

# SenseDesign: a Design Work Dataset Annotated with Binary Polarity Semantic Descriptors to Evaluate VLLMs’ Aesthetic Sense

Anonymous ICME submission

**Abstract**—Recent advances in Vision-Language Large Models (VLLMs) have demonstrated impressive capabilities in various multimodal tasks. However, these models often struggle with capturing subtle aesthetic nuances, particularly in the context of design evaluation. Existing benchmark datasets also could not support the evaluation of VLLMs’ capabilities in aesthetic understanding of design works. We present SenseDesign, a carefully curated dataset of 1,019 professional design images annotated by experts using an adapted Semantic Differential Method with binary polarity semantic descriptors. Based on this newly curated dataset, we propose a novel framework to evaluate VLLMs’ aesthetic sense in industrial design works by combining vision-language large models with emotional word-pair classification using two different evaluators. Through extensive experimentation across multiple VLLMs, we provide empirical evidence that guides future improvements in multimodal understanding of aesthetics in design work. Our code and the dataset would be released upon publication acceptance.

**Index Terms**—Vision-Language Large Models (VLLMs), Multimodal Understanding, Aesthetic and Emotional Assessment.

## I. INTRODUCTION

In recent years, Large Language Models (LLMs) have achieved remarkable success across a wide range of linguistic tasks, including natural language understanding, machine translation, dialogue generation, and information retrieval. By leveraging extensive textual corpora, these models have demonstrated impressive capabilities in both comprehension and generation of human language. As research progresses toward more complex, multimodal contexts, the need to jointly process visual and textual information has led to the emergence of Vision-Language Large Models (VLLMs). While VLLMs have shown promise in areas such as image captioning, visual question answering, and cross-modal retrieval, they still face significant limitations. On the one hand, they often exhibit “hallucinations”, producing descriptions misaligned with the actual visual content; on the other hand, current methods of visual feature extraction struggle to capture the subtle emotional and aesthetic nuances inherent in design works.

This study focuses on the challenging task of aesthetic evaluation of design works, a concept closely aligned with the principles of Kansei Engineering [1]—a user-centered methodology that inherently values the emotional and aesthetic aspects of product experiences. **In this context, the aim is to evaluate mature designs by distilling their embedded aesthetic and emotional qualities, and articulating these insights in precise, contextually appropriate language.** Given that VLLMs often struggle with such nuanced aesthetic

dimensions, employing aesthetic evaluation as a testing ground enables us to rigorously assess their aesthetic sense.

However, direct verification on whether VLLMs possess deep aesthetic awareness is not straightforward. Aesthetic judgments are inherently shaped by individual preferences and cultural backgrounds, and without appropriate high-level knowledge constraints, models prompted with naive inputs may produce superficial, vacuous responses lacking genuine insight. To address this challenge, we incorporate the Semantic Differential Method (SDM) to impose a structured and well-defined evaluative framework, enhancing semantic precision and discriminability in aesthetic evaluation. Moreover, given the ambiguity and overlapping meanings of certain descriptors (e.g., “warm” and “energizing”), we introduce a textual discrimination mechanism using binary polarity word pairs, which ensures methodological rigor and interpretability even when handling multifaceted, potentially overlapping concepts. These insights necessitate an annotated design work dataset with binary polarity semantic descriptors. Unfortunately, such a dataset is not yet existent to enable nuanced and comprehensive aesthetic evaluation using VLLMs.

To tackle the above challenges, we make the following major contributions in this work:

- 1) **Creating Benchmark Dataset:** We collected 9,229 design-related images and curated a final dataset consisting of 1,019 images, and called it SenseDesign. Images in SenseDesign are carefully annotated by design experts using SDM, establishing a high-quality foundation for aesthetic evaluation of design works.
- 2) **Evaluating VLLMs’ Aesthetic Sense:** We propose a framework to evaluate VLLMs’ sense of aesthetic evaluation in industrial design works.
- 3) Through comprehensive experiments using the most representative VLLMs, we systematically analyze their performance using two different evaluators, and find that existing VLLMs still have rooms to improve in aesthetic evaluation of design works.

## II. RELATED WORK

### A. VLLMs in Visual Question Answering

With the rapid development of vision-language tasks in recent years, researchers begin to explore the integration of visual inputs into language models. Through the incorporation of visual encoding techniques like CLIP [2], VLLMs have

emerged as powerful tools for multimodal understanding and generation.

As an important multimodal task, VQA [3], [4] requires a wide range of models. From basic object recognition (e.g. “How many apples are there in the picture?”), to complex visual reasoning (“What emotion does this scene convey?”), VQA tasks demonstrate a variety of levels in difficulty. Simple VQA tasks may only require the model to have basic object detection and counting capabilities, while advanced VQA tasks require the model to have deep scene understanding, contextual reasoning, and even grasp of abstract concepts.

Various types of VQA tasks have been proposed to assess the ability of VLLMs, including factual question answering [5], relational reasoning [6], and common sense reasoning [7]. These tasks test the depth of VLLMs’ visual understanding from different perspectives. However, very few are dedicated to the assessment of the aesthetic characteristics in images. Based on this observation, we propose a new paradigm in VQA task that focuses on evaluating the model’s ability to understand the aesthetic features of images. This task not only requires the model to recognize basic visual elements, but also to have a deep understanding of abstract aesthetic concepts such as expressiveness and emotion.

### B. Existing Aesthetic Benchmarks

Several notable datasets and evaluation frameworks have been proposed [8]–[12] to evaluate VLLMs’ aesthetic understanding capabilities. These efforts have made important contributions to our understanding of how VLLMs perceive and process aesthetic information.

However, existing aesthetic benchmarks face several significant limitations: **First.** The inclusion of AI-generated images (AGI) in existing datasets, such as those in [13], introduces undesired noise, as even after data cleaning, these images may contain regions that are not aesthetically well-designed. Unlike carefully crafted photographs or designs where each element serves a purpose, AGIs may contain arbitrary or inconsistent aesthetic elements that compromise the dataset’s quality. **Second.** Current datasets typically fall into three categories: aesthetic scoring [11], [12], aesthetic captions [10], and composition scoring [8]. For score-based annotations, using only numerical values severely limits the dimensionality of aesthetic information. For caption-based annotations (e.g., the AVA dataset [10]), descriptions often merely state what is visible in the image without capturing deeper aesthetic principles or design intentions. For instance, a caption might simply describe “a red charging cable” without acknowledging its user-centric design philosophy. **Third.** Existing datasets often contain distracting background information beyond the main subject matter, making it challenging for VLLMs to focus on and evaluate the intended aesthetic elements [14]. This noise in visual information can hinder the model’s ability to identify and assess the aesthetic aspects that are actually relevant to the evaluation task.

Based on the above observations, we curate the SenseDesign dataset, which contains industrial design works with manually

annotated semantic word pairs using binary polarity descriptors. These annotated design works better reflect creator’s aesthetic design intentions, and mitigate the aforementioned superficial label limitation by multi-dimensional aesthetic annotations. Using our newly curated dataset in design works, we propose a framework to evaluate existing most representative VLLMs’ ability to sense aesthetics.

## III. METHOD

In this section, we provide the details in our data curation and methodology to evaluate VLLMs’ aesthetic sense. First, we present the construction and annotation process of the **SenseDesign** dataset, which forms the basis for our experiments. Next, we outline the overall workflow in our proposed evaluation framework. Finally, we introduce the two evaluators adopted in our framework.

### A. SenseDesign Dataset

We constructed our SenseDesign dataset through systematic web crawling, utilizing search keywords related to design, aesthetic design, product design, and industrial design across various online platforms. Following the initial data collection, we conducted a rigorous manual filtering process to remove irrelevant images, including daily photographs, advertisements, and text-only images, as well as those with cluttered backgrounds or unclear design subjects. Notably, our final curated collection primarily consists of design works that have received recognition from prestigious international design competitions, including the Red Dot Award, iF Design Award, and Good Design Award, ensuring the dataset’s quality and professional credibility.

Our annotation methodology draws inspiration from the Semantic Differential Method (SDM) in Kansei Engineering [15], while specifically adapting it for evaluating VLLMs. We develop a more straightforward approach using semantic word pairs that capture design characteristics through opposing concepts. Rather than employing traditional multi-point rating scales, we adopt a binary semantic polarity approach where annotators make clear choices between opposing descriptors. This adaptation not only preserves the fundamental strength of contrasting semantic concepts, but also creates more distinct and unambiguous training data. Such binary classification aligns particularly well with how VLLMs process and understand design characteristics, potentially enabling more accurate assessment of their ability to perceive and interpret design elements.

The annotations were carried out by a diverse team of five experts, including university art and design professors, senior product designers with over ten years of experiences, and user researchers. They selected the most relevant descriptors for each design work from predefined semantic word pair repository and made decisive choices within each selected pair to best capture the design’s characteristics. This approach differs from traditional scale-based evaluations, as it generates precise binary annotations that are particularly well-suited for training and evaluating VLLMs, effectively bridging the gap

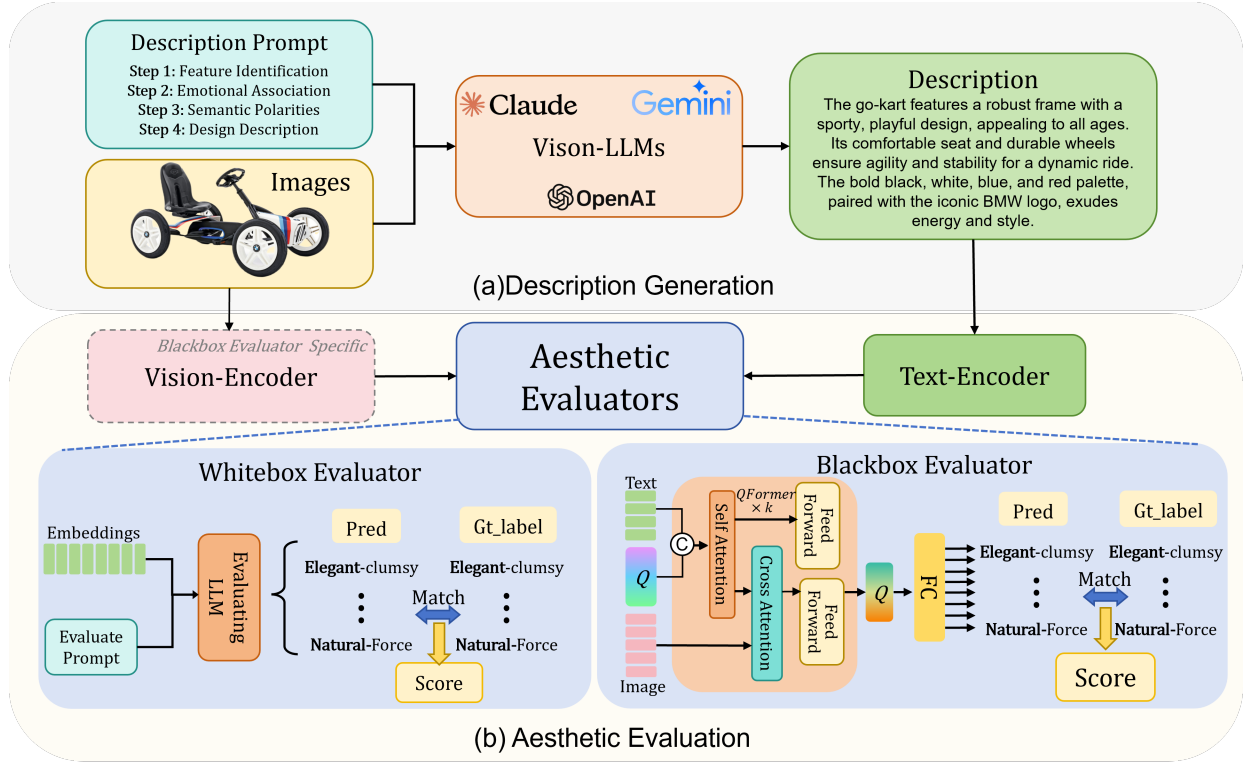


Fig. 1: The proposed framework diagram, demonstrating workflow for evaluating VLLMs' aesthetic understanding capabilities, and the structural design of the Whitebox and Blackbox evaluators.

between traditional design evaluation methods and modern AI assessment requirements.

## B. Evaluation Framework

Our aesthetic VQA task consists of two main components: Description Generation and Aesthetic Evaluation. In the Description Generation component, we use a fixed-prompt VLLM to generate descriptive text about the input image. In the Aesthetic Evaluation component, we utilize evaluators to analyze the consistency between the emotional word pairs extracted from the generated description and the input image, comparing them with ground truth labels to evaluate the VLLM's aesthetic understanding capabilities. For comprehensive evaluation, we designed two complementary evaluators: the Whitebox Evaluator (which ensures transparency and traceability by analyzing the generated text for linguistic assessment), which analyzes the generated text for transparent linguistic assessment, and the Blackbox Evaluator, which integrates image and text inputs for multimodal analysis.

To systematically evaluate VLLM's aesthetic understanding in design works, we further break down the evaluator into two steps to form a three-stage workflow as follows:

- **Description Generation Stage:** Given an input image  $I$  and a prompt  $P$ , the VLLM generates a description:  $D = \text{VLLM}(I, P)$ .
- **Classification Stage:** The image-description pair is processed by the classifier:  $C = \text{Evaluator}(I, D)$ , where  $C$  represents the predicted aesthetic word pairs, and

the Evaluator operates only on the description input, as illustrated in fig. 1

- **Evaluation Stage:** The classification results are matched against ground truth labels  $L$  to compute the final score  $\text{Score} = \text{Match}(C, L)$ , where  $L$  contains the ground truth aesthetic word pairs.

## C. Aesthetic Evaluators

We propose two evaluators to assess the aesthetic understanding capabilities of VLLMs: **Whitebox Evaluator**. The Whitebox Evaluator employs a constrained LLM to analyze generated descriptions. Given a predefined vocabulary of word pairs, the LLM systematically identifies and extracts relevant word pairs present in the descriptions. These extracted word pairs are then matched against the ground truth annotations to compute accuracy scores. This method directly evaluates the VLLM's ability to express design characteristics through appropriate semantic descriptors. **Blackbox Evaluator**. The Blackbox evaluator utilizes a trained classifier with a Q-former-like architecture [16] that processes both image and description inputs. The classifier has Q-former blocks, each of which includes two self-attention blocks and one cross attention block with a learnable query in the first Q-former block. For each semantic word pair in our vocabulary, the classifier employs two distinct classification heads, forming a multi-task learning setup: one for presence detection ( $A_i$ ) and another for polarity determination ( $B_i$ ). The classification is performed using the query embedding from the final Q-former block.


Image	Label	Description
	<b>Clean</b> -Cluttered <b>Modern</b> -Outdated <b>Elegant</b> -Gaudy <b>Plush</b> -Hard <b>Premium</b> -Basic <b>Durable</b> -Fragile <b>Flexible</b> -Rigid <b>Balanced</b> -Awkward <b>Pure</b> -Muddled <b>Refined</b> -Crude <b>Harmonious</b> -Discordant <b>Premium</b> -Basic <b>Professional</b> -Amateur <b>Smooth</b> -Rough <b>Sophisticated</b> -Simple <b>Intuitive</b> -Confusing <b>Seamless</b> -Disruptive <b>Solid</b> -Flimsy <b>Premium</b> -Cheap <b>Refined</b> -Crude <b>Comfortable</b> -Uncomfortable	Minimalistic white headphones with smooth, rounded contours offer a modern, fresh design. Comfortable ear cups and effective noise cancellation ensure an immersive experience, while the adjustable headband provides a personalized fit. Lightweight materials and metal accents add elegance, blending style with functionality for a premium audio experience.
	<b>Sturdy</b> -Weak, <b>Elegant</b> -Clumsy <b>Balanced</b> -Unstable <b>Premium</b> -Basic <b>Durable</b> -Fragile <b>Secure</b> -Loose <b>Gentle</b> -Harsh <b>Solid</b> -Wobbly <b>Weighted</b> -Light <b>Smooth</b> -Sticky <b>Precise</b> -Imprecise <b>Organized</b> -Messy <b>Thoughtful</b> -Careless <b>Sleek</b> -Bulky <b>Professional</b> -Amateur <b>Modern</b> -Outdated	This sleek headphone design features cushioned ear cups for comfort and an adjustable microphone for versatile use. Lightweight yet durable, the dark blue and black color scheme adds sophistication. Paired with a stand for organized storage, it offers a stylish and immersive audio experience.

Fig. 2: Examples of design images, its annotated labels, and descriptions generated by gpt-4o with prompt.

We introduce a dynamic thresholding mechanism, where the base threshold ( $T_{base}$ , typically set to 0.5), is adjusted by the confidence of the direction prediction through an adjustment factor  $k$ :  $T_i = T_{base} - k \times |B_i - 0.5|$ .

This adjustment allows the threshold to adapt based on the confidence of the polarity prediction: when  $B_i$  approaches 0.5 (uncertain direction), the threshold remains close to  $T_{base}$ ; when  $B_i$  deviates significantly from 0.5 (clear direction), the threshold is lowered, making it easier to pass the presence check.

The final decision for each semantic pair follows:

$$\text{Decision}_i = \begin{cases} \text{Positive} & \text{if } A_i > T_i \text{ and } B_i > 0.5 \\ \text{Negative} & \text{if } A_i > T_i \text{ and } B_i < 0.5 \\ \text{None} & \text{otherwise} \end{cases} \quad (1)$$

This multi-task approach provides a more nuanced evaluation of the VLLM’s aesthetic comprehension by simultaneously considering both the presence of semantic pairs and their directional tendency, which are then compared with ground truth labels for performance assessment.

Both evaluators yield quantitative metrics that reflect different aspects of the VLLM’s capability to understand and articulate design aesthetics. The Whitebox Evaluator focuses on emotional expression in the text and excels at capturing labels conveyed by the text—especially more abstract labels—but remains susceptible to “hallucination” issues. Meanwhile, the Blackbox Evaluator, through multimodal fusion and a trainable FC layer, more accurately captures visual features under instruction and avoids being confounded by stylistic or extraneous elements in the image, though it may be influenced by training data biases. These two evaluators contribute to a more comprehensive and reliable assessment of VLLMs’ ability to capture multi-dimensional design features.

## IV. EXPERIMENTS

### A. Experimental Setup

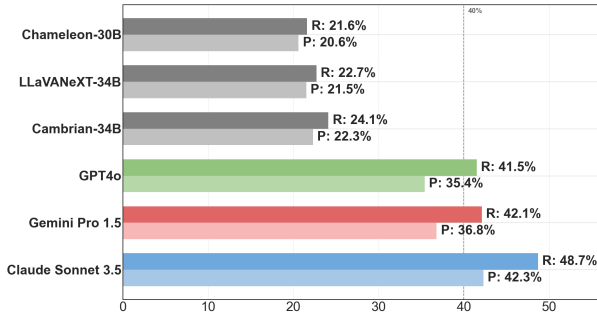
**Model Selection.** We evaluated six VLLMs, including three leading proprietary models (GPT-4o, Google Gemini Pro 1.5, and Claude Sonnet 3.5) and three open-source models (Cambrian-34B [17], LLaVaNExT-34B [18], and Chameleon-30B [19]) in vision-language understanding.

**Evaluation Setup.** We categorized the 84 pairs of emotional-polar descriptors into seven dimensions (textttVisual, Spatial, Abstract, Material, Tactile, Practical, and Craftsmanship) to enable a more systematic analysis of VLLMs’ performance across different aspects of design aesthetics. GPT-4o was selected in the Whitebox Evaluator because preliminary experiments showed it exhibits relatively better ability to comprehend and extract semantic emotional word pairs from descriptive texts. A specialized training process was implemented to optimize aesthetic evaluation performance for the Blackbox Evaluator. The training pipeline incorporates image-word pair labels directly into prompts, leveraging multiple LLMs to generate precise descriptions. These descriptions, combined with their corresponding images and word-pair labels, form our comprehensive training dataset. Pretrained weights from InstructBLIP were used as our foundation, and the training efforts were focused on the fully connected layers. The classifier was trained using a classification loss function specifically designed to predict the presence of aesthetic word pairs, ensuring accurate assessment of design characteristics.

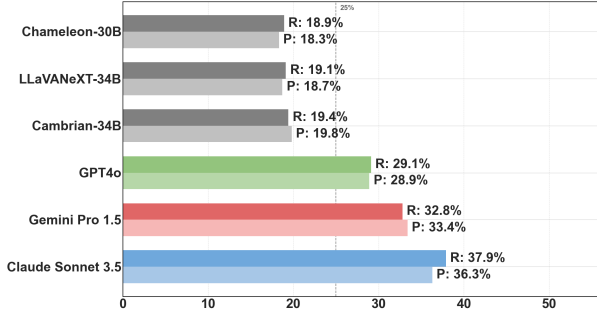
### B. Can VLLMs Express Design Aesthetics?

*a) Overall Classification Performance:* We begin by evaluating each model on the SenseDesign dataset using the Whitebox and Blackbox Evaluator (see section III-C). Our findings indicate that all tested models face substantial





(a) Results about Whitebox Evaluator



(b) Results about Blackbox Evaluator

Fig. 3: Performance comparison on different VLLMs using Precision (P) and Recall (R). The Blackbox Evaluator provides a stricter evaluation strategy than the Whitebox Evaluator.

challenges in design aesthetic understanding. As illustrated in fig. 3, both precision and recall remain relatively low, even under different evaluation strategies, reflecting VLLMs’ inherent limitations in capturing design nuances. In particular, recall generally surpasses precision, which can be attributed to the evaluators’ distinct mechanisms: (1) Whitebox Evaluator employs prompts that generate broader emotional descriptors, potentially capturing latent design features, (2) Blackbox Evaluator adopts a moderately relaxed threshold policy, ensuring the VLLMs’ outputs receive a sufficiently comprehensive assessment. Among the six models, the open-source ones—Cambrian-34B, LLaVaNExT-34B, and Chameleon-30—are significantly outperformed by proprietary counterparts. In contrast, **GPT-4o**, **Google Gemini Pro 1.5**, and **Claude Sonnet 3.5** demonstrate superior performance; hence, we focus on these three commercial models for further dimensional analysis. All reported data in the following sections represent the average performance of these three models.

*b) Dimensional Analysis via Radar Chart:* Figure 4 depicts the average recall for each dimension across models. Our analysis reveals four key limitations:

- **Text-Visual Gap.** Both evaluators exhibit alignment issues between textual and visual content, albeit in different forms and degrees. Under Whitebox Evaluator, the *Abstract* and *Spatial* dimensions show markedly higher recall, revealing how high-level descriptors (e.g., *conceptual*, *layered*) often lack sufficient visual grounding. Because Whitebox Evaluator matches text literally, it amplifies this misalignment whenever VLLMs invent ab-

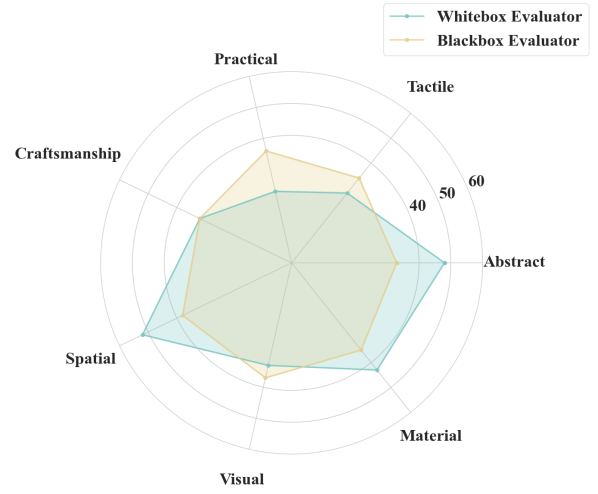


Fig. 4: Radar chart showing recall scores for Whitebox and Blackbox evaluators across seven dimensions.

stract terms. Although Blackbox Evaluator’s multimodal fusion partially alleviates such discrepancies, the fundamental alignment gap persists—merely being masked by a trainable classifier head. This points to a foundational shortcoming in VLLMs’ ability to connect language with corresponding visual evidence.

- **Limited Recognition of Fine Visual Details.** Both evaluators also exhibit weaknesses in capturing subtle design features, but manifest them differently. Blackbox Evaluator leverages fused visual-text features, potentially mitigating over-reliance on surface-level visual signals (e.g., brightness, contrast). However, its cautious approach leads to lower recall (around 30%), especially for complex material textures or intricate craftsmanship details. In contrast, Whitebox Evaluator achieves higher recall by permissively matching textual labels, rather than by accurately detecting fine-grained visual cues. Regardless of the evaluation approach, VLLMs still face significant hurdles in accurately recognizing nuanced design elements.
- **Material-Craftsmanship Understanding.** Under both evaluators, the *Material* and *Craftsmanship* categories consistently rank among the worst in performance. Even with explicit prompt descriptors, the models struggle to (i) visually differentiate complex surfaces (e.g., metallic sheen vs. matte texture), and (ii) contextualize their design implications (e.g., handcrafted details vs. mass-produced finishing).
- **Practical-Tactile Assessment.** Although *Practical* and *Tactile* fare slightly better than *Material* or *Craftsmanship*, the models remain inconsistent. Under Whitebox Evaluator, hints of “noise-canceling comfort” or “comfortable for prolonged wear” can inflate recall, even if the image itself provides minimal evidence for such user-centric claims. Conversely, when the tactile or usability cues are visually subtle (e.g., the cushion softness of a headset), Blackbox Evaluator frequently

misses these important features. Such observations underscore the inherent difficulty of extracting user-oriented or sensory attributes from purely visual data and relatively shallow text prompts.

### C. Discussion

We conducted an extended ablation study by selectively providing different types of emotional-polar word pairs to the VLLMs. Our experiments reveal two important findings: (1) when the model’s prompt includes additional *Visual* or *Spatial* labels, recall improves substantially (reaching 70–80% in certain configurations), and (2) when only *Abstract* or similar high-level descriptors are introduced, performance remains virtually unchanged. This outcome suggests that *Visual* or *Spatial* cues help VLLMs align textual labels with observable product attributes, whereas abstract descriptors demand richer context or specialized fine-tuning to effectively capture subtle or conceptual design features.

In our evaluations, three major factors emerge as key constraints on VLLMs’ ability to interpret design qualities:

- 1) **Inadequate Extraction of Subtle Visual Signals.** Although VLLMs can recognize basic shapes and colors, they frequently fail to register finer-grained indicators, such as specialized materials and intricate craftsmanship details. Consequently, their outputs may overlook critical design elements or resort to generalities.
- 2) **Mismatch of Visual Cues and Linguistic Constructs.** Even when VLLMs detect relevant visual attributes, they often lack domain-specific vocabulary to translate these observations into suitable aesthetic or emotional terminology (e.g., “hand-polished surfaces,” “artisan-like joinery”), leading to incomplete or inaccurate descriptions.
- 3) **Weak Abstract Reasoning in Emotional Word-Pairs.** Lastly, VLLMs struggle to bridge *visual/spatial* features and *emotional/aesthetic* implications, especially with high-level descriptors (e.g., “futuristic vs. retro”). Unlike human evaluators, who can intuitively map particular forms to symbolic or emotional meanings, current models rely heavily on superficial textual patterns, resulting in insufficient conceptual inference.

### V. CONCLUSION

In this work, we propose a structured evaluative framework using SDM and binary polarity semantic descriptors to more accurately assess aesthetics in design works, and we introduce the SenseDesign dataset under this framework to evaluate representative VLLMs. While we mitigate hallucinations and improve precision through prompt design, we acknowledge the prompts’ limited robustness. Nevertheless, the distinct performance gaps among VLLMs suggest that these shortcomings extend beyond prompt quality, pointing to deeper challenges for VLLMs. Our framework can thus facilitate future research in enhancing VLLMs’ aesthetic sense.

### REFERENCES

- [1] Mitsuo Nagamachi, “Kansei engineering: a new ergonomic consumer-oriented technology for product development,” *International Journal of industrial ergonomics*, vol. 15, no. 1, pp. 3–11, 1995.
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [3] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra, “Vqa: Visual question answering,” *International Journal of Computer Vision*, vol. 123, pp. 4 – 31, 2015.
- [4] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang, “An empirical study of gpt-3 for few-shot knowledge-based vqa,” *ArXiv*, vol. abs/2109.05014, 2021.
- [5] Drew A. Hudson and Christopher D. Manning, “Gqa: A new dataset for real-world visual reasoning and compositional question answering,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019.
- [6] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick, “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.
- [7] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi, “From recognition to cognition: Visual commonsense reasoning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6720–6731.
- [8] Bo Zhang, Li Niu, and Liqing Zhang, “Image composition assessment with saliency-augmented multi-pattern pooling,” *arXiv preprint arXiv:2104.03133*, 2021.
- [9] Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen, “Aesthetic critiques generation for photos,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3514–3523.
- [10] Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic, “Aesthetic image captioning from weakly-labelled photographs,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [11] Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma, “Thinking image color aesthetics assessment: Models, datasets and benchmarks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21838–21847.
- [12] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L. Rosin, “Towards artistic image aesthetics assessment: a large-scale dataset and a new method,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22388–22397.
- [13] Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin, “Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception,” *arXiv preprint arXiv:2401.08276*, 2024.
- [14] Yuxuan Lei, Dingkang Yang, Zhaoyu Chen, Jiawei Chen, Peng Zhai, and Lihua Zhang, “Large vision-language models as emotion recognizers in context awareness,” *ArXiv*, vol. abs/2407.11300, 2024.
- [15] Simon Schütte, Jörgen Anders Evert Eklund, Jan R. C. Axelsson, and Mitsuo Nagamachi, “Concepts, methods and tools in kansei engineering,” *Theoretical Issues in Ergonomics Science*, vol. 5, no. 3, pp. 214–232, 2004.
- [16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19730–19742.
- [17] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al., “Cambrian-1: A fully open, vision-centric exploration of multimodal llms,” *arXiv preprint arXiv:2406.16860*, 2024.
- [18] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li, “Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models,” *arXiv preprint arXiv:2407.07895*, 2024.
- [19] Chameleon Team, “Chameleon: Mixed-modal early-fusion foundation models,” *arXiv preprint arXiv:2405.09818*, 2024.