this user study analysis, we recruited 32 volunteers from diverse backgrounds including university students and researchers to evaluate 120 generated image pairs (30 pairs per dimension). For each pair, participants compared generations from the baseline Qwen model and our aligned model side-by-side (using identical generation seeds). Image positions (left vs. right) and the overall sequence order were fully randomized for each participant. In total, this study gathered 3,840 pairwise comparisons, providing a robust sample size for human preference evaluation.

E.1 User Study Examples

This section presents example image pairs used in the user study where 32 volunteers were asked to select the preferred image-without knowing which model produced it—or choose "Hard to Tell" when undecided.



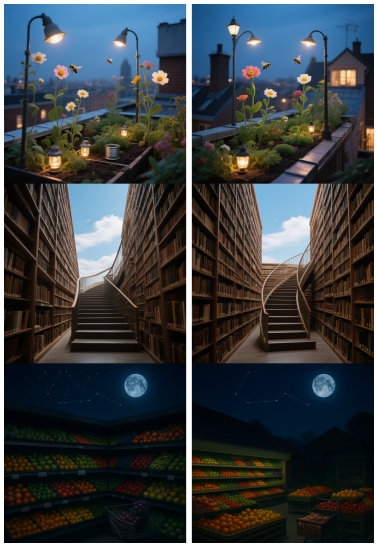
(a) Aesthetics



(b) Funniness



(c) Emotional Valence



(d) Memorability

Figure 6: Demonstration of the user study set up, each pair is generated with the same prompt using Qwen-Image using consistent seed.

E.2 Statistical Significance Analysis

We conducted a statistical analysis to assess the significance of the user study results. We defined the **preference rate** for each participant i as the proportion of decisive comparisons in which images generated by our model were preferred over those generated by the baseline. "Hard-to-Tell" responses were excluded from this metric:

$$p_i = \frac{N_{\text{Ours},i}}{N_{\text{Ours},i} + N_{\text{Baseline},i}}$$

For each of the four cognitive dimensions, we first averaged the participant-level preference rates across all 32 participants to obtain the overall preference rate. To quantify the uncertainty and statistical significance of these estimates, we then performed participant-level bootstrap resampling with 20,000 iterations to estimate the 95% confidence intervals (CIs). We additionally computed one-sided bootstrap p-values against the null hypothesis that our model was not preferred over the baseline, i.e., that the mean preference rate was less than or equal to chance level ($\bar{p} \leq 0.5$).

Dimension	**Preference Rate**	**95% CI**	p**-value**
Aesthetics	**0.546**	[0.514, 0.579]	**0.004**
Emotion	0.532	[0.485, 0.578]	0.084
Funniness	0.521	[0.472, 0.569]	0.181
Memorability	**0.553**	[0.505, 0.601]	**0.019**

Table 10: Decisive preference rates and bootstrap significance analysis across the four cognitive dimensions. "Hard-to-Tell" responses were excluded when computing participant-level preference rates.

Interpretation of Statistical Significance. Aesthetics ($p = 0.004$) and Memorability ($p = 0.019$) showed statistically significant preference rates above the chance level of 0.5. Their 95% bootstrap confidence intervals also remained entirely above 0.5, providing consistent evidence that participants preferred images generated by our model over those generated by the baseline along these two dimensions.

Emotion ($p = 0.084$) and Funniness ($p = 0.181$) exhibited positive preference margins, with mean preference rates of 0.532 and 0.521, respectively. However, these effects did not reach statistical significance at the conventional threshold of $p < 0.05$, and their confidence intervals included 0.5. One possible explanation is that emotional and humorous content admits a wider range of valid human interpretations, resulting in greater variability in image preferences and consequently noisier preference signals.

Overall, all four dimensions yielded mean preference rates above 0.5, indicating a consistent directional trend in favor of our model. However, statistically significant evidence was observed only for Aesthetics and Memorability.

F EXAMPLES OF SFT DATASET

In this section, we provide example samples from our constructed CogIP-Bench dataset used for supervised fine-tuning, as shown in Figures 7 and 8. The dataset spans four cognitive dimensions.

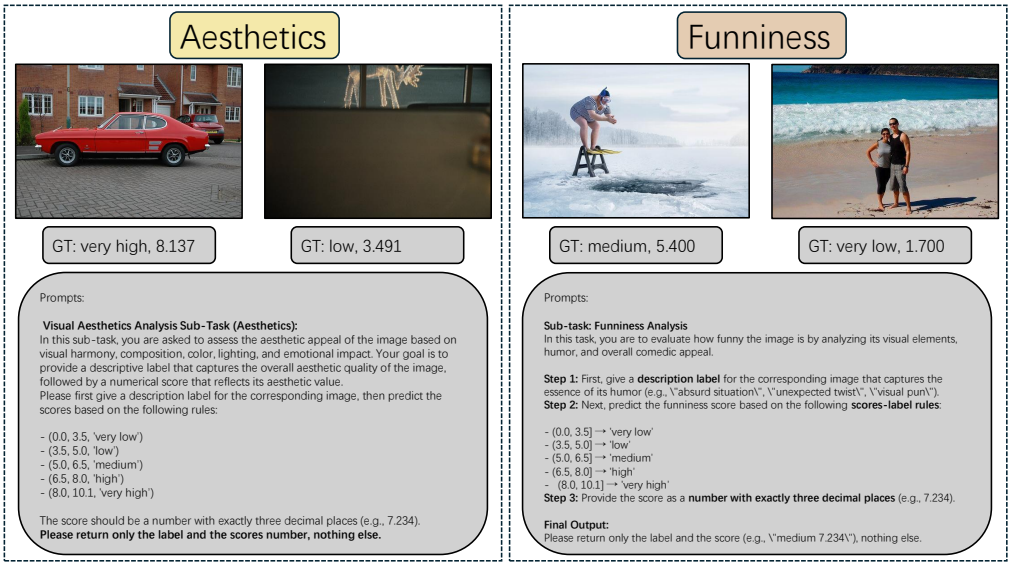


Figure 7: Examples of CogIP-Bench samples for Aesthetics and Funniness

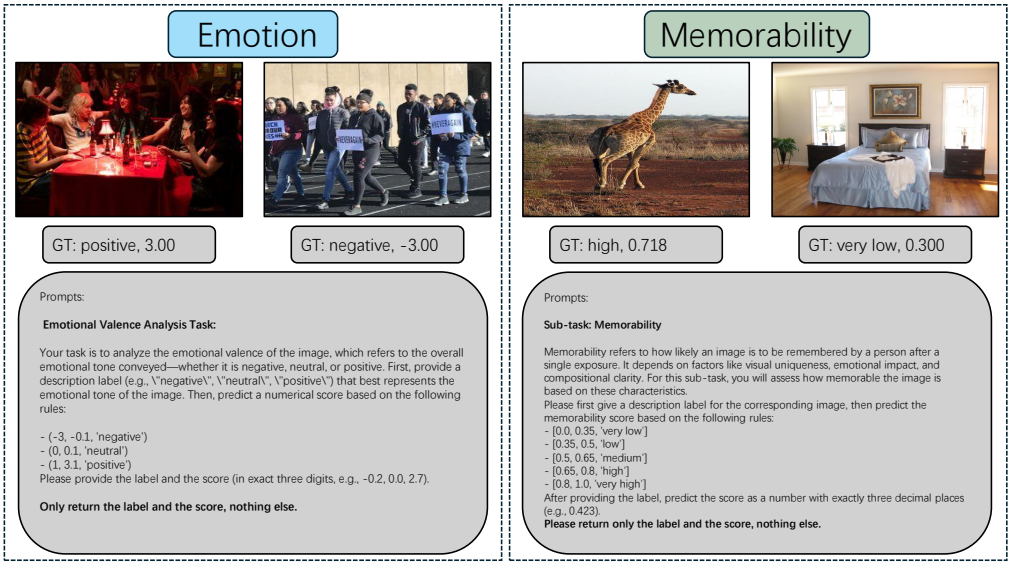


Figure 8: Examples of CogIP-bench samples of Emotional Valence and Memorability.

G PROMPTS

In this section, we systematically present all prompts used to construct the CogIP-Bench dataset and to evaluate the SFT effects on Qwen-Image. Specifically:

- Figure 9 illustrates the prompt used to rewrite the original simple instruction prompts for the dataset—shown here using aesthetics as an example—including the reformulation of the "from human" instruction.
- Figure 10 presents the prompt employed to rewrite the "from gpt" ground-truth responses.
- Figure 11 provides the prompts used to generate text-to-image instructions designed to elicit specific cognitive attributes, along with a general-purpose version for comparison.

Prompt for rewriting "from user" instruction:

You will see a visual-analysis instruction.
Rewrite it in fresh language **for the same sub-task** (Aesthetics).
• Important: Keep the **<image>** tag.
• Add brief explanation for this specific sub-task (Aesthetics) to the instruction.
• And also (in this specific order):
  • Ask the model to give a description label for the corresponding image, then
  • Ask the model to predict the scores based on this scores-label rules:
{
- (0.0, 3.5, 'very low')
- (3.5, 5.0, 'low')
- (5.0, 6.5, 'medium')
- (6.5, 8.0, 'high')
- (8.0, 10.1, 'very high')
}.
• Describe the scores prediction rule (e.g. exact three digits) after asking the model to return the label and the scores number.
• Ask the model to return only the label and the scores number, nothing else.

ORIGINAL:
<image> You are an expert in visual aesthetics. Please analyze the aesthetic value of the following image and assign a score from 1.000 to 10.000, using exactly three decimal places. The score should reflect nuanced judgment — not just whole or half values. Be precise and consider subtle details. Avoid defaulting to scores ending in .000 or .500 unless genuinely appropriate. Please only return a number without other comment

Figure 9: Prompts for rewriting "from human" instruction in the dataset.

Prompt for rewriting "from MLLM" (GT):

A dataset stores only the numeric score **{5.241}** for the sub-task {Aesthetics}.
Rewrite that answer by:
• keeping the SAME numeric value
• Augment the answer with a short natural-language qualifier according to the label rules:
{
- (0.0, 3.5, 'very low')
- (3.5, 5.0, 'low')
- (5.0, 6.5, 'medium')
- (6.5, 8.0, 'high')
- (8.0, 10.1, 'very high')
}
• Returning only ONE line with the numeric score and the qualifier, nothing else.

ORIGINAL ANSWER:
{5.241}

Figure 10: Prompts for rewriting "from gpt" GT in the dataset.

Prompt for generating Qwen-Image T2I prompts:

- **Aesthetics**: describe something that require understanding of beauty, whether it's people or nature or plants.
- **Funniness**, require humorous cognition, know what is dramatic/funny.
- **Emotional**, can capture the emotion of a certain scene/facial expression.
- **Memorability**, describe a scene which will challenge the model to generate one that is more memorable (may give some details/vague description) and make the prompt a bit vague (but not too much) so it can test whether they have good understanding of the cognition.
- **General**, give general text-to-image generation prompts that test the quality of the generated images.

Figure 11: Prompt to generate T2I generation prompts for Qwen-Image