What Do Vision–Language Models Encode for Personalized Image Aesthetics Assessment?

Anonymous ACL submission

Abstract

Personalized image aesthetics assessment (PIAA) is an important research problem with practical real-world applications. While methods based on vision-language models (VLMs) are promising candidates for PIAA, it remains unclear whether they internally encode rich, multi-level aesthetic attributes required for effective personalization. In this paper, we first analyze the internal representations of VLMs to examine the presence and distribution of such aesthetic attributes, and then leverage them for lightweight, individual-level personalization without model fine-tuning. Our analysis reveals that VLMs encode diverse aesthetic attributes that propagate into the language decoder layers. Building on these representations, we demonstrate that simple linear models can perform PIAA effectively. We further analyze how aesthetic information is transferred across layers in different VLM architectures and across image domains. Our findings provide insights into how VLMs can be utilized for modeling subjective, individual aesthetic preferences.

1 Introduction

Image aesthetics assessment (IAA) is the task of evaluating the aesthetic quality of an input image. Recently, personalized image aesthetics assessment (PIAA) has attracted increasing attention in the IAA field. In this setting, models are trained to predict the aesthetics assessment that a specific user would assign to an image. Several datasets for PIAA have been proposed (Ren et al., 2017; Yang et al., 2022; Maerten et al., 2025), and they revealed substantial variation in aesthetic preferences across individuals. Given practical applications such as social media platforms, it is necessary to align assessment models with individual preferences.

In PIAA, image-level aesthetic attributes such as lighting and color have been leveraged to better reflect individual preferences. Many existing PIAA methods rely on training with large-scale, single-domain general image aesthetics assessment (GIAA) datasets (Shi et al., 2024; Zhu et al., 2022; Yun and Choo, 2024; Ren et al., 2017) to extract such attributes from the target images. However, such approaches require additional training costs. Their transferability across different domains, such as photographs and artworks, is also questionable.

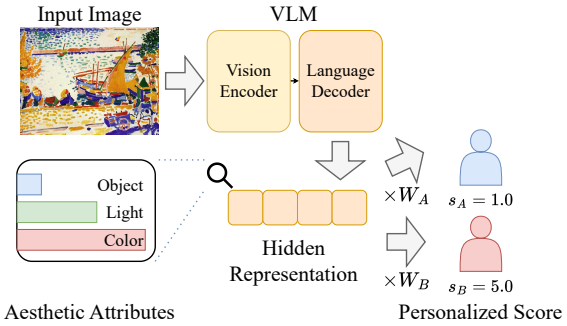

Figure 1: Overview of PIAA using VLM representations. s and W denote user-specific aesthetics scores and linear estimators respectively. Aesthetic attributes encoded in VLM hidden representations are linearly transformed to predict user-specific aesthetic scores without model fine-tuning.

To overcome these limitations, we propose a PIAA framework based on general vision-language models (VLMs). For GIAA, prior studies have already employed VLMs to leverage their rich text-based data source (Ke et al., 2023) or caption generation ability (Zhong et al., 2023). However, their application to PIAA tasks has been limited to optimization at the demographic-group level rather than the individual level (Li et al., 2025).

In this work, we aim to achieve individual-level personalization by leveraging aesthetic attributes implicitly encoded in VLMs through large-scale pretraining.[1] Previous studies using linear probing have shown that VLM hidden representations capture semantic information (Tao et al., 2024; Chandhok et al., 2025) as well as overall image

[1] Our code will be released publicly upon paper acceptance.

aesthetic scores (Hentschel et al., 2022). However, it remains unclear whether these representations also encode the diverse and continuous aesthetic attributes required for effective PIAA. Therefore, we first conduct linear probing on the hidden layers of VLMs to verify the existence of rich aesthetic attributes. We then apply linear regression to these representations and predict individual-level aesthetic scores, enabling personalized assessment without additional model fine-tuning. We further analyze the domain-dependent behavior of our approach using datasets from two distinct domains: photographs and artworks.

Our contributions are summarized as follows:

- We demonstrate that VLMs encode multiple aesthetic attributes beyond a single global aesthetic score within their hidden representations. Notably, this information propagates into the language decoder layers. Furthermore, we reveal that models with different architectures encode aesthetic attributes in different model regions.

- We show that simple linear regression on VLM hidden representations achieves strong PIAA performance, substantially outperforming methods based on text outputs such as few-shot prompting or fine-tuning. Our analysis indicates that the limited set of aesthetic attributes identified via linear probing plays a central role in personalization for photographs. Experiments on artwork datasets further suggest the presence of additional information in VLMs that contributes to PIAA but is not captured by probing based on photographs.

2 Related Work

2.1 Image Aesthetics Assessment

Several datasets have been proposed for general image aesthetics assessment (GIAA) over the past decade (Murray et al., 2012; Kong et al., 2016; Chang et al., 2017). To extend GIAA to personalized image aesthetics assessment (PIAA), several datasets with user-specific annotations have been introduced. FLICKER-AES (Ren et al., 2017) was among the earliest datasets to include personalized aesthetic ratings. PARA (Yang et al., 2022) and LAPIS (Maerten et al., 2025) further expanded this line of work by providing richer image attribute annotations and user-specific ratings for photographs and artworks, respectively.

A variety of methods have also been proposed for PIAA. Zhu et al. (2020) introduced a meta-learning approach to adapt models to individual users with limited annotations. Shi et al. (2024) and Zhu et al. (2022) modeled personalized aesthetics as interactions between image- and user-level attributes. Yun and Choo (2024) proposed representing personalization as parameter changes induced by fine-tuning on general IAA tasks, referred to as a "task vector," which is subsequently applied to user-specific prediction.

While these approaches have shown promising results, they typically require training on large-scale GIAA datasets followed by additional adaptation for each target user. Such multi-stage pipelines incur substantial computational cost, and their cross-domain transferability (e.g., from photographs to artworks) remains unclear.

2.2 Aesthetics Assessment with VLMs

Several studies have explored the use of textual descriptions to train aesthetic assessment models based on VLM architectures. Ke et al. (2023) pre-trained CoCa (Yu et al., 2022) using aesthetics-related captions and demonstrated its effectiveness for downstream aesthetic assessment tasks. Zhou et al. (2024a) generated synthetic textual descriptions of image aesthetic attributes using VLMs and leveraged these data to train an aesthetics-aware image encoder.

Several benchmarks and fine-tuning approaches have also been proposed to evaluate and improve VLMs for aesthetic assessment (Wu et al., 2023; Zhou et al., 2024b; Huang et al., 2024; Wu et al., 2024; Qi et al., 2025). More recently, Li et al. (2025) analyzed VLM behavior on PIAA datasets with respect to user attributes such as gender and age, revealing tendencies to over-align with specific demographic groups.

However, existing evaluations of VLMs' ability to perceive fine-grained aesthetic attributes are primarily limited to multiple-choice or binary formats. As a result, it remains unclear whether VLMs encode continuous, multi-level aesthetic attributes required for personalized image aesthetics assessment. Furthermore, to the best of our knowledge, no prior work has investigated individual-level PIAA using VLMs as the backbone model.

2.3 Representation Analysis of VLMs

Linear probing has been widely adopted as a tool for analyzing the internal representations of founda-

tion models across various visual tasks (El Banani et al., 2024; Kaltampanidis et al., 2025). Hentschel et al. (2022) applied linear probing to CLIP (Radford et al., 2021) to evaluate its understanding of general image aesthetics.

Another line of work emphasizes the importance of integrated analyses spanning multiple transformer layers across both vision encoders and language decoders. Tong et al. (2024) demonstrated that certain visual attributes are challenging for vision encoders trained with contrastive image-text objectives to capture. In contrast, Chandhok et al. (2025) demonstrated that information relevant to fine-grained recognition tends to diminish in language decoder layers while remaining more stable in vision encoders. Tao et al. (2024) conducted layer-wise probing of language decoders, suggesting that different layers encode different types of information. Several studies have further analyzed how visual information is transferred from vision tokens to text tokens through the language decoder (Kaduri et al., 2025; Zhang et al., 2025).

Despite these insights, most existing analyses involving language decoders focus on a limited set of tasks, such as object recognition or visual question answering. Although Hentschel et al. (2022) examined aesthetic understanding at the representation level, it remains unclear how multiple, fine-grained aesthetic attributes are encoded and propagated across different layers of VLMs. Importantly, our probing is designed not merely to assess overall aesthetic awareness, but to reveal multi-attribute representations that enable personalization.

3 Probing Aesthetic Attributes in VLMs

In this section, we investigate whether VLMs encode rich, multi-level aesthetic attributes that are relevant to PIAA.

3.1 Method Overview

We perform linear probing on the internal representations of VLMs to quantify the extent to which aesthetic attribute information is encoded. Formally, let I denote an input image, $\mathbf{v}_I \in \mathbb{R}^K$ the K-dimensional ground-truth aesthetic attribute vector (e.g., object, lighting, and color), and $\mathbf{h}(I) \in \mathbb{R}^D$ the D-dimensional hidden representation extracted from a VLM for image I. Our objective is to learn an image-agnostic linear transformation $M \in \mathbb{R}^{K \times D}$ such that

$$M\mathbf{h}(I) \approx \mathbf{v}_I. \quad (1)$$

In practice, we estimate M using ridge regression, which performs a stable estimation for high-dimensional representations while mitigating overfitting through L2 regularization. Detailed implementation is provided in Appendix A.3. For the hidden representation $\mathbf{h}(I)$, we extract output hidden vectors from each transformer layer of the VLM, using the image together with the prompt *"Assess the aesthetics of this image."* as input. To account for potential prompt sensitivity, we examine the robustness of probing results to alternative prompt formulations in Appendix C.3.

We use average pooling to obtain a single representation for each transformer layer, as visual information in VLMs is distributed across tokens and no dedicated image-level token exists. This provides a simple, modality-agnostic aggregation that enables fair comparison across vision and language representations.

We consider three layer-wise representations obtained via average pooling: $\mathbf{V}_i$ (vision encoder), $\mathbf{LT}_i$ (language decoder, text tokens), and $\mathbf{LV}_i$ (language decoder, vision tokens). We primarily focus on comparisons involving $\mathbf{LT}_i$, since aesthetic attribute encoding in language decoder layers remains less explored in prior work (Hentschel et al., 2022) and $\mathbf{LT}_i$ representations directly support text-based outputs such as aesthetic scores and captions. For completeness, we also compare $\mathbf{LT}_i$ with the last text-token representation in the language decoder in Appendix B.1.

3.2 Settings

Datasets We use two photographic datasets for the probing experiments. Our primary dataset is AADB (Kong et al., 2016), which provides 11-dimensional aesthetic attribute annotations alongside overall aesthetic scores, with continuous values in the range $[-1, 1]$. We also conduct probing experiments on PARA (Yang et al., 2022), which includes several general aesthetic attribute annotations. For each hidden representation, we train attribute regressors on the training split and report evaluation metrics on the test split.

Note that the attribute annotations in PARA exhibit strong inter-attribute correlations as shown in Appendix A.5. Such correlations are not well aligned with our goal of identifying diverse and independent aesthetic attributes encoded in VLM representations. Therefore, we report results on AADB, where inter-attribute correlations are relatively low, as the primary probing results in the fol-

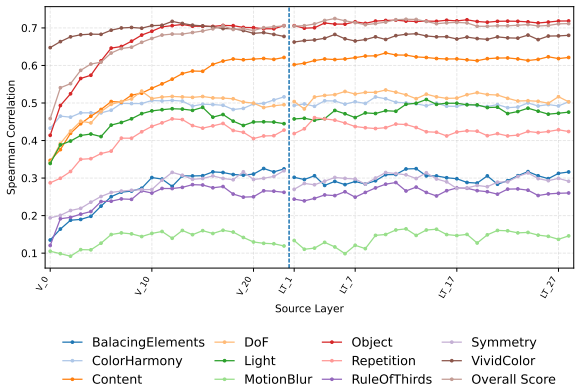

Figure 2: Layer-wise probing performance across **V** and **LT** layers for Qwen3-VL 2B on AADB. The dotted line indicates the boundary between V and LT.

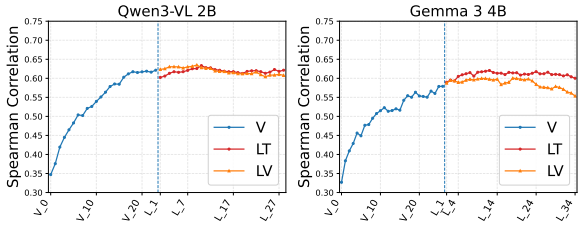

Figure 3: Layer-wise probing performance for the *Content* attribute in Qwen3-VL 2B and Gemma 3 4B.

lowing section, and present PARA-based probing results in Appendix C.2 as supplementary material.

Models We evaluate two state-of-the-art open-source VLMs with distinct architectures: Qwen3-VL (Bai et al., 2025) and Gemma 3 (Team et al., 2025). Gemma 3 uses a fixed number of visual tokens and feeds only the final vision encoder output into the language decoder, whereas Qwen3-VL produces resolution-dependent vision tokens and integrates multi-level vision representations via DeepStack (Meng et al., 2024). All models are instruction-tuned variants, and we evaluate multiple model sizes (2B, 4B, and 8B for Qwen3-VL; 4B and 12B for Gemma 3).

We additionally include DINOv3 (ViT-B/16, ViT-L/16) (Siméoni et al., 2025) as a vision-only foundation model for comparison.

Evaluation Methods We use Spearman's rank correlation coefficient as the primary evaluation metric. For each model, we first report the best correlation obtained across different $\mathbf{LT}_i$ layers (or $\mathbf{V}_i$ layers for DINOv3). We then compare these values with the corresponding results from $\mathbf{LV}_i$ and $\mathbf{V}_i$ representations.

3.3 Results

Overall Results The main results on **LT** layers (and **V** layers for DINOv3) are summarized in Table 1. Results for the other representations are reported in Appendix B.1. Layer-wise results for all **V** and **LT** layers of Qwen3-VL 2B are further visualized in Figure 2.

Across more than half of the aesthetic attributes, all VLMs achieve moderate positive correlations (greater than 0.4). For the remaining attributes, which exhibit sparse label distributions (see Appendix A.5), the models still achieve consistently positive correlations. We observe similar trends across alternative prompt formulations reported in Appendix C.3. These results indicate that VLMs encode a diverse set of aesthetic attributes in their hidden representations in a manner that is linearly accessible. To address the possibility that correlated attributes or spurious visual cues drive these effects, we conduct additional robustness analyses based on controlled image augmentations, which are reported in Appendix C.1.

When comparing different aesthetic attributes, probing performance for the overall aesthetic score is consistently higher than that for most fine-grained attributes. This observation suggests that the ability to capture general aesthetics, previously verified through the probing study by Hentschel et al. (2022), does not necessarily imply a robust encoding of fine-grained aesthetic attributes.

Comparison of Models and Components Interestingly, Qwen3-VL 2B, 4B, and 8B achieve the best performance for different attributes, respectively. The fact that smaller models within the same model family can outperform larger variants suggests that aesthetic attribute encoding is not directly correlated with conventional VLM benchmark performance, such as visual question answering.

Another notable observation from Figure 2 is that language decoder representations achieve higher correlations for a larger number of attributes than vision encoder representations. While vision encoder layers achieve better performance for specific attributes (e.g., *VividColor* in Qwen3-VL 2B), no substantial performance degradation is observed in the language decoder layers. Moreover, as Table 1 shows, DINOv3 consistently yields the lowest correlations across nearly all attributes. Together, these findings suggest that language decoder layers play an important role in encoding aesthetic attribute information beyond what is captured by

Attribute	Qwen3-VL			Gemma 3		DINOv3	
	2B	4B	8B	4B	12B	ViT-B/16	ViT-L/16
BalancingElements	0.325 (13)	**0.332** (0)	0.300 (14)	0.309 (24)	0.307 (14)	0.317 (11)	0.307 (22)
ColorHarmony	0.516 (9)	**0.523** (9)	0.515 (6)	0.493 (24)	0.504 (32)	0.479 (6)	0.482 (10)
Content	**0.633** (10)	0.632 (13)	0.627 (15)	0.621 (12)	0.624 (35)	0.551 (12)	0.579 (17)
DoF	**0.535** (10)	0.518 (10)	0.530 (16)	0.512 (9)	0.515 (15)	0.506 (7)	0.507 (18)
Light	**0.509** (14)	0.507 (12)	0.490 (14)	0.452 (18)	0.468 (6)	0.439 (8)	0.436 (11)
MotionBlur	0.165 (12)	0.134 (5)	**0.188** (9)	0.155 (1)	0.152 (37)	0.161 (7)	0.143 (3)
Object	**0.722** (18)	0.719 (19)	0.716 (16)	0.706 (18)	0.714 (7)	0.688 (12)	0.696 (19)
Repetition	**0.461** (3)	0.446 (14)	0.451 (4)	0.415 (8)	0.430 (20)	0.438 (8)	0.451 (19)
RuleOfThirds	**0.288** (11)	0.267 (5)	0.266 (0)	0.267 (12)	0.273 (15)	0.230 (9)	0.230 (20)
Symmetry	0.315 (10)	**0.329** (14)	0.307 (6)	0.281 (11)	0.302 (33)	0.299 (5)	0.313 (11)
VividColor	0.686 (0)	0.695 (11)	**0.696** (0)	0.671 (10)	0.687 (18)	0.686 (3)	0.685 (10)
Overall Score	0.725 (5)	**0.727** (19)	0.720 (10)	0.700 (10)	0.719 (13)	0.636 (9)	0.666 (17)

Table 1: Highest Spearman correlation achieved by linear probing on **LT** layers for each aesthetic attribute in AADB. Values in parentheses indicate the corresponding layer indices.

vision-only foundation models.

Layer-wise Analysis We further analyze layer-wise probing performance for the *Content* attribute across **V**, **LT**, and **LV** representations in Figure 3. For Gemma 3, performance in **LT** representations improves notably in the early to middle layers of the language decoder as layer depth increases. This observation is consistent with prior studies (Kaduri et al., 2025; Zhang et al., 2025), which report that visual information relevant to textual outputs is transferred to text tokens in the lower to middle layers of the language decoder. In contrast, this trend is not observed for Qwen3-VL, where probing performance for **LT** and **LV** representations remains comparable across layers. This pattern is consistently observed across multiple attributes.

We hypothesize that this architectural difference stems from the way aesthetic information is integrated across different modalities. Specifically, Gemma 3 may process aesthetics information primarily within the language decoder, exhibiting its potential dependence on text supervision. At the same time, Qwen3-VL may encode a larger part of such information within the vision encoder due to its DeepStack architecture.

4 PIAA with VLMs

In this section, we perform PIAA using hidden representations of VLMs, leveraging the rich aesthetic attributes identified in Section 3.

4.1 Method Overview

We aim to predict personalized aesthetic scores, which may vary across users for the same image.

Formally, let u denote a target user, I an image, $\mathbf{h}(I)$ a hidden representation extracted from a VLM, and $s_{I,u}$ the personalized aesthetic score assigned by user u to image I. Our objective is to learn a user-specific, image-agnostic linear transformation M_u such that

$$M_u \mathbf{h}(I) \approx s_{I,u}. \tag{2}$$

We train M_u using user-specific training data and evaluate its performance on held-out test data for the same user.

As in Section 3, we estimate M_u via ridge regression. For $\mathbf{h}(I)$, we reuse the $\mathbf{V}_i$, $\mathbf{LT}_i$, and $\mathbf{LV}_i$ representations defined in Section 3. Based on the probing analysis in Section 3, we observe that the layer at which aesthetic attribute information peaks varies across models and attributes. In contrast, language decoder representations in the middle layers consistently contain substantial information. Accordingly, we use $\mathbf{LT}_{15}$ as a representative layer for reporting the main PIAA results, as it provides a stable and informative representation across different models. The same prompt used for representation extraction in Section 3 is also adopted here. We refer to this primary approach as **Linear-Hidden**.

4.2 Baselines

We compare Linear-Hidden against several VLM-based baselines designed to evaluate different aspects of personalization.

To better understand the source of personalization effects, we first consider two variants of the Linear-Hidden model. In **Linear-Hidden (GIAA)**, we replace personalized scores with non-user-specific GIAA scores as regression targets. This setting isolates general aesthetic perception from user-specific preference modeling.

In **Linear-Hidden (Reduce)**, we first train a user-agnostic regressor M on AADB to predict

aesthetic attributes excluding the overall score, and then train a user-specific regressor M'_u such that

$$M'_u(M\mathbf{h}(I)) \approx s_{I,u}. \quad (3)$$

Since the intermediate transformation M substantially reduces representation dimensionality, this variant evaluates whether the aesthetic attributes identified in Section 3 are sufficient to support PIAA prediction.

In addition, we include text-based baselines that do not directly access hidden representations. We prompt VLMs to output GIAA scores as text without user-specific conditioning (**Raw Text**), as well as conventional adaptation methods (**Few-shot** and **LoRA**). We also include the **Adjust-Bias** baseline, which applies a user-specific additive bias correction to **Raw Text** predictions based on training-set errors, allowing us to disentangle score calibration from preference ordering. Implementation details of all baselines are provided in Appendix A.4.

4.3 Other Settings

We conduct experiments on two PIAA datasets from different domains introduced in Section 2.1: PARA (Yang et al., 2022), which consists of photographs, and LAPIS (Maerten et al., 2025), which focuses on artworks. For each dataset, we randomly sample 200 users. For each user, we construct a personalized support set with either 10 images (small setting) or 100 images (large setting), and reserve 50 images as a personalized test set. Unless otherwise specified, we report results under the 100-shot setting. Due to the high memory cost of long-context prompting, we use the 10-shot setting for the **Few-shot** baseline. Full results for both 10-shot and 100-shot settings are provided in Appendix B.2. Since LAPIS annotations are provided on a $[0, 100]$ scale whereas PARA uses a $[1, 5]$ scale, we linearly rescale LAPIS annotations to the $[1, 5]$ range prior to training and evaluation.

As target models, we use the same Qwen3-VL and Gemma 3 models evaluated in Section 3.

While prior work on PARA (Yang et al., 2022) reports correlation values aggregated over all test subjects as the primary evaluation metric, we observe that such metrics can be artificially improved through simple per-user, image-agnostic numeric adjustments applied uniformly across all images. To disentangle this numeric calibration effect from genuine user-specific preference modeling, we evaluate performance at the individual user level.

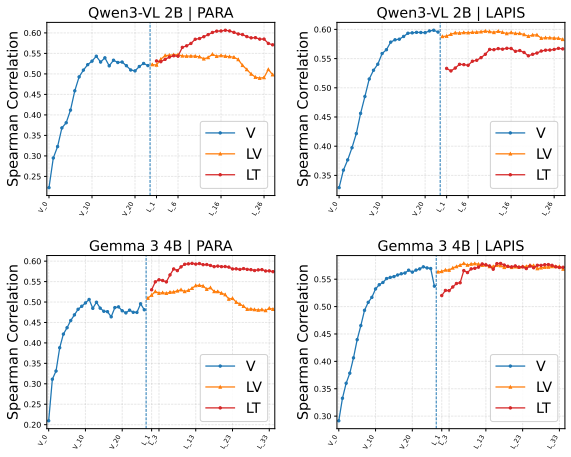

Figure 4: Layer-wise PIAA performance across **V**, **LV**, and **LT** representations for multiple models and datasets.

Specifically, we compute Spearman's rank correlation coefficient (ρ) and the coefficient of determination (R^2) separately for each user to capture complementary aspects of personalization performance: ρ measures the consistency of relative preference ordering. At the same time, R^2 reflects the accuracy of absolute score prediction. We then report the instance-averaged metrics across users as our main evaluation results.

4.4 Results

Overall Results Table 2 summarizes the main PIAA results. Complete results for different support set sizes are provided in Appendix B.2.

Across all models and both datasets, the Linear-Hidden approach consistently outperforms text-based baselines, including Raw Text, Few-shot prompting, and LoRA fine-tuning. Notably, on the LAPIS dataset, while text-based baselines such as Raw Text and Few-shot prompting already exhibit low performance, LoRA performs even worse. One possible explanation is the difficulty of learning fine-grained image–score relationships using token-level likelihood objectives under limited data. Despite this difficulty, Linear-Hidden achieves high Spearman correlations (above 0.5). These results indicate that language decoder representations in VLMs encode sufficiently rich information for image aesthetics assessment.

Domain Comparison Results obtained with Linear-Hidden variants reveal domain-specific characteristics. While Linear-Hidden (GIAA) yields substantially worse R^2 values than the full

Method	Support	Qwen3VL 2B ρ / R^2	Qwen3VL 4B ρ / R^2	Qwen3VL 8B ρ / R^2	Gemma 3 4B ρ / R^2	Gemma 3 12B ρ / R^2
PARA						
Raw Text		0.504 / -0.571	0.570 / -1.277	0.528 / -0.729	0.462 / -1.107	0.493 / -1.879
Few-shot	10-shot	0.319 / -1.850	0.197 / -1.576	0.372 / -0.547	0.241 / -0.537	0.407 / -0.185
Adjust-Bias	100-shot	0.504 / -0.310	0.570 / -0.672	0.528 / -0.441	0.462 / -0.321	0.493 / -1.562
LoRA	100-shot	0.487 / -1.970	0.578 / -1.751	0.568 / -0.978	0.489 / -0.893	0.524 / -0.525
Linear-Hidden	100-shot	**0.604** / 0.363	**0.611** / 0.362	0.591 / **0.341**	0.591 / 0.346	0.594 / 0.329
Linear-Hidden (GIAA)	100-shot	0.596 / 0.041	0.603 / 0.057	**0.596** / 0.043	0.584 / -0.014	**0.594** / 0.036
Linear-Hidden (Reduce)	100-shot	0.585 / **0.367**	0.597 / **0.382**	0.558 / 0.322	**0.592** / **0.365**	0.593 / **0.373**
LAPIS						
Raw Text		0.098 / -0.778	0.176 / -0.937	0.175 / -0.763	0.119 / -1.340	0.233 / -1.335
Few-shot	10-shot	0.142 / -1.265	0.221 / -0.380	0.264 / -0.480	0.127 / -0.354	0.227 / -0.459
Adjust-Bias	100-shot	0.098 / -0.264	0.176 / -0.231	0.175 / -0.206	0.119 / -0.162	0.233 / -0.442
LoRA	100-shot	0.026 / -0.701	0.153 / -1.580	0.164 / -1.386	0.116 / -0.936	0.201 / -1.022
Linear-Hidden	100-shot	**0.568** / **0.321**	**0.568** / **0.319**	**0.573** / **0.313**	**0.568** / **0.328**	**0.571** / **0.323**
Linear-Hidden (GIAA)	100-shot	0.418 / -0.148	0.420 / -0.148	0.420 / -0.151	0.413 / -0.153	0.416 / -0.155
Linear-Hidden (Reduce)	100-shot	0.480 / 0.224	0.468 / 0.202	0.459 / 0.197	0.469 / 0.220	0.446 / 0.189

Table 2: User-averaged PIAA performance on PARA and LAPIS. Each cell reports Spearman's ρ and R^2. Best values per column are highlighted in bold.

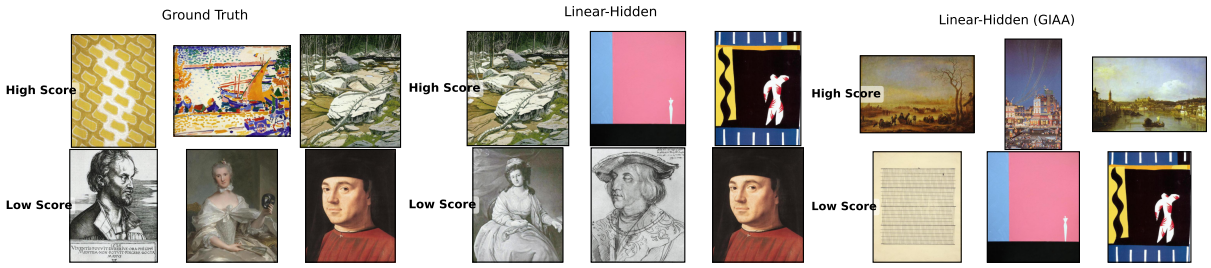

Figure 5: Examples of the LAPIS images assigned high (top) and low (bottom) scores by different methods for a representative user.

PIAA setting, the two settings exhibit only minor differences in Spearman correlation. This suggests that general aesthetics scores primarily contribute to image-agnostic numerical calibration across users, but do not fully capture individual preference ordering.

In contrast, training with user-specific PIAA labels results in significant improvements in both Spearman correlation and R^2 on the LAPIS dataset. Figure 5 illustrates qualitative differences between GIAA-based and PIAA-based predictions for a representative user in LAPIS. While the GIAA-based model favors fine-grained drawings, the PIAA-based model more accurately reflects the user's preference for colorful and abstract artworks, as well as their disinterest in realistic human portraits.

Comparisons with Linear-Hidden (Reduce) further highlight domain-specific behavior. On PARA, the Reduced variant achieves performance comparable to the complete Linear-Hidden model. However, on LAPIS, a clear performance gap exists between the two methods. This observation suggests that the 11-dimensional aesthetic attribute space identified in Section 3 is sufficient for PIAA in photographs, whereas personalization in the artwork domain relies on additional attributes that are also encoded in VLM representations but are not captured by the probed attributes.

Layer-wise Analysis Layer-wise comparisons across **V**, **LV**, and **LT** representations are shown in Figure 4. For PARA, both Qwen3-VL and Gemma 3 exhibit peak performance in the middle layers of **LT**, with **LT** consistently outperforming **V** and **LV** across layers. In contrast, this trend does not hold for LAPIS. Moreover, for Qwen3-VL, **LV** representations consistently outperform **LT** across layers on LAPIS.

Method	Support	Qwen3VL 2B ρ / R^2	Qwen3VL 4B ρ / R^2	Qwen3VL 8B ρ / R^2	Gemma 3 4B ρ / R^2	Gemma 3 12B ρ / R^2
Raw Text		0.380 / -2.292	0.405 / -4.308	0.387 / -2.943	0.347 / -3.109	0.330 / -6.295
Linear-Hidden	100-shot	**0.467** / 0.058	**0.472** / 0.053	**0.463** / 0.059	**0.463** / 0.079	**0.458** / 0.018
Linear-Hidden (GIAA)	100-shot	0.428 / -0.908	0.435 / -0.867	0.440 / -0.894	0.422 / -1.093	0.432 / -0.930
Linear-Hidden (Reduce)	100-shot	0.431 / **0.150**	0.447 / **0.157**	0.447 / **0.085**	0.436 / **0.145**	0.450 / **0.159**

Table 3: PIAA performance on PARA for users with low agreement to GIAA ("hard" users).

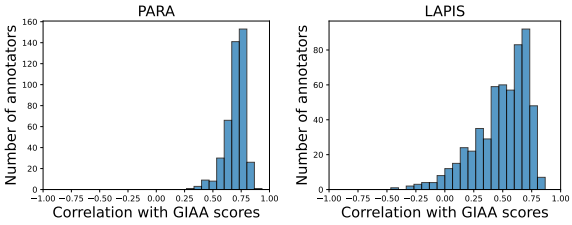

Figure 6: Distribution of Spearman correlation between PIAA and GIAA scores across users in PARA and LAPIS.

We hypothesize that these differences are due to the domain-specific availability of aesthetics-related textual supervision. While photographs benefit from rich caption and critique datasets containing aesthetic content (Ghosal et al., 2019; Qi et al., 2025; Chang et al., 2017), comparable resources are scarce for artworks. As a result, instruction-tuned VLMs may integrate aesthetic information into text tokens for photographs, but rely more heavily on vision-side representations for artworks.

4.5 Additional Analysis on "Hard" Users

We observe that the personalization effect of Linear-Hidden is weaker on PARA than on LAPIS. However, this difference may partly be because of the label distribution characteristics. As shown in Figure 6, annotator preferences in PARA exhibit substantially higher agreement with GIAA scores compared to those in LAPIS. This indicates that personalization is fundamentally more challenging on PARA, as user-specific variations from general aesthetics are limited. To isolate this effect from the domain difference, we resample 50 users from PARA with the lowest Spearman correlation between their annotations and GIAA scores, and repeat the PIAA experiments.

Results for these "hard" users are presented in Table 3. Although the performance gap remains smaller than that observed on LAPIS, regressors trained with PIAA labels show clear improvements over GIAA-based baselines. This finding demonstrates that VLM-based personalization is feasible for photographs when annotator preferences are sufficiently distinct from one another.

Notably, under this setting, the Reduced variant consistently underperforms the full Linear-Hidden in terms of Spearman correlation. This suggests that personalization for specific preferences in photographs requires additional attributes beyond those identified through our linear probing. At the same time, the Reduced variant exhibits improved R^2 scores compared to the full Linear-Hidden, which may be attributed to enhanced numerical stability resulting from the reduced number of explanatory variables.

5 Conclusion

In this paper, we investigated what aesthetic attributes are encoded within VLMs and how such representations can be leveraged for PIAA.

Through linear probing, we demonstrated that VLMs encode a diverse set of aesthetic attributes, including those that propagate into language decoder layers. We further demonstrated that these hidden representations provide effective signals for individual-level personalization. Our analysis also revealed domain-dependent differences in how aesthetic attributes relevant to PIAA are represented across vision and language components of VLMs.

We identify two main directions for future work. First, further investigation is needed to uncover additional aesthetic attributes that are not captured by current probes, particularly those encoded in vision tokens for artwork domains. Second, an important next step is to translate these representation-level insights into improved personalization of text-based behaviors in VLMs, such as aesthetic judgments and caption generation.

We hope that our findings contribute to a deeper understanding of aesthetic representations in multimodal language models and the future development of VLMs that are better aligned with subjective, individual user preferences.

Limitations

Due to the difficulty of obtaining annotations for rich aesthetic attributes and personalized image aesthetics assessment, our experiments rely on a limited number of existing datasets. Accordingly, the following limitations should be considered when interpreting our results:

- The set of aesthetic attributes used in this study is not exhaustive. In addition, dataset-specific correlations among attributes may have influenced the linear probing results.
- For PIAA evaluation, we use a single dataset for each domain. As a result, trends interpreted as domain-specific may partly reflect dataset-specific biases, including those introduced by the annotator populations.

Regarding linear probing, it is important to note that the presence of linearly accessible information in hidden representations does not necessarily imply that VLMs directly utilize such information during text generation or scoring. A more fine-grained analysis at the module or neuron level would be required to establish a closer connection between representation-level findings and model behavior at the output level.

Finally, our implementation of VLM-based PIAA baselines leaves room for improvement. In particular, methods that directly interpret textual outputs as numeric scores discard information relevant to regression objectives, such as the relative proximity between scores. More task-aligned optimization strategies could improve PIAA performance even for approaches that rely on text-based outputs.

Ethical Considerations

Although the datasets used in this work include information related to annotator identity, such information is not utilized in our experiments. Our analysis relies solely on subjective aesthetic scores associated with images, which poses minimal risk of personal data leakage.

Nevertheless, the proposed methods and findings could be combined with downstream recommendation or ranking systems that incorporate user attributes. In such scenarios, any information that could enable personal identification should be handled with appropriate care and in accordance with relevant privacy regulations.

Additionally, personalization may introduce or amplify biases in aesthetics assessments for specific demographic groups. As highlighted in prior studies, more comprehensive bias analyses are necessary before deploying personalized aesthetic assessment systems in real-world or industrial settings.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. 2020. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125.

Shivam Chandhok, Wan-Cyuan Fan, Vered Shwartz, Vineeth N. Balasubramanian, and Leonid Sigal. 2025. Response wide shut? surprising observations in basic vision language model capabilities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25530–25545, Vienna, Austria. Association for Computational Linguistics.

Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen. 2017. Aesthetic critiques generation for photos. In *Proceedings of the IEEE international conference on computer vision*, pages 3514–3523.

Alex Clark. 2015. Pillow (pil fork) documentation.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. 2024. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806.

Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic. 2019. Aesthetic image captioning from weakly-labelled photographs. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0.

Simon Hentschel, Konstantin Kobs, and Andreas Hotho. 2022. Clip knows image aesthetics. *Frontiers in Artificial Intelligence*, 5:976235.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang,

Leida Li, and Weisi Lin. 2024. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*.

Omri Kaduri, Shai Bagon, and Tali Dekel. 2025. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14549–14558.

Yannis Kaltampanidis, Alexandros Doumanoglou, and Dimitrios Zarpalas. 2025. Which direction to choose? an analysis on the representation power of self-supervised vits in downstream tasks. In *World Conference on Explainable Artificial Intelligence*, pages 376–399. Springer.

Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. 2023. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10041–10051.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*.

Kun Li, Lai Man Po, Hongzheng Yang, Xuyuan Xu, Kangcheng Liu, and Yuzhi Zhao. 2025. AesBias-Bench: Evaluating bias and alignment in multimodal language models for personalized image aesthetic assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7618–7631, Suzhou, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Anne-Sofie Maerten, Li-Wei Chen, Stefanie De Winter, Christophe Bossens, and Johan Wagemans. 2025. Lapis: A novel dataset for personalized image aesthetic assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6302–6311.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. 2024. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. 2025. The photographer's eye: Teaching multimodal large language models to see, and critique like photographers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24807–24816.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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

Huiying Shi, Jing Guo, Yongzhen Ke, Kai Wang, Shuai Yang, Fan Qin, and Liming Chen. 2024. Personalized image aesthetics assessment based on graph neural network and collaborative filtering. *Knowledge-Based Systems*, 294:111749.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, and 1 others. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. 2024. Probing multimodal large language models for global and local semantic representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13050–13056, Torino, Italia. ELRA and ICCL.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and 1 others. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and 1 others. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, and 1 others. 2024. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25490–25500.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. 2022. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19861–19869.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Jooyeol Yun and Jaegul Choo. 2024. Scaling up personalized image aesthetic assessment via task vector customization. In *European Conference on Computer Vision*, pages 323–339. Springer.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19781–19791.

Zhipeng Zhong, Fei Zhou, and Guoping Qiu. 2023. Aesthetically relevant image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3733–3741.

Hantao Zhou, Longxiang Tang, Rui Yang, Guanyi Qin, Yan Zhang, Yutao Li, Xiu Li, Runze Hu, and Guangtao Zhai. 2024a. Uniqa: Unified vision-language pre-training for image quality and aesthetic assessment. *arXiv preprint arXiv:2406.01069*.

Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. 2024b. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*.

Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. 2020. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics*, 52(3):1798–1811.

Hancheng Zhu, Yong Zhou, Zhiwen Shao, Wenliang Du, Guangcheng Wang, and Qiaoyue Li. 2022. Personalized image aesthetics assessment via multi-attribute interactive reasoning. *Mathematics*, 10(22):4181.

A Implementation Details

A.1 Models

All experiments load VLMs using the Transformers (Wolf et al., 2020) library.[2][3][4]. For experiments with high memory requirements, such as LoRA fine-tuning, we fix the floating-point precision to bfloat16. For other experiments, we preserve the original precision of the released model weights by specifying "auto" for the `torch_dtype` parameter.

For all experiments, we fix random seeds for data sampling and training procedures where applicable, and use deterministic decoding for text generation.

A.2 Computational Resources

All experiments were conducted on a single computing node equipped with one NVIDIA H100 GPU. The few-shot (100-shot) baseline experiments reported in Appendix B.2 required approximately 24 GPU-hours. All other experiments, including linear probing and PIAA prediction, were completed within approximately 8 GPU-hours.

A.3 Regressor Implementation

All experiments that involve ridge regression, including linear probing in Section 3 and the

[2] https://huggingface.co/collections/Qwen/qwen3-vl

[3] https://huggingface.co/docs/transformers/main/model_doc/dinov3

[4] https://huggingface.co/collections/google/gemma-3-release

```
You are an expert judge of image aesthetics.
I will show you some example images with
this user's ratings on a 1 to 5 scale.
From these examples, infer the user's
personal  preferences.
Then I will show you a new image; please
predict this user's rating for it.
For the examples, each rating is this user's
own rating, already mapped to a 1 to 5 scale.
When you answer for the final image, respond
with a single number from 1 to 5, and
nothing else.

{% for idx in range(len(support_set)) %}
{{ images[idx] }}
Example {{ idx + 1}}. This user rated this
image {{ images[idx].score }} out of 5.
{% endfor %}

Now, based on the user's previous ratings,
what is this user's rating for THIS image?
Answer with a single number from 1 to 5,
and do not output any other text.
{{ target_image }}
```

Figure 7: Prompt template used for the Few-shot PIAA baseline.

Linear-Hidden methods in Section 4, are implemented using a scikit-learn (Pedregosa et al., 2011) pipeline consisting of *StandardScaler* followed by *RidgeCV*.

Formally, the regression objective minimizes the following loss [5] with respect to the parameter vector $\mathbf{w}$:

$$\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \alpha\|\mathbf{w}\|_2^2, \tag{4}$$

where $\mathbf{X}$ denotes the matrix of VLM features standardized to zero mean and unit variance, and $\mathbf{y}$ denotes the corresponding ground-truth labels. The regularization coefficient α controls the strength of the L2 penalty.

In our experiments, α is selected via cross-validation on the training set from 13 logarithmically spaced candidates in the range $[10^{-3}, 10^{3}]$.

A.4 PIAA Baselines

This section lists the baselines used in Section 4 and provides implementation details for each method.

Raw Text For the **Raw Text** baseline, we prompt the VLM to output a general image aesthetics assessment (GIAA) score without any user-specific conditioning. We use the following instruction to obtain a scalar score from the model: *"Assess the overall aesthetic quality of this image. Please rate it on a scale from 1 to 5. Output only the numeric score, and do not output any other text."*

Adjust-Bias For the **Adjust-Bias** baseline, we first obtain GIAA predictions for both the support and test sets using the same prompt as in **Raw Text**. We then estimate a user-specific bias term based on the support set and subtract it from test-time predictions to obtain personalized scores.

Formally, let N denote the size of the support set, I_i the i-th image in the support set, s_i the score assigned by the target user u, and $v(I_i)$ the GIAA score predicted by the VLM. The bias term b_u and the personalized prediction $s_u(I)$ for a test image I are computed as:

$$b_u = \frac{1}{N}\sum_{i=1}^{N}\big(v(I_i) - s_i\big), \tag{5}$$

$$s_u(I) = v(I) - b_u. \tag{6}$$

Few-shot For the **Few-shot** baseline, we construct prompts that interleave example images with corresponding personalized scores from the support set, enabling the model to infer user-specific scoring tendencies from a small number of demonstrations. Figure 7 illustrates a pseudo-prompt showing how few-shot image-score pairs are integrated into the prompt.

LoRA For the **LoRA** baseline, we perform user-specific LoRA fine-tuning using the PEFT (Mangrulkar et al., 2022) library. We apply LoRA to all linear layers of the VLM. Unless otherwise specified, we set the LoRA hyperparameters to $\alpha = 16$, rank $r = 8$, and dropout $= 0.1$, following standard practice. All other parameters use the default settings provided by the PEFT library.

We use the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 1×10^{-4}. A linear learning-rate scheduler with warmup is applied using `get_linear_schedule_with_warmup` from Transformers (Wolf et al., 2020), with the warmup period set to 10% of the total training steps.

For each user, we train the model for three epochs. The batch size is set to 4 by default, but reduced to 2 for Qwen3-VL 8B due to memory constraints.

Due to the high resolution of images in PARA and the variable vision token length in Qwen3-VL, we resize images for the LoRA baseline on Qwen3-

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

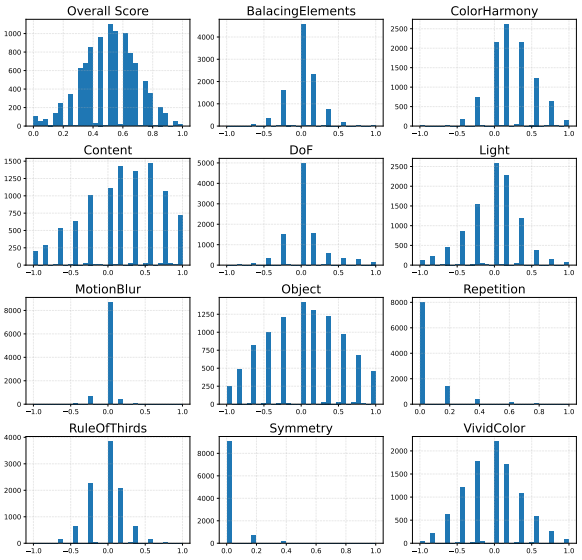

Figure 8: Distribution of aesthetic attribute annotations in AADB.

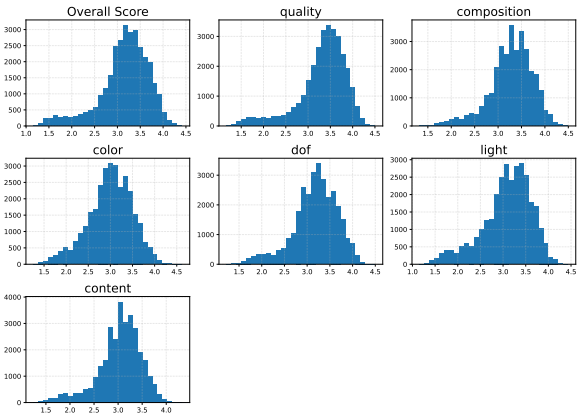

Figure 9: Distribution of aesthetic attribute annotations in PARA.

VL models such that the longer side does not exceed 1024 pixels.

A.5 Aesthetic Attribute Distribution

Figures 8 and 9 show the distributions of aesthetic attribute annotations in AADB and PARA, respectively. In AADB, although annotations are provided as continuous values, several attributes such as *MotionBlur*, *Repetition*, and *Symmetry* take on only a limited number of distinct values. For these attributes, the most frequent value accounts for more than half of the annotations. When interpreting the probing results, this characteristic should be taken into account, and direct comparisons across attributes based solely on correlation values should be avoided.

Figure 10 illustrates the Spearman correlation between aesthetic attributes in the two datasets. While inter-attribute correlations are generally moderate in AADB, the aesthetic attributes annotated in PARA exhibit substantially higher correlations with each other. We hypothesize that this difference stems from the annotation protocol used in PARA. First, each attribute score is obtained by averaging subjective ratings from more than 20 annotators per image. Second, all attribute annotations are collected within a single annotation interface. These factors may encourage consistent scoring patterns across attributes, leading to higher inter-attribute correlations.

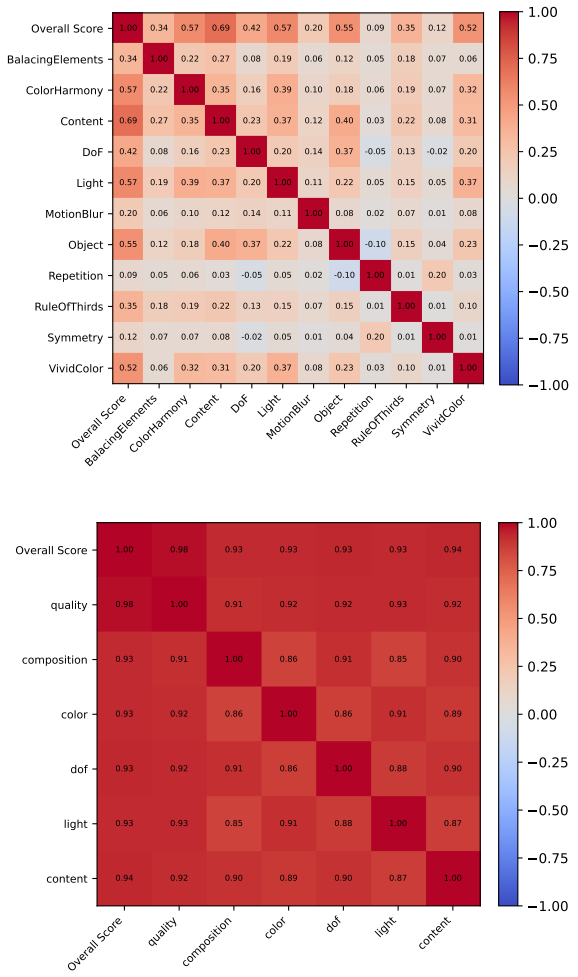

Figure 10: Spearman correlation between aesthetic attributes in AADB (top) and PARA (bottom).

Due to this strong correlation structure, the general aesthetic attributes in PARA are less suitable for analyzing the diversity of aesthetic attributes encoded in VLMs, which is the primary focus of our probing experiments. Accordingly, we adopt AADB as the primary dataset for the probing results reported in Section 3.3, and treat PARA-based

probing results as supplementary analyses.

A.6 Other Software and Artifacts

All experiments were implemented in Python 3.12.11. For VLM inference and training, we used PyTorch (Paszke et al., 2019) 2.6.0+cu126 and the Transformers (Wolf et al., 2020) library version 4.57.1. Additional libraries used in our experiments include Albumentations (Buslaev et al., 2020) 2.0.8, `peft` (Mangrulkar et al., 2022) 0.18.0, Pillow (Clark, 2015) 12.0.0, and OpenCV (Bradski, 2000) 4.11.0.86. Evaluation metrics were computed using scikit-learn (Pedregosa et al., 2011) 1.7.2 and SciPy (Virtanen et al., 2020) 1.16.3.

B Detailed Results

B.1 Full AADB Probing Results

Tables 4 and 5 summarize the results of our probing experiments on AADB using **V** and **LV** representations, respectively. While some attribute-specific differences can be observed (e.g., *VividColor* is detected more strongly with **V** than with **LT**), the overall trends are consistent across all three representation types.

We additionally examine $\mathbf{L}\tau_i$, defined as the representation of the last text token at the i-th language decoder layer. The probing results for $\mathbf{L}\tau_i$ are shown in Table 6. Consistent with the results for **V**, **LV**, and **LT**, no substantial differences are observed in the overall probing trends.

B.2 Full PIAA Experiment Results

Tables 7 and 8 present the complete results of our PIAA experiments, including both 10-shot and 100-shot settings.

Due to the considerable context length required for the **Few-shot** baseline in the 100-shot setting, we initially evaluated this configuration using the relatively lightweight Gemma 3 4B model. After confirming that the 100-shot Few-shot setting did not yield substantial performance improvements over the 10-shot setting, we did not run additional 100-shot Few-shot experiments for other models.

When comparing the 10-shot and 100-shot results, all Linear-Hidden variants consistently benefit from larger support sets. However, for the PARA dataset, the 10-shot Linear-Hidden results are lower than those of the **Raw Text** baseline. We attribute this behavior to the fact that VLMs capture GIAA aspects of photographs more effectively than those of artworks, and that annotator agreement with GIAA scores in PARA is relatively strong.

Figure 11: Example images from AADB with applied augmentations.

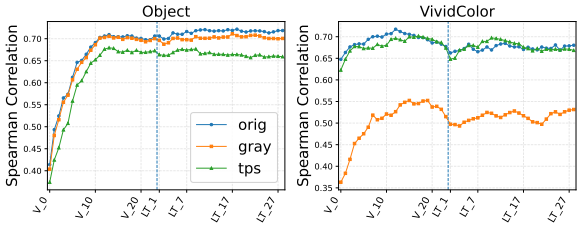

Figure 12: Probing performance under different image augmentations for Qwen3-VL 2B.

C Additional Experiments

C.1 Probing with Augmented Images

Since the aesthetic attributes used for probing on AADB exhibit non-negligible correlations, the results presented in Section 3.3 may be driven by a limited set of underlying visual traits. To examine this possibility, we conduct an additional analysis based on controlled image augmentations.

Specifically, we apply the following two augmentations to images in AADB and repeat the same probing experiments:

- **Grayscale**: Input images are converted to grayscale using Pillow (Clark, 2015), largely removing color information while preserving overall image structure.
- **Thin Plate Spline**: Thin Plate Spline transformations are applied using Albumentations (Buslaev et al., 2020) to introduce geometric distortions while approximately preserving color statistics.

Figure 12 shows probing results of Qwen3-VL 2B for the *Object* and *VividColor* attributes under different augmentations. For the *Object* attribute, performance degrades substantially under Thin Plate Spline augmentation but remains relatively stable under Grayscale conversion. In contrast, probing performance for *VividColor* drops

Attribute	Qwen3-VL			Gemma 3		DINOv3	
	2B	4B	8B	4B	12B	ViT-B/16	ViT-L/16
BalancingElements	0.325 (21)	**0.328** (23)	0.307 (26)	0.300 (26)	0.300 (26)	0.317 (11)	0.307 (22)
ColorHarmony	**0.516** (23)	0.513 (23)	0.508 (22)	0.502 (10)	0.502 (10)	0.479 (6)	0.482 (10)
Content	**0.621** (23)	0.621 (23)	0.608 (26)	0.579 (26)	0.579 (26)	0.551 (12)	0.579 (17)
DoF	0.532 (9)	0.530 (9)	**0.535** (12)	0.513 (10)	0.513 (10)	0.506 (7)	0.507 (18)
Light	0.488 (15)	**0.500** (12)	0.497 (14)	0.474 (10)	0.474 (10)	0.439 (8)	0.436 (11)
MotionBlur	0.161 (17)	0.152 (7)	**0.180** (10)	0.139 (19)	0.139 (19)	0.161 (7)	0.143 (3)
Object	0.709 (13)	**0.709** (23)	0.705 (15)	0.685 (15)	0.685 (15)	0.688 (12)	0.696 (19)
Repetition	0.458 (12)	0.461 (12)	**0.462** (26)	0.446 (15)	0.446 (15)	0.438 (8)	0.451 (19)
RuleOfThirds	**0.282** (14)	0.272 (11)	0.278 (17)	0.247 (13)	0.247 (13)	0.230 (9)	0.230 (20)
Symmetry	**0.319** (23)	0.304 (12)	0.279 (26)	0.301 (9)	0.301 (9)	0.299 (5)	0.313 (11)
VividColor	0.717 (12)	**0.719** (13)	0.718 (13)	0.698 (13)	0.698 (13)	0.686 (3)	0.685 (10)
Overall Score	0.706 (23)	**0.707** (23)	0.701 (22)	0.673 (25)	0.673 (25)	0.636 (9)	0.666 (17)

Table 4: Highest Spearman correlation achieved by linear probing on **V** layers for each aesthetic attribute in AADB. Values in parentheses indicate the corresponding layer indices.

Attribute	Qwen3-VL			Gemma 3	
	2B	4B	8B	4B	12B
BalancingElements	0.331 (7)	**0.349** (13)	0.314 (23)	0.307 (1)	0.300 (1)
ColorHarmony	0.517 (9)	**0.519** (1)	0.517 (9)	0.478 (21)	0.489 (12)
Content	**0.635** (9)	0.627 (1)	0.622 (26)	0.600 (18)	0.599 (19)
DoF	0.533 (6)	**0.534** (3)	0.516 (1)	0.471 (14)	0.478 (17)
Light	0.470 (10)	**0.478** (11)	0.474 (8)	0.423 (20)	0.428 (7)
MotionBlur	**0.158** (18)	0.147 (25)	0.132 (5)	0.118 (15)	0.137 (27)
Object	**0.716** (14)	0.712 (2)	0.710 (2)	0.695 (1)	0.692 (21)
Repetition	0.442 (2)	0.446 (3)	**0.446** (7)	0.400 (2)	0.405 (15)
RuleOfThirds	0.301 (14)	**0.303** (25)	0.285 (2)	0.257 (13)	0.247 (16)
Symmetry	**0.320** (3)	0.304 (16)	0.287 (16)	0.280 (25)	0.291 (26)
VividColor	0.710 (2)	0.714 (3)	**0.718** (8)	0.668 (4)	0.675 (2)
Overall Score	**0.729** (8)	0.721 (11)	0.718 (1)	0.685 (2)	0.694 (2)

Table 5: Highest Spearman correlation achieved by linear probing on **LV** layers for each aesthetic attribute in AADB. Values in parentheses indicate the corresponding layer indices.

significantly under Grayscale while being less affected by Thin Plate Spline.

These contrasting behaviors indicate that different aesthetic attributes rely on distinct visual traits and are differentially affected by color and geometric transformations. This observation supports our claim that the probing results in Section 3.3 reflect the presence of multiple, disentangled aesthetic attributes encoded in VLM representations, rather than being driven by a single correlated feature.

C.2 Probing on PARA

As described in Section 3.2, we also perform linear probing on the general aesthetic attributes provided in the PARA dataset. The results are summarized in Table 9 and visualized in Figure 13.

Consistent with our findings on AADB, the models strongly encode PARA aesthetic attributes, including in the language decoder layers. However, as discussed in Appendix A.5, the aesthetic attributes in PARA exhibit strong inter-attribute correlations. As a result, these probing results alone do not provide strong evidence for the presence of diverse and disentangled aesthetic attributes encoded in VLM representations.

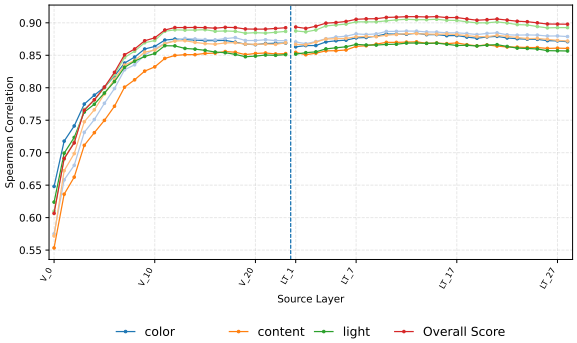

Figure 13: Layer-wise probing performance across **V** and **LT** layers for Qwen3-VL 2B on PARA.

Attribute	Qwen3-VL			Gemma 3	
	2B	4B	8B	4B	12B
BalancingElements	0.303 (5)	**0.322** (6)	0.295 (1)	0.306 (14)	0.302 (14)
ColorHarmony	0.511 (11)	**0.526** (22)	0.526 (28)	0.493 (12)	0.496 (15)
Content	**0.629** (11)	0.625 (13)	0.618 (25)	0.612 (6)	0.621 (34)
DoF	**0.531** (15)	0.520 (13)	0.521 (34)	0.500 (1)	0.524 (15)
Light	0.483 (16)	**0.510** (26)	0.485 (25)	0.442 (10)	0.469 (5)
MotionBlur	0.154 (2)	0.155 (29)	0.169 (4)	**0.174** (21)	0.161 (43)
Object	0.711 (16)	**0.719** (19)	0.717 (13)	0.705 (18)	0.711 (10)
Repetition	0.453 (7)	**0.456** (18)	0.456 (13)	0.422 (16)	0.421 (29)
RuleOfThirds	0.277 (13)	**0.294** (33)	0.268 (3)	0.264 (22)	0.257 (16)
Symmetry	0.309 (5)	**0.315** (18)	0.306 (12)	0.267 (7)	0.292 (33)
VividColor	0.680 (18)	0.692 (3)	**0.697** (21)	0.670 (7)	0.685 (18)
Overall Score	0.713 (11)	**0.726** (24)	0.717 (23)	0.693 (3)	0.714 (15)

Table 6: Highest Spearman correlation achieved by linear probing on $\mathbf{L}\tau$ layers for each aesthetic attribute in AADB. Values in parentheses indicate the corresponding layer indices.

Method	Support	Qwen3VL			Gemma 3	
		2B	4B	8B	4B	12B
		ρ / R^2	ρ / R^2	ρ / R^2	ρ / R^2	ρ / R^2
Raw Text		0.504 / -0.571	0.570 / -1.277	0.528 / -0.729	0.462 / -1.107	0.493 / -1.879
Adjust-Bias	10-shot	0.504 / -0.385	0.570 / -0.765	0.528 / -0.517	0.462 / -0.425	0.493 / -1.727
Few-shot	10-shot	0.319 / -1.850	0.197 / -1.576	0.372 / -0.547	0.241 / -0.537	0.407 / -0.185
LoRA	10-shot	0.476 / -1.867	0.581 / -1.383	0.567 / -1.206	0.503 / -1.464	0.542 / -1.231
Linear-Hidden	10-shot	0.396 / 0.069	0.401 / 0.071	0.402 / 0.085	0.402 / 0.067	0.389 / 0.044
Linear-Hidden (GIAA)	10-shot	0.446 / -0.059	0.447 / -0.112	0.456 / -0.066	0.462 / -0.092	0.444 / -0.086
Linear-Hidden (Reduce)	10-shot	0.454 / -0.049	0.466 / -0.271	0.416 / -0.257	0.460 / -0.127	0.476 / -0.067
Adjust-Bias	100-shot	0.504 / -0.310	0.570 / -0.672	0.528 / -0.441	0.462 / -0.321	0.493 / -1.562
Few-shot	100-shot	–	–	–	0.254 / -0.533	–
LoRA	100-shot	0.487 / -1.970	0.578 / -1.751	0.568 / -0.978	0.489 / -0.893	0.524 / -0.525
Linear-Hidden	100-shot	**0.604** / 0.363	**0.611** / 0.362	0.591 / **0.341**	0.591 / 0.346	0.594 / 0.329
Linear-Hidden (GIAA)	100-shot	0.596 / 0.041	0.603 / 0.057	**0.596** / 0.043	0.584 / -0.014	**0.594** / 0.036
Linear-Hidden (Reduce)	100-shot	0.585 / **0.367**	0.597 / **0.382**	0.558 / 0.322	**0.592** / **0.365**	0.593 / **0.373**

Table 7: Full PIAA results on PARA.

C.3 Prompt Sensitivity of the Probing Results

While Section 3 demonstrates that VLMs encode diverse aesthetic attributes, it remains unclear to what extent these representations are sensitive to the specific instructions provided to the model. Such sensitivity could potentially affect the validity of the probing results, as well as downstream experiments in Section 4, where different prompts are required for baselines such as **Few-shot**.

To examine this issue, we repeat the probing experiments from Section 3 on AADB using Qwen3-VL 2B with several different instruction variants:

Base The same instruction used in Section 3: *"Assess the aesthetics of this image."*

Numeric Format An instruction that explicitly enforces numeric output formatting: *" Assess the aesthetics of this image. Please rate it on a scale from 1 to 5. Output only the numeric score, and do not output any other text. "*

Attribute List An instruction that explicitly lists aesthetic attribute names to encourage attribute-aware assessment: *" Assess the aesthetics of this image with respect to the following attributes: {attrs}. You do not need to output the attributes explicitly; use them only as internal criteria. "* Here, *{attrs}* denotes a concatenation of attribute names (e.g., *BalancingElements, ColorHarmony, . . .*).

Unrelated An instruction that is unrelated to aesthetic assessment: *"Describe the weather today in one sentence."*

The results are summarized in Table 10. Across all prompt variants, the final probing performance exhibits no significant differences. Based on this observation, we conclude that the probing results are robust to reasonable variations in prompt design, and we therefore adopt flexible prompt for-

Method	Support	Qwen3VL 2B ρ / R^2	Qwen3VL 4B ρ / R^2	Qwen3VL 8B ρ / R^2	Gemma 3 4B ρ / R^2	Gemma 3 12B ρ / R^2
Raw Text		0.098 / -0.778	0.176 / -0.937	0.175 / -0.763	0.119 / -1.340	0.233 / -1.335
Adjust-Bias	10-shot	0.098 / -0.399	0.176 / -0.382	0.175 / -0.345	0.119 / -0.284	0.233 / -0.587
Few-shot	10-shot	0.142 / -1.265	0.221 / -0.380	0.264 / -0.480	0.127 / -0.354	0.227 / -0.459
LoRA	10-shot	-0.011 / -0.533	0.188 / -1.290	0.137 / -1.465	0.174 / -1.315	0.249 / -1.169
Linear-Hidden	10-shot	0.392 / 0.003	0.390 / 0.002	0.402 / 0.013	0.407 / 0.042	0.409 / 0.018
Linear-Hidden (GIAA)	10-shot	0.336 / -0.240	0.338 / -0.237	0.348 / -0.221	0.337 / -0.227	0.336 / -0.237
Linear-Hidden (Reduce)	10-shot	0.312 / -0.334	0.264 / -0.562	0.277 / -0.427	0.296 / -0.332	0.255 / -0.418
Adjust-Bias	100-shot	0.098 / -0.264	0.176 / -0.231	0.175 / -0.206	0.119 / -0.162	0.233 / -0.442
Few-shot	100-shot	–	–	–	0.093 / -0.402	–
LoRA	100-shot	0.026 / -0.701	0.153 / -1.580	0.164 / -1.386	0.116 / -0.936	0.201 / -1.022
Linear-Hidden	100-shot	**0.568 / 0.321**	**0.568 / 0.319**	**0.573 / 0.313**	**0.568 / 0.328**	**0.571 / 0.323**
Linear-Hidden (GIAA)	100-shot	0.418 / -0.148	0.420 / -0.148	0.420 / -0.151	0.413 / -0.153	0.416 / -0.155
Linear-Hidden (Reduce)	100-shot	0.480 / 0.224	0.468 / 0.202	0.459 / 0.197	0.469 / 0.220	0.446 / 0.189

Table 8: Full PIAA results on LAPIS.

Attribute	Qwen3-VL 2B	Qwen3-VL 4B	Qwen3-VL 8B	Gemma 3 4B	Gemma 3 12B	DINOv3 ViT-B/16	DINOv3 ViT-L/16
color	0.884 (13)	**0.886** (18)	0.886 (17)	0.874 (8)	0.880 (13)	0.845 (8)	0.855 (16)
composition	0.887 (12)	**0.890** (19)	0.889 (13)	0.880 (10)	0.887 (16)	0.845 (8)	0.855 (17)
content	0.871 (13)	**0.874** (17)	0.872 (17)	0.863 (9)	0.867 (13)	0.818 (8)	0.832 (17)
dof	0.883 (13)	**0.887** (19)	0.886 (19)	0.874 (9)	0.879 (15)	0.838 (8)	0.847 (17)
light	0.869 (13)	**0.876** (18)	0.875 (18)	0.861 (10)	0.866 (14)	0.828 (8)	0.839 (17)
quality	0.905 (15)	**0.910** (17)	0.909 (19)	0.893 (7)	0.899 (15)	0.853 (8)	0.863 (17)
Overall Score	0.910 (15)	**0.913** (18)	0.913 (18)	0.900 (9)	0.906 (15)	0.861 (8)	0.872 (17)

Table 9: Highest Spearman correlation achieved by probing on **LT** layers for PARA. Layer indices are shown in parentheses.

mulations in Section 4.

C.4 Effect of Image Resizing on PIAA

As described in Appendix A.4, we resize PARA images when running the **LoRA** baseline with Qwen3-VL models due to memory constraints. To validate that this design choice does not significantly affect the PIAA results, we additionally evaluate the **Raw Text** baseline using the resized images and compare its performance with that obtained on the original images.

Table 11 summarizes the results. Across all model sizes, the **Raw Text** baseline shows no substantial differences in either Spearman correlation or R^2 between the original and resized images. This observation indicates that image resizing alone does not materially affect PIAA performance for Qwen3-VL models on PARA, thereby supporting the validity of the resizing strategy adopted for the **LoRA** baseline in Section 4.

D License and Intended Use of Scientific Artifacts

All scientific artifacts used in this work, including datasets, pretrained models, and software libraries, are utilized in accordance with their respective licenses and terms of use. This study does not release new datasets or models. The experiments are conducted solely for academic research purposes, and no artifacts are used in a manner that violates their original licensing conditions.

E AI Assistance Usage

AI-assisted tools, including ChatGPT[6] and Google Gemini[7], were used to support writing refinement and code development in accordance with ACL Policy on AI Writing/Coding Assistance.

[6] https://chatgpt.com/
[7] https://gemini.google.com/

Attribute	Base ρ / R^2	Numeric Format ρ / R^2	Attribute List ρ / R^2	Unrelated ρ / R^2
BalancingElements	**0.325** (13)	0.323 (10)	0.323 (11)	0.319 (0)
ColorHarmony	0.516 (9)	0.510 (12)	0.518 (16)	**0.531** (11)
Content	0.633 (10)	**0.634** (12)	0.629 (10)	0.626 (9)
DoF	0.535 (10)	**0.541** (11)	0.531 (21)	0.535 (9)
Light	**0.509** (14)	0.505 (19)	0.508 (17)	0.491 (13)
MotionBlur	0.165 (12)	0.142 (11)	**0.176** (12)	0.167 (10)
Object	0.722 (18)	0.723 (11)	**0.727** (11)	0.726 (24)
Repetition	0.461 (3)	0.463 (3)	0.456 (4)	**0.463** (4)
RuleOfThirds	0.288 (11)	0.286 (11)	**0.295** (13)	0.288 (10)
Symmetry	0.315 (10)	0.312 (0)	0.314 (6)	**0.331** (12)
VividColor	0.686 (0)	0.688 (5)	**0.696** (23)	0.688 (28)
score	**0.725** (5)	0.721 (12)	0.724 (25)	0.718 (7)

Table 10: Probing performance on Qwen3-VL 2B under different prompt formulations.

Model	Method	ρ / R^2
Qwen3-VL 2B	Raw Text (Original)	0.504 / -0.571
	Raw Text (Resized)	0.505 / -0.511
	LoRA (Resized, 100-shot)	0.487 / -1.970
Qwen3-VL 4B	Raw Text (Original)	0.570 / -1.277
	Raw Text (Resized)	0.568 / -1.117
	LoRA (Resized, 100-shot)	0.578 / -1.751
Qwen3-VL 8B	Raw Text (Original)	0.528 / -0.729
	Raw Text (Resized)	0.532 / -0.724
	LoRA (Resized, 100-shot)	0.568 / -0.978

Table 11: Effect of image resizing on PIAA performance for Qwen3-VL models on PARA.