

VITAL: Vision-Encoder-centered Pre-training for LMMs in Visual Quality Assessment

Ziheng Jia^{1*}, Linhan Cao^{1*}, Jinliang Han¹, Zicheng Zhang², Jiaying Qian¹, Jiarui Wang¹,
 Zijian Chen¹, Guangtao Zhai^{1,2}, Xiongkuo Min^{1†}

¹Shanghai Jiaotong University, ²Shanghai Artificial Intelligence Laboratory

*Equal Contribution, †Corresponding Author

Abstract

Developing a robust visual quality assessment (VQualA) large multi-modal model (LMM) requires achieving **versatility**, **powerfulness**, and **transferability**. However, existing VQualA LMMs typically focus on a single task and rely on full-parameter fine-tuning, which makes them prone to overfitting on specific modalities or task types, thereby limiting their generalization capacity and transferability. To address this, we propose a **vision-encoder-centered generative pre-training** pipeline and develop the **VITAL-Series LMMs**. (1) We adopt a machine-executed annotation-scrutiny paradigm, constructing over 4.5M vision-language (VL) pairs—the **largest VQualA training dataset to date**. (2) We employ a multi-task training workflow that simultaneously enhances the model’s quantitative scoring precision and strengthens its capability for quality interpretation across both image and video modalities. (3) Building upon the vision encoder, we realize an **efficient model zoo extension**: the model zoo exhibits strong zero-shot performance, and each paired decoder requires only a swift warm-up using less than 1/1000 of the pre-training data to achieve performance comparable to the fully trained counterpart. Overall, our work lays a cornerstone for advancing toward the **foundation LMM for VQualA**. The project is at <https://github.com/jzhws/VITAL-Series>.

1. Introduction

With the rapid advancement of large multi-modal models (LMMs) [1, 5, 12, 26, 60, 70], the field of computer vision has undergone a transformative evolution. In parallel, leveraging the **visual-language instruction tuning** [31] to facilitate low-level perception-based vision tasks also shows significant promise [25, 28, 54, 74–76]. A typical subfield of this is visual quality assessment (VQualA). The development of a robust **VQualA LMM** depends on the **versatility**—the ability to process commonly encountered visual

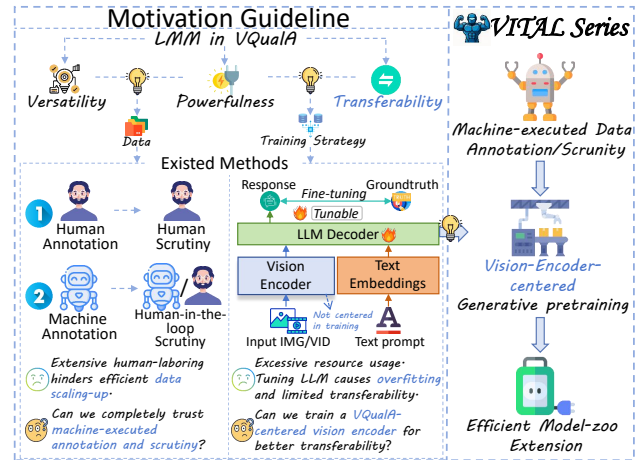


Figure 1. Comprehensive **dataset** construction underpins the versatility and performance of VQualA LMMs, while an effective **training strategy** is also essential for improving transferability and scalability. Most existing works focus on a single visual modality or task and depend heavily on human annotation, limiting scalability and versatility. In addition, fine-tuning the LLM decoder often causes overfitting, reducing transferability. These limitations serve as our motivation.

signals and to adapt to a range of tasks. Meanwhile, the model should demonstrate **powerfulness**, achieving high performance across multiple scenarios. Furthermore, as different hardware affordances necessitate models with varying parameter sizes, it is equally important—though frequently overlooked—to ensure **transferability**: the potential to efficiently perform structural adaptation.

Training data constitute the foundation that fuels the development of versatile and powerful VQualA LMMs. From the functional perspective, the diversity of training data determines the capability boundary. However, existing works are typically trained on datasets focused on a single modality or task, resulting in a restricted capability. In terms of performance, the development of a

high-performing VQualA LMM is also contingent upon the large-scale, high-quality datasets. However, the reliance on labor-intensive human annotation processes imposes significant constraints on data scalability. Consequently, current VQualA MIDBs remain limited in scale and diversity, hindering their applicability. These observations lead to a critical question: *Is it feasible to construct large-scale and task-comprehensive datasets with satisfactory quality solely through machine-executed annotation?* We believe the answer is affirmative: (1) Recent LMMs demonstrate remarkably low-level perception abilities and exhibit reduced bias to external uncertainties compared to human annotation. (2) Machine-generated annotations can still exhibit enhanced robustness when supported by effective formatting strategies and weak-to-strong generalization mechanisms.

Moreover, a **well-designed training strategy** is also essential to achieving both powerfulness and transferability. Existing LMMs for VQualA, often trained via full-parameter (with the large language model (LLM) decoder tuned) supervised fine-tuning (SFT) on a fixed structure, tend to overfit on specific data or task types, resulting in poor generalization performance. Through extensive analysis (detailed in the supplementary materials (*Supp. Sec. 8.2*)), we identify the vision encoder as the core component of VQualA-centric LMMs. Meanwhile, the pre-training process [38, 39] has been shown to facilitate inter-domain and cross-structural transferability. Consequently, we argue that adopting a vision-encoder-centered pre-training paradigm constitutes a key solution.

The above discussions (summarized in Fig. 1) motivate our work. We introduce the **Vision-Encoder-centered Pre-training for LMMs in VQualA (VITAL)**, a vital training framework built entirely upon machine-annotation. Our main contributions are in threefold:

- We construct a machine-annotated large-scale pre-training dataset comprising 4.58M vision-language (VL) pairs. The dataset covers two major tasks: **visual quality scoring** and **text generation**. We also design detailed data processing procedures to maximize the reliability of these machine-generated labels.
- We conduct the **vision-encoder-centered generative pre-training** on the tasks. For quality scoring, we employ the **proxy machine opinion distribution (PMOD) prediction** strategy to mitigate potential biases introduced by machine-produced weak labels. For text generation, we modify the loss function to dynamic *focal loss*, which balances the learning rates of simpler and more complex targets while alleviating structure overfitting. The pre-training procedure produces the **VITAL Vision Encoder**.
- This vision encoder facilitates **efficient model structural transfer**, enabling the construction of a scalable model zoo. The model zoo demonstrates strong zero-shot capabilities and, following post-training or efficient few-shot

Table 1. Statistic summary of our pre-training dataset. Note that the total counts of images and videos include the overlapping content. Each *pairwise-training* sample represents a single VL pair.

Modality	Subtask Types	# Original / Sampled	# V-L Pairs
Image	Distribution	5,000,000 / 1,000,000	1,000,000
	Pair	5,000,000 / 1,000,000	500,000
	Distortion	5,000,000 / 500,000	500,000
Video	Depict	5,000,000 / 400,000	400,000
	Distribution	4,000,000 / 845,000	845,000
	Pair	4,000,000 / 1,000,000	500,000
Video	Distortion	4,000,000 / 435,144	435,144
	Depict	4,000,000 / 400,000	400,000
Overall	# Images: 2,900,000 # Videos: 2,680,144		4,580,144

warm-up (with only 4000 samples), achieves superior performance and generalization across diverse VQualA tasks and datasets, against state-of-the-art (SOTA) baselines.

2. Related Works

Fundamental visual quality assessment can be categorized into Image Quality Assessment (IQA) [67] and Video Quality Assessment (VQA) [35]. IQA can be classified by content type, such as in-the-wild IQA [17], AI-generated content (AIGC) IQA [27], panoramic IQA [42], among others. VQA extends IQA by incorporating temporal factors, covering various types as well, such as user-generated content (UGC) [49], professionally generated content (PGC) [9, 19, 32, 61], and AI generated videos [8, 71, 73].

Studies have already leveraged LMMs to address specific VQualA tasks. *Q-Align* [58] lays the foundation of the LMM-based quality scoring using the one-hot probability estimation strategy. *Compare2Score* [77] tackles subjective label scarcity using pairwise preference as supplementary labels, and *DeQA-Score* [65] further explores leveraging soft-labeled MOS distribution to boost IQA scoring. *Q-Instruct* [57], *Aes-Expert* [18], and *VQA²* [20] pioneer in training LMMs with quality interpreting capabilities in image technical quality, image aesthetics, and video quality, respectively. *Co-Instruct* [59] and *DepictQA* [63, 64] focus on image pair quality comparison and joint analysis tasks. However, as mentioned above, these LMMs are limited to single modalities or tasks and lack a transferable strategy for efficiently extending the model zoo.

Diverse VQualA MIDBs have also been constructed for instruction tuning. These include *Q-Align-DB* [58] and *VQA²-Stage2* [20], which are reformulated from subjective labels used for quality scoring. Additionally, fully human-annotated MIDBs, such as *Q-Pathway-200K* [57], *AesMMIT* [18], and *VQA²-Stage3* [20] exemplify human annotation and revision approaches to dataset construction. Other datasets, like *OmniVQA-Chat-400K* [21] and *DepictQA-495K* [64], are based on machine-generated annotations/scutiny with human-in-the-loop. Nevertheless, the construction of these MIDBs still relies on manual labor, which prevents rapid data scaling-up. All the above issues also motivate our development of the *VITAL Series*.

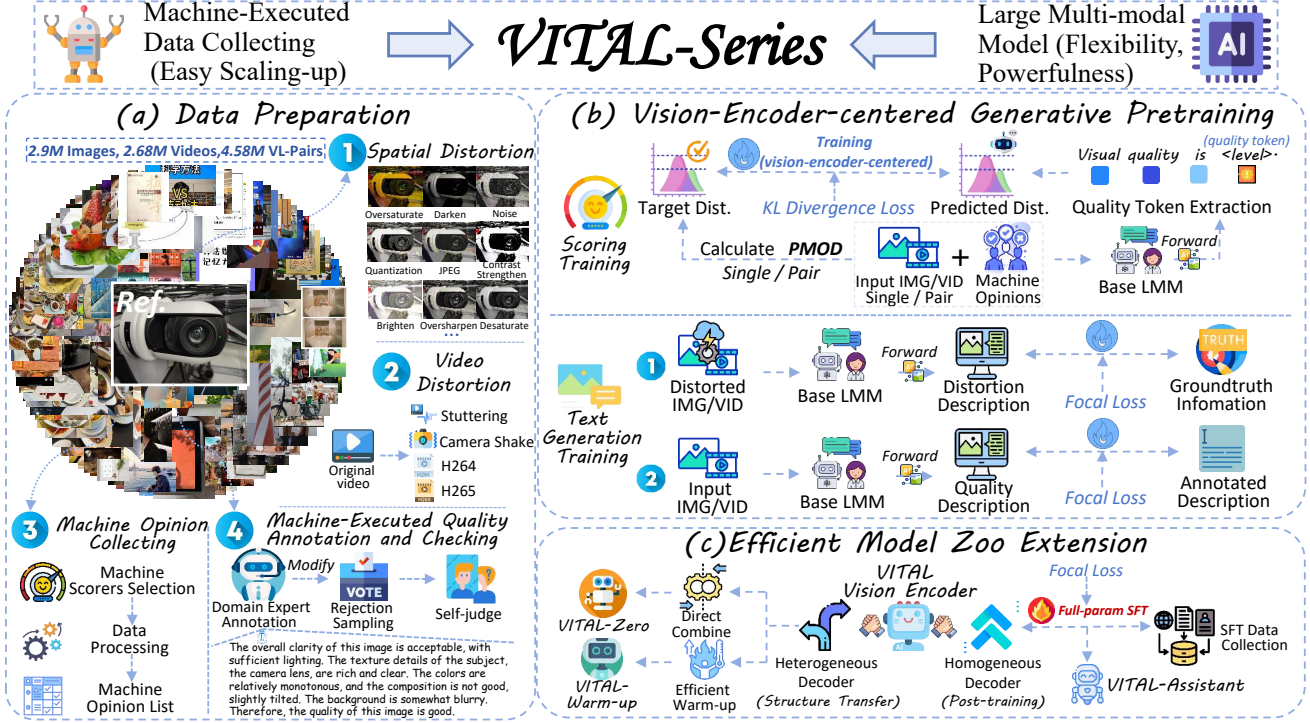


Figure 2. Overall workflow of the dataset preparation and the pretraining process of our VITAL Series.

3. The VITAL Series

3.1. Data Preparation

Our dataset is selected from a huge candidate pool containing 5M images and 4M videos. It consists of two major task types: visual quality scoring and text generation (with distortion recognition and quality interpretation subtasks), comprising 4.58M VL pairs. Statistics are shown in Tab. 1. The data preparation pipeline is provided in Fig. 2.

Candidate Pool Selection. Considering both the content diversity and the data scalability, we advocate sourcing data from in-the-wild content. We employ a minimally constrained selection strategy. Specifically, we retain only the basic controls—such as video duration (1-20s) and image resolution (no lower than 360×480 and no higher than 4K)—to ensure compliance with training efficiency. Video data are randomly sampled from the *Panda70M* [7], while image data are crawled from online platforms.

Quality Scoring Task. Earlier quality scoring datasets typically rely on subjective experiments. Multiple human experts are required to rate the designated visual signals’ quality under strictly controlled environments and standardized protocols, thereby producing the opinion scores [67]. We propose that machine annotators can substitute the procedures of human scoring, since the differences in model design and architecture among various machine annotators represent distinct **perception perspectives**, effectively mir-

roring human individuality and diversity. Meanwhile, aggregating multiple machine labels into an opinion distribution provides a more robust representation. This stems from the fact that the uncertainty of machine annotations is inherently reflected in the aggregated distribution, and estimating uncertainty can significantly enhance the LMM robustness [45]. Specifically, we select six no-reference models that have demonstrated strong performance in objective VQA and IQA tasks, respectively¹. To ensure consistency, all VQA models are pretrained on the *LSVQ (train)* [62], with the IQA ones pretrained on the *KonIQ-10K (train)* [17]. The output scores of each model are all mapped to $[0, 1]$ (with videos and images separated, detailed in *Supp. Sec. 8.5*) and discretized into 5 quality levels (*high*, *good*, *fair*, *poor*, and *low*) with an interval of 0.2.

Text Generation Task. This task is designed to enhance the LMM’s **sensitivity to visual quality cues** in text generation. Since human VQualA primarily focuses on **visual distortions**, the most perceptually decisive factors, we prioritize this aspect in our dataset construction. Specifically, we select 25 types of spatial distortions (following the methodology in *KADIS-700K* [29]) and four types of video-specific distortions. For spatial distortions, we define five severity levels: *mild*, *noticeable*, *relatively severe*, *severe*, and *very*

¹VQA: *Minimalistic-VQA-VII* [44], *Minimalistic-VQA-IX*, *FAST-VQA* [55], *DOVER* [56], *Q-Align* [58], *KVQ* [37]. IQA: *TOPIQ-NR* [6], *TReS* [15], *LIQE* [69], *ARNIQA* [3], *QualiCLIP* [2], *Q-Align*.

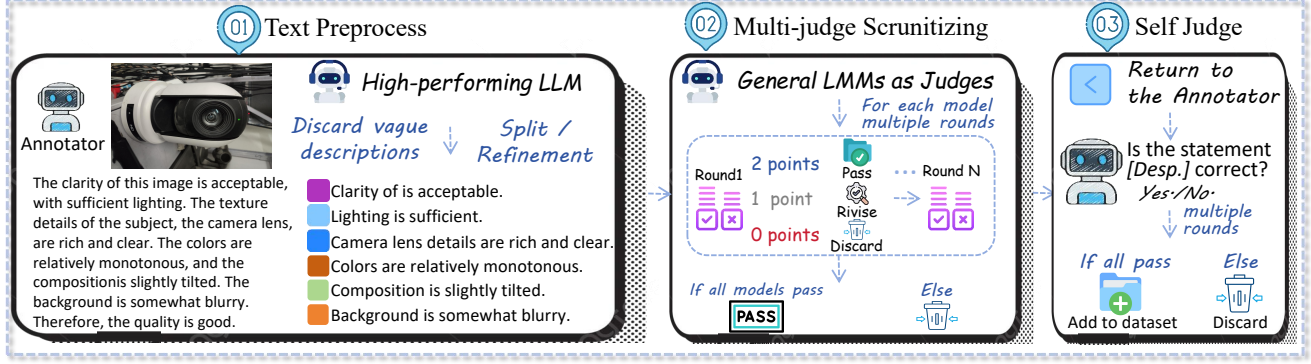


Figure 3. Machine-executed data annotation and quality scrutiny for the quality interpreting subtask.

severe. For video distortions, we set three levels: *mild*, *noticeable*, and *severe*. For each image or video in the candidate pool that satisfies the selection criteria, we randomly assign a distortion type and a severity level, which are then applied uniformly across the entire extent of the sample (detailed in *Supp. Sec. 9.3*). Each distortion is recorded in text format as “[*severity level*]-[*distortion type*]”, which is then refined to form the training-ready VL pair.

In addition to distortion recognition, the capacity to deliver a textual description of visual quality attributes is equally essential. We employ a **rejection-sampling** approach characterized by “domain-expert annotation with general LMM judgment” in the machine annotation stage. Specifically, we utilize the *VQA²-Assistant-Enhanced* [20] to generate comprehensive visual quality annotations for each sample. Subsequently, a proprietary LLM (we use the *GPT-4o-mini*) is employed to refine and paraphrase these annotations. All vague statements (e.g., “the quality of this image/video is good”) are removed during this process. The yielded descriptive paragraph is then decomposed into concise, sentence-level statements, each paired with the corresponding visual sample to form a VL pair. For each pair, we prompt high-performing LLMs (*GPT-5-250807*, *Gemini-2.5-Flash* [34], and *Qwen-VL-Max-250813* [4]) to serve as the “judges” and independently vote on their correctness for three rounds. If all judges completely agree with the match of the statement and the visual signal in all rounds, the pair is passed. Conversely, if any of the judges identifies a substantial deviation in any round, the corresponding VL pair is discarded. In cases where a judge reports minor inaccuracies or partial inconsistencies between the statement and the visual content, we adopt its revision suggestions.

Inspired by [52], we find that the annotator itself can also serve as an effective self-critical judge. Therefore, in the next step, we adopt a “**self-judge**” strategy for further data filtering. All accepted statements in the former procedure are re-evaluated by the annotator using semantically equivalent but structurally varied prompts. Each statement undergoes three rounds of consecutive binary judgments; only

those receiving consistent positive responses are retained and included in the dataset, while all others are discarded (the annotation and scrutiny process is visualized in Fig. 3 and related prompts are all provided in *Supp. Sec. 6*).

3.2. Generative Pre-training

The **vision-to-text generative pre-training** paradigm [23, 31] is a commonly adopted LMM pre-training strategy. Given that full-parameter fine-tuning is computationally expensive and often leads to overfitting, it diminishes the model’s transferability. Hence, we center the training process on the vision encoder by **freezing the LLM component and all projectors** (the workflow is shown in Fig. 2). We adopt the architecture in *VQA²* [20], wherein the **vision encoder** comprises the **image encoder** and the **motion extractor** (*SlowFast-R50* [11]). The extracted image and motion feature tokens are concatenated to form the input visual token sequence. We utilize *InternVL-3-8B-Instruct* [78] as the base model for the image encoder (*InternViT-300M-448px*), the vision projector (*MLP*), and the LLM component (*Qwen2.5-7B* [47]). The specific details of the model architecture are presented in *Supp. Sec. 7*. During image-only input cases, the *SlowFast* component is deactivated.

Quality Scoring Training. Inspired by and following *DeQA-Score* [65], we adopt the *PMOD prediction*—an effective way to leverage machine-generated scores as a form of weak supervision. Specifically, for each visual input V , we first compute the mean μ and standard deviation σ of the corresponding machine opinion list. Using these parameters, we initialize a *Gaussian* distribution $\mathcal{N}(\mu, \sigma^2)$ to represent the sample’s *PMOD*. We then perform linear adjustment to adapt the probability mass within the 5 predefined quality intervals (*high* (0.8–1.0), *good* (0.6–0.8), *fair* (0.4–0.6), *poor* (0.2–0.4), and *low* (0–0.2)), ensuring that the adjusted probabilities $p_i, i \in \{0, 1, 2, 3, 4\}$ sum to 1 and preserve the same mean value. During training, we adopt the commonly used target template “The quality of this image/video is [level]”. For all tokens preceding the [level] token (at index i_{level}), we apply the *cross-entropy* (CE) loss

to supervise text generation. For the $[level]$ token, the model outputs five logits corresponding to the predefined quality levels. These logits are converted into the predicted probabilities p_i^{pred} using the softmax processing. The *Kullback–Leibler (KL) divergence* between the predicted and target *PMOD* ($\mathcal{L}_{kl} = \sum_{i=0}^4 p_i \log(p_i/p_i^{pred})$) is then employed. Finally, the loss of the single visual signal input is designed as the weighted sum of the *CE* and the *KL* losses:

$$\mathcal{L}_{Scoring-single} = -\frac{1}{L} \left(\gamma \sum_{\ell=0}^{i_{level}-1} \log p(\mathbf{z}_\ell | \mathbf{Z}_\ell) - \mathcal{L}_{kl} \right), \quad (1)$$

where $\mathbf{z}_\ell$ represents the ℓ -th target label, $\mathbf{Z}_\ell$ denotes the input and target token sequence before $\mathbf{z}_\ell$, and L represents the length of the target token sequence. Empirically, $i_{level} = L - 1$ and γ is set to 0.01.

In subjective experiments, pairwise preferences are generally easier to obtain than absolute MOSs and are less susceptible to criteria and content biases. Analogously, training on *pairwise preference PMOD* prediction can improve the LMM’s generalization capability. Specifically, following *DeQA-Score* [65] and *Thurstone Model* [48], we treat each visual input’s *PMOD* in a pair as independent. Consequently, the preference relationship between a visual signal pair ($V_I \sim \mathcal{N}(\mu_I, \sigma_I^2)$, $V_{II} \sim \mathcal{N}(\mu_{II}, \sigma_{II}^2)$) can be modeled as the difference between two independent *Gaussian* variables $\mathcal{N}(\mu_I - \mu_{II}, \sigma_I^2 + \sigma_{II}^2)$. Consequently the *pairwise PMOD* of the V_I surpassing V_{II} in visual quality can be modeled as:

$$p^{pred}(I > II) = \Phi \left(\frac{\mu_I^{pred} - \mu_{II}^{pred}}{\sqrt{(\sigma_I^{pred})^2 + (\sigma_{II}^{pred})^2}} \right), \quad (2)$$

where $\Phi(\dots)$ represents the standard *Gaussian*. To stabilize training, we manually set all the σ^{pred} as 1. As the model outputs 5 predicted quality level logits at index i_{level} , the discrete *pairwise PMOD* prediction can be derived from this.

When constructing the label for such preference pairs, we also account for **tie cases** in which the two visual signals exhibit similar quality (we set the ratio of *better*, *worse*, and *tie* samples to 4 : 4 : 2 when preparing the dataset):

$$p^{true}(I > II) = \begin{cases} 1 & \text{if } \mu_I^{true} - \mu_{II}^{true} > \sqrt{(\sigma_I)^2 + (\sigma_{II})^2} \\ 0 & \text{if } \mu_{II}^{true} - \mu_I^{true} > \sqrt{(\sigma_I)^2 + (\sigma_{II})^2} \\ 0.5 & \text{else (tie cases).} \end{cases} \quad (3)$$

Thus, we define the predicted *pairwise PMOD* as $P_{pred} = [p^{pred}(I > II), 1 - p^{pred}(I > II)]$ and the label *pairwise PMOD* as $P_{true} = [p^{true}(I > II), 1 - p^{true}(I > II)]$. Since the prefix preceding the quality token can be easily fitted (with Eq. (??)), the *CE* loss is unnecessary during pairwise training. Therefore, we employ only the *KL divergence* between the predicted and label *pairwise PMOD* as the training loss:

$$\mathcal{L}_{PMOD-pair} = \mathcal{D}_{KL}(P_{true} \| P_{pred}). \quad (4)$$

Text Generation Training. The text generation training consists of two subtasks: distortion identification and visual quality description. The training corpus comprises statements exhibiting diverse lengths and linguistic complexity. For relatively short and semantically simple descriptions, such as those specifying particular distortion types and severity, the model tends to converge more readily, reflected by a rapid increase in output token probabilities. In contrast, longer and semantically richer statements pose greater representational challenges, causing the model to exhibit slower or less pronounced probability gains across target tokens. Under such conditions, the model tends to overfit toward shorter and simpler output forms. To handle this issue, we adopt the *focal loss* [30] that dynamically adjusts the learning rate for each output token based on its instantaneous target token output probability. This effectively mitigates the imbalance between simple and complex VL pairs by emphasizing tokens that are harder to predict and down-weighting those that are already well fitted:

$$\mathcal{L}_{Interp} = -\frac{1}{L} \left(\sum_{\ell=0}^{L-1} \alpha (1 - p(\mathbf{z}_\ell | \mathbf{Z}_\ell))^\beta \log p(\mathbf{z}_\ell | \mathbf{Z}_\ell) \right), \quad (5)$$

where $\alpha = 1$ and $\beta = 2$.

Pre-training Details. All the training data is randomly mixed. We set the EPOCH-NUM to 1 and BATCH-SIZE-PER-DEVICE to 2. The training requires approximately 1920 GPU hours using 8-GPU H200. After training, we refer to the full LMM—comprising the *VITAL Vision Encoder* and the other frozen part (the **homogeneous decoder**)—as *VITAL-Base-8B*. Moreover, to mitigate potential biases in model outputs induced by identical textual guidance, we employ the **prompt disentanglement** strategy. Specifically, we rely exclusively on the vision token sequence as the input tokens, excluding any textual guides. This setting enables the LMM to directly evoke visual quality understanding through the vision tokens themselves, avoiding overfitting to frequently encountered textual prefixes.

3.3. Efficient Model Zoo Extension

General LMMs typically undergo extensive pre-training, during which they acquire and share rich VL priors. This representational foundation provides the practical basis for **efficient model zoo extension** (in Fig. 2). Here, we utilize the *VITAL Vision Encoder* as the base socket and integrate it with diverse decoders to construct a model zoo encompassing variations in functionality and structure.

For the *VITAL-Base-8B* with the **homogeneous decoder**, we further enhance it via post-training to improve its comprehensiveness in visual quality interpreting. To this end, we aggregate a collection of publicly available MIDBs (including *Q-Pathway-200K*, *AesMMIT-400K*, *VQA²-Stage3-DB-115K*, and *OmniVQA-Chat-400K* with a total of 1120K

Table 2. Performance on the video quality scoring task. Datasets marked in *italics* denote *OOD* datasets. *Mini-VQA* refers to *Minimalistic-VQA*. *DNN-Based* models and *In-domain LMMs* are trained on human-labeled *LSVQ (train)*. *General LMMs* are used as the reference models of our *VITAL-Series* with zero-shot inference. [Per column: highest in **red**, second in **blue**, third in **boldface**.]

Datasets	LSVQ(1080p)		LSVQ(test)		LIVE-VQC		KoNViD-1K		YT-UGC		YT-Gaming		CGVDS		KVQ		AVG.↑
# of videos	3,573		7,182		585		1,200		1,098		600		357		2,926		
Models	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	
DNN-Based																	
FAST-VQA (TPAMI 2023) [55]	0.765	0.793	0.880	0.871	0.830	0.822	0.869	0.870	0.725	0.742	0.631	0.677	0.725	0.747	0.518	0.526	0.749
Mini-VQA-VII (TPAMI 2024) [44]	0.740	0.784	0.861	0.859	0.757	0.813	0.843	0.841	0.775	0.779	0.596	0.682	0.681	0.733	0.604	0.659	0.750
DOVER (ICCV 2023) [56]	0.797	0.821	0.893	0.892	0.835	0.857	0.885	0.879	0.801	0.814	0.647	0.728	0.694	0.747	0.559	0.593	0.778
KVQ (CVPR 2025) [37]	0.814	0.846	0.896	0.897	0.820	0.843	0.890	0.892	0.829	0.818	0.625	0.632	0.754	0.792	0.559	0.576	0.780
In-domain LMMs																	
Q-ALIGN (ICML 2024) [58]	0.797	0.830	0.883	0.882	0.773	0.829	0.865	0.877	0.831	0.847	0.611	0.681	0.756	0.798	0.555	0.606	0.776
VQA-UGC-SCORER (MM 2025) [20]	0.794	0.821	0.878	0.872	0.785	0.830	0.881	0.880	0.811	0.823	0.613	0.698	0.656	0.741	0.678	0.689	0.778
INTERVL3-SFT-8B-LSVQ [78]	0.749	0.777	0.849	0.839	0.743	0.790	0.846	0.843	0.798	0.800	0.632	0.693	0.801	0.808	0.560	0.592	0.758
General LMMs (reference)																	
INTERVL3-INSTRUCT-1B	0.437	0.477	0.352	0.382	0.422	0.476	0.516	0.538	0.480	0.511	0.276	0.282	0.338	0.388	0.285	0.312	0.405
INTERVL3-INSTRUCT-2B	0.468	0.492	0.463	0.468	0.418	0.434	0.575	0.563	0.510	0.495	0.301	0.291	0.114	0.149	0.287	0.297	0.395
INTERVL3-INSTRUCT-8B	0.523	0.453	0.571	0.485	0.411	0.363	0.512	0.453	0.595	0.520	0.293	0.294	0.263	0.145	0.278	0.257	0.401
INTERVL3-INSTRUCT-14B	0.499	0.562	0.602	0.582	0.374	0.413	0.580	0.587	0.596	0.601	0.241	0.268	0.293	0.361	0.365	0.335	0.454
VITAL-Series																	
VITAL-BASE-8B	0.786	0.815	0.883	0.879	0.800	0.843	0.878	0.881	0.854	0.856	0.710	0.764	0.830	0.854	0.721	0.766	0.820
VITAL-ZERO-1B	0.655	0.636	0.715	0.709	0.665	0.670	0.672	0.680	0.643	0.637	0.481	0.503	0.679	0.674	0.428	0.330	0.611
VITAL-ZERO-2B	0.660	0.686	0.765	0.750	0.688	0.713	0.783	0.747	0.676	0.655	0.582	0.615	0.759	0.693	0.562	0.595	0.683
VITAL-ZERO-14B	0.663	0.689	0.802	0.732	0.719	0.729	0.782	0.717	0.790	0.732	0.567	0.604	0.766	0.761	0.553	0.505	0.694
VITAL-WARM-UP-1B	0.786	0.819	0.869	0.866	0.771	0.812	0.873	0.875	0.846	0.848	0.705	0.754	0.808	0.824	0.730	0.743	0.808
VITAL-WARM-UP-2B	0.787	0.817	0.869	0.866	0.768	0.816	0.874	0.878	0.843	0.844	0.700	0.750	0.807	0.819	0.654	0.669	0.798
VITAL-WARM-UP-14B	0.787	0.806	0.870	0.865	0.773	0.819	0.875	0.877	0.843	0.841	0.688	0.722	0.804	0.815	0.740	0.765	0.795

Table 3. Performance on the image quality scoring task. Datasets marked in *italics* denote *OOD* datasets. *DNN-Based* models and *In-domain LMMs* are trained on human-labeled *KoniQ (train)*.

Datasets	KonIQ		SPAQ		LIVE-C		AGIQA		KADID		TID		CSIQ		AVG.↑
# of images	2,010		2,224		1,169		2,982		2,000		3,000		866		
Models	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	
DNN-Based															
NIMA (TIP 2018) [46]	0.859	0.896	0.856	0.838	0.771	0.814	0.654	0.715	0.535	0.532	0.381	0.426	0.604	0.607	0.678
DBCNN (TCSVT 2020) [68]	0.875	0.884	0.806	0.812	0.755	0.773	0.641	0.730	0.484	0.497	0.385	0.514	0.572	0.573	0.664
HYPERIQA (CVPR 2020) [41]	0.906	0.917	0.788	0.791	0.749	0.772	0.640	0.702	0.468	0.506	0.632	0.701	0.719	0.759	0.718
MUSIQ (ICCV 2021) [22]	0.929	0.924	0.863	0.868	0.830	0.789	0.630	0.722	0.556	0.575	0.629	0.704	0.721	0.754	0.750
TRES (WACV 2022) [15]	0.916	0.926	0.860	0.861	0.771	0.805	0.623	0.716	0.468	0.492	0.393	0.508	0.572	0.573	0.677
CLIP-IQA+ (AAAI 2023) [50]	0.895	0.909	0.864	0.866	0.805	0.832	0.685	0.736	0.654	0.653	0.575	0.681	0.710	0.770	0.760
LIQE (CVPR 2023) [69]	0.928	0.912	0.833	0.846	0.870	0.830	0.708	0.772	0.662	0.667	0.658	0.703	0.808	0.782	0.784
TOPIQ (TIP 2024) [6]	0.930	0.944	0.870	0.874	0.811	0.826	0.720	0.796	0.511	0.546	0.445	0.562	0.748	0.767	0.739
In-domain LMMs															
Q-ALIGN [58]	0.940	0.941	0.887	0.886	0.860	0.853	0.735	0.772	0.684	0.674	0.568	0.671	0.737	0.785	0.785
DEQA-SCORE (CVPR 2025) [65]	0.941	0.953	0.896	0.895	0.879	0.892	0.729	0.809	0.687	0.694	0.607	0.673	0.744	0.787	0.799
INTERVL3-SFT-8B-KonIQ [78]	0.913	0.914	0.863	0.870	0.817	0.822	0.678	0.773	0.521	0.553	0.459	0.565	0.762	0.790	0.736
General LMMs (Reference)															
INTERVL3-INSTRUCT-1B	0.630	0.672	0.750	0.757	0.625	0.648	0.566	0.602	0.410	0.404	0.417	0.434	0.452	0.511	0.563
INTERVL3-INSTRUCT-2B	0.698	0.671	0.605	0.603	0.667	0.637	0.683	0.686	0.490	0.500	0.524	0.483	0.596	0.573	0.601
INTERVL3-INSTRUCT-8B	0.766	0.769	0.623	0.698	0.705	0.654	0.686	0.791	0.579	0.573	0.536	0.491	0.650	0.596	0.658
INTERVL3-INSTRUCT-14B	0.737	0.701	0.560	0.652	0.618	0.586	0.655	0.674	0.536	0.541	0.478	0.498	0.509	0.496	0.589
VITAL-Series															
VITAL-BASE-8B	0.931	0.931	0.884	0.886	0.851	0.866	0.736	0.811	0.759	0.708	0.680	0.707	0.823	0.851	0.816
VITAL-ZERO-1B	0.819	0.744	0.765	0.762	0.649	0.610	0.647	0.665	0.614	0.636	0.663	0.650	0.650	0.653	0.681
VITAL-ZERO-2B	0.897	0.855	0.862	0.856	0.810	0.783	0.725	0.738	0.625	0.630	0.662	0.648	0.764	0.781	0.760
VITAL-ZERO-14B	0.876	0.878	0.868	0.841	0.800	0.765	0.709	0.721	0.706	0.665	0.610	0.652	0.774	0.765	0.759
VITAL-WARM-UP-1B	0.918	0.917	0.871	0.877	0.851	0.859	0.716	0.783	0.674	0.659	0.674	0.667	0.787	0.793	0.789
VITAL-WARM-UP-2B	0.919	0.916	0.873	0.878	0.852	0.859	0.719	0.781	0.674	0.646	0.652	0.663	0.782	0.792	0.786
VITAL-WARM-UP-14B	0.918	0.917	0.870	0.875	0.853	0.857	0.720	0.784	0.670	0.645	0.676	0.691	0.786	0.796	0.791

instruction VL pairs) and apply full-parameter SFT using the *focal loss*, yielding the *VITAL-Assistant-8B*.

For **heterogeneous decoders** (here we choose the LLM in *InternVL-1B*, *InternVL-2B*, and *InternVL-14B* with their corresponding projectors) that are not used during pre-training, we propose **two transfer strategies**. (1) The pre-trained *VITAL Visual Encoder* is directly combined with the target decoder, producing the *VITAL-Zero* series. (2) After transferring the decoder, we sample 4000 instances from the pre-training data (preserving the original task distribution) and perform efficient decoder-only **warm-up training**, resulting in the *VITAL-Warm-up* series.

Extensive experiments justify that all models deliver satisfactory performance on their respective VQualA tasks, underscoring the strong transferability.

4. Experiments

To evaluate the performance and transferability of the *VITAL-Series*, we conduct comparison experiments across multiple key tasks. Moreover, our evaluation also highlights the discussions and ablation studies of key attributes.

4.1. Experiments Settings

For VQA-scoring, we evaluate on 8 open-sourced datasets: *LSVQ (test)* [62], *LSVQ (1080P)*, *KoNViD-1K* [16], *LIVE-VQC* [14], *YouTube-UGC* [53], *LIVE-YouTube-Gaming* [66], *CGVDS* [40], and *KVQ* [33]. These datasets collectively cover diverse scenarios. We compare against multiple SOTA methods based on deep neural network (DNN) or LMM (in Tab. 2). All compared models use

Table 4. Evaluation results on the *Qbench-video-test-single*. *Multi.* refers to the question type with four options and a single answer.

Sub-categories	Question Types			Quality Concerns				Overall↑
Models	Binary ↑	Multi. ↑	Open-ended ↑	Technical ↑	Aesthetic ↑	Temporal ↑	AIGC ↑	
Open-sourced General LLMs								
INTERNVL3.5-8B (BASE MODEL) [78]	50.84%	44.25%	33.54%	41.82%	52.37%	45.07%	37.58%	42.67%
INTERNVL3.5-8B [51]	56.23%	49.83%	38.61%	47.33%	58.77%	51.36%	38.51%	48.00%
INTERNVL3.5-14B	55.89%	48.43%	39.56%	45.22%	55.92%	51.36%	42.55%	47.78%
INTERNVL3.5-38B	56.23%	48.08%	41.14%	46.03%	58.06%	50.51%	43.17%	48.33%
QWEN3VL-4B [60]	54.21%	46.34%	45.09%	48.22%	59.95%	50.85%	36.65%	48.50%
QWEN3VL-8B	56.57%	46.34%	42.56%	47.24%	59.95%	49.49%	41.30%	48.39%
QWEN3VL-32B	64.31%	54.70%	43.51%	52.43%	64.93%	56.97%	45.03%	53.94%
Domain Specific LLMs								
VQA ² -ASSISTANT [20]	67.12%	59.93%	39.56%	55.19%	56.87%	57.99%	43.79%	55.56%
OMNIVQA-CHATTER [21]	68.69%	68.29%	44.15%	58.67%	62.56%	60.54%	53.73%	59.94%
Proprietary LLMs								
GPT-4o (24-11-20) [1]	59.26%	51.57%	47.63%	51.38%	61.37%	51.87%	49.38%	52.72%
GPT-5 (25-08-07)	56.57%	55.05%	48.73%	52.03%	66.11%	55.78%	51.55%	53.33%
GEMINI-2.5-PRO [34]	69.70%	63.07%	54.75%	61.43%	68.48%	56.97%	66.46%	62.33%
QWEN-VL-MAX (25-08-13) [4]	56.57%	55.05%	43.99%	50.49%	64.45%	50.34%	49.69%	51.67%
VITAL-Series								
VITAL-BASE-8B	61.28%	58.19%	35.76%	52.43%	53.08%	50.34%	41.93%	51.33%
VITAL-ASSISTANT-8B	72.12%	71.71%	46.35%	61.60%	65.69%	63.57%	58.65%	62.94%

publicly released weights trained on the *LSVQ* (train) (28056 samples). For IQA-scoring, we utilize 7 representative datasets: *KonIQ-10K* (test) [17], *SPAQ* [10], *KADID-10K* [29], *AGIQA-3K* [27], *TID-2013* (TID) [36], *CSIQ* [24], and *LIVE-Challenge* (LIVE-C) [13]. Similarly, we include various high-performing methods for comparison (in Tab. 3). All compared models are trained on *KonIQ-10K* (train) (7046 samples) by default. Furthermore, since our pre-training strategy follows an *opinion-unaware* (OU) paradigm requiring no human-annotated labels for training, we conduct comparison experiments on the IQA-scoring task against several recent OU approaches. All baselines are implemented according to their official open-sourced configurations to ensure experimental fairness. The results are depicted in Tab. 5. We use two widely used metrics for evaluation: the *Pearson Linear Correlation Coefficient* (PLCC) and the *Spearman Rank Correlation Coefficient* (SRCC).

For the text generation task, we evaluate on the *Qbench-video-test-single* [72], which consists of 900 questions targeting various video and distortion types with multiple quality concerns. Alongside our models, we include both recently released open-source and proprietary general and in-domain LLMs as baselines to enhance the credibility of the experimental results (demonstrated in Tab. 4).

During **inference**, we generally follow the settings of VQA² with a little modification (detailed in Supp. Sec. 8.1).

4.2. Main Results Analysis

In the scoring task, the *VITAL-Base-8B* achieves outstanding performance across all IQA and VQA datasets. It surpasses both its base model and the reference SFT model trained on human-annotated datasets (video: *LSVQ* (train); image: *KonIQ* (train)). Furthermore, on most of the *out-of-distribution* (OOD) datasets (video: *LIVE-YouTube-Gaming*, *CGVDS*, and *KVQ*; image: *KADID-10k*, *AGIQA-3K*, *TID-2013*, and *CSIQ*), *VITAL-Base-8B* demonstrates

clear superiority over the strongest baselines (VQA: *KVQ*; IQA: *DeQA-Score*). For the *VITAL-Zero* series, despite the absence of fine-tuning, the models still achieve substantial performance gains compared to their base models. This underscores the crucial role of generative pre-training in achieving robust zero-shot performance. Moreover, the results of the *VITAL-Warm-up* series demonstrate strong transferability. The *1B*, *2B*, and *14B* models achieve consistently competitive scoring performance. Notably, their *OOD* performance (in most cases) shows only marginal degradation compared with *VITAL-Base-8B*. In comparison with OU methods, both the *VITAL-Base* and *VITAL-Warm-up* models achieve markedly higher performance. The above observations indicate that generative pre-training, in the absence of human annotations, can still significantly improve the VQualA LLM’s generalization and transferability to various types of visual content.

In the text generation task, the *VITAL-Base-8B* largely preserves its instruction-following capability (through vision-encoder-centered training without tuning the LLM) and outperforms the base model. After post-training, the *VITAL-Assistant-8B* achieves superior performance compared with both the latest general and domain-specific LLMs. We also provide a series of case studies (in Supp. Sec. 11.2) focused on real-world applications to further demonstrate the functionality of the *VITAL-Assistant*.

4.3. Discussions

Linear Probe Performance. Beyond compatibility with LLM decoders, verifying whether the *VITAL Vision Encoder* can adapt to decoders with simpler architectures is also crucial to assessing its transferability. To this end, we employ the *linear probe* setup. Specifically, we extract image feature tokens from different layers of *InternViT* and motion tokens from *SlowFast*, apply token-wise average pooling to each, concatenate the averaged feature

Table 5. Comparison with *OU* IQA methods (SRCC / PLCC).

Models	KoniQ	KADID	SPAQ	CSIQ
<i>OU Models</i>				
CLIP-IQA [50]	0.642 / 0.709	0.501 / 0.520	0.735 / 0.734	0.681 / 0.716
ARNIQA [3]	0.795 / 0.832	0.725 / 0.717	0.788 / 0.797	0.629 / 0.594
QUALICLIP [2]	0.715 / 0.765	0.665 / 0.665	0.841 / 0.851	0.781 / 0.815
<i>VITAL-Series</i>				
BASE-8B	0.931 / 0.931	0.759 / 0.708	0.884 / 0.886	0.823 / 0.851
WARM-UP-1B	0.918 / 0.917	0.674 / 0.659	0.871 / 0.877	0.787 / 0.793
WARM-UP-2B	0.919 / 0.916	0.674 / 0.646	0.873 / 0.878	0.782 / 0.792
WARM-UP-14B	0.918 / 0.917	0.670 / 0.645	0.870 / 0.875	0.786 / 0.796

Table 6. Linear probe performance compared with the referring *Simple-VQA*.

Version	Tunable #	LIVE-VQC	KoNViD-1k	YT-Gaming
LINEAR-PROBE	1.61 M	0.760 / 0.786	0.853 / 0.832	0.631 / 0.716
FULL	339.26 M	0.780 / 0.808	0.870 / 0.873	0.692 / 0.735
SIMPLE-VQA [43]	86.91 M	0.749 / 0.789	0.826 / 0.820	0.656 / 0.717

Table 7. Ablation study of key training attributes.

ATTRIBUTES	LIVE-VQC	KoNViD-1k	YT-Gaming	SPAQ	KADID	CSIQ
BASE-8B	0.800 / 0.843	0.878 / 0.881	0.710 / 0.764	0.884 / 0.886	0.759 / 0.708	0.823 / 0.851
w/o PMOD	0.725 / 0.773	0.835 / 0.840	0.598 / 0.657	0.865 / 0.872	0.602 / 0.668	0.743 / 0.785
w/o PAIR	0.785 / 0.817	0.856 / 0.867	0.703 / 0.760	0.868 / 0.870	0.725 / 0.687	0.810 / 0.832
w/o TEXT	0.790 / 0.837	0.868 / 0.873	0.712 / 0.760	0.876 / 0.880	0.743 / 0.712	0.818 / 0.854

tokens, and input them into the linear probe. We fine-tune the model on *LSVQ (train)*. Consistent with the pre-training task, the linear probe outputs probabilities corresponding to five quality levels, and we use the *CE loss* as the training objective. Detailed implementations are in *Supp. Sec. 8.2*. We evaluate two training configurations: *linear probe tuning* and *full-param tuning*, and compare them against *Simple-VQA* [43] (with a similar architecture but without VQualA-specific pre-training) on the *KoNViD-1k*, *LIVE-VQC*, and *LIVE-YT-Gaming*. Results are shown in Tab. 6. We observe that even with lightweight linear probe tuning, the model surpasses the performance of the reference *Simple-VQA*, while enabling full-param tuning further enhances its performance. This demonstrates the strong adaptability of the vision encoder within lightweight DNN decoders.

Data Scaling Effects. Recording the data scaling effects is essential for optimizing training efficiency. We measure how the average scoring performance of the *VITAL-Base-8B* across the 15 datasets varies as the pre-training data volume is proportionally scaled up. We also analyze the post-training data scaling effect for *VITAL-Assistant*, referencing direct SFT on *InternVL3-8B (from-scratch)*. The curves are displayed in Fig. 4 (a) and (b). For pre-training, the performance exhibits a positive correlation with the data volume, with performance gains gradually becoming marginal in the later training period. In post-training, both SFT strategies show an initial performance drop, while the pre-training-based approach remains much more stable. Moreover, this strategy exhibits superior data efficiency compared to the *from-scratch* approach. This highlights the crucial role of text-generation pretraining in the post-training phase.

Key Training Attributes Ablation. We also conduct ablation studies to validate the contributions of various training tasks and tricks. Specifically, we report the results under different configurations: whether the *PMOD* is used (oth-

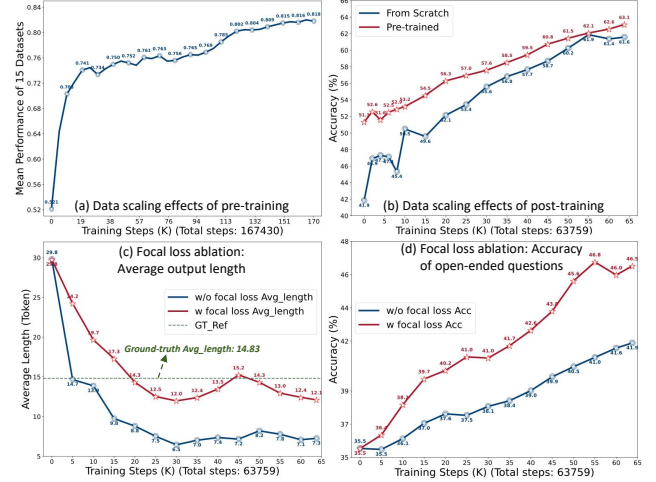


Figure 4. Plots of data scaling effects ((a) and (b)) and the focal loss ablation study ((c) and (d)).

erwise, the model is trained only with a single visual input using the quality level corresponding to the mean machine opinion with *CE loss*), whether *pair-wise training* is included, and whether the *text generation* tasks are involved. For all ablations, the training settings are kept identical to ensure fair comparison (results shown in Tab. 7). Results demonstrate that the *PMOD* prediction serves as the primary contributor to the improvement in scoring performance. In addition, the incorporation of pairwise training further calibrates the scoring, while the inclusion of the text generation tasks also provides an additional enhancement in most cases. To investigate the effect of *focal loss*, we extract the 316 open-ended questions in *QBench-video-test-single*. Fig. 4 (c) and (d) present the comparison between the post-training with *focal loss* and *CE loss* in terms of the average output length and overall accuracy. Evidently, the model trained with focal loss consistently preserves longer output lengths and higher accuracy.

5. Conclusion

We construct a large-scale machine-annotated dataset comprising 4.58M VL pairs. Building upon this, we devise the *VITAL-Series*, a suite of **versatile**, **powerful**, and **transferable** VQualA LMMs, through the **vision-encoder-centered generative pre-training**. For quality scoring, our models significantly outperform SOTA baselines on *OOD* scenarios. In text generation tasks, our post-trained model significantly outperforms various general and in-domain LMM competitors. Furthermore, the models exhibit strong structural transferability—demonstrating robust **zero-shot** and efficient **few-shot warm-up** performance when paired with diverse decoders. Overall, the *VITAL-Series* provides a new direction for developing **deployment-friendly** foundation LMMs tailored to VQualA.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62522116, Grant 62271312 and Grant 62132006, and in part by STCSM under Grant 22DZ2229005. This work was also supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0124104) in collaboration with Shanghai Artificial Intelligence Laboratory.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 7
- [2] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for real-world image quality assessment. *arXiv e-prints*, pages arXiv–2403, 2024. 3, 8
- [3] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *WACV*, pages 189–198, 2024. 3, 8
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 4, 7
- [5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1
- [6] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE TIP*, 2024. 3, 6
- [7] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pages 13320–13331, 2024. 3
- [8] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Wang Jiarui, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. *NIPS*, 37:40111–40144, 2024. 2
- [9] Shan Cheng, Huanqiang Zeng, Jing Chen, Junhui Hou, Jianqing Zhu, and Kai-Kuang Ma. Screen content video quality assessment: Subjective and objective study. *IEEE TIP*, 29: 8636–8651, 2020. 2
- [10] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, pages 3677–3686, 2020. 7
- [11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR*, pages 6202–6211, 2019. 4
- [12] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xianguyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1
- [13] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE TIP*, 25(1):372–387, 2015. 7
- [14] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. Incapture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE TCSVT*, 28(9):2061–2077, 2017. 6
- [15] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *WACV*, pages 1220–1230, 2022. 3, 6
- [16] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanzt natural video database (konvid-1k). In *QoMEX*, pages 1–6, 2017. 6
- [17] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP*, 29:4041–4056, 2020. 2, 3, 7
- [18] Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *ACM MM*, pages 5911–5920, 2024. 2
- [19] Ziheng Jia, Xiongkuo Min, and Guangtao Zhai. Dsa-qoe: Quality of experience evaluation for streaming video based on dual-stage attention. In *ISCAS*, pages 1–5. IEEE, 2024. 2
- [20] Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. Vqa2: visual question answering for video quality assessment. In *ACM MM*, pages 6751–6760, 2025. 2, 4, 6, 7
- [21] Ziheng Jia, Zicheng Zhang, Zeyu Zhang, Yingji Liang, Xiaorong Zhu, Chunyi Li, Jinliang Han, Haoning Wu, Bin Wang, Haoran Zhang, et al. Scaling-up perceptual video quality assessment. *AAAI*, 2026. 2, 7
- [22] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 6
- [23] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *CVPR*, pages 10041–10051, 2023. 4
- [24] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *JoEI*, 19(1):011006–011006, 2010. 7
- [25] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1
- [26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and

- Chunyuan Li. Llava-onevision: Easy visual task transfer. *TMLR*, 2024. 1
- [27] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE TCSVT*, 34(8):6833–6846, 2023. 2, 7
- [28] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024. 1
- [29] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *QoMEX*, pages 1–3, 2019. 3, 7
- [30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, pages 2980–2988, 2017. 5
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NIPS*, 36:34892–34916, 2023. 1, 4
- [32] Wei Lu, Wei Sun, Xiongkuo Min, Wenhan Zhu, Quan Zhou, Jun He, Qiyuan Wang, Zicheng Zhang, Tao Wang, and Guangtao Zhai. Deep neural network for blind visual quality assessment of 4k content. *IEEE TBC*, 69(2):406–421, 2022. 2
- [33] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *CVPR*, pages 25963–25973, 2024. 6
- [34] Shreshtha Basu Mallick and Logan Kilpatrick. Gemini 2.5: Updates to our family of thinking models, 2025. 4, 7
- [35] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual video quality assessment: A survey. *SCIS*, 67(11):211301, 2024. 2
- [36] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *SPIC*, 30:57–77, 2015. 7
- [37] Yunpeng Qu, Kun Yuan, Qizhi Xie, Ming Sun, Chao Zhou, and Jian Wang. Kvq: Boosting video quality assessment via saliency-guided local perception. In *CVPR*, pages 2150–2160, 2025. 3, 6
- [38] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2
- [39] 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 *ICML*, pages 8748–8763. PmlR, 2021. 2
- [40] Avinab Saha, Yu-Chih Chen, Chase Davis, Bo Qiu, Xiaoming Wang, Rahul Gowda, Ioannis Katsavounidis, and Alan C Bovik. Study of subjective and objective quality assessment of mobile cloud gaming videos. *IEEE TIP*, 32:3295–3310, 2023. 6
- [41] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, 2020. 6
- [42] Wei Sun, Xiongkuo Min, Guangtao Zhai, Ke Gu, Huiyu Duan, and Siwei Ma. Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment. *IEEE JSTSP*, 14(1):64–77, 2019. 2
- [43] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *ACM MM*, pages 856–865, 2022. 8
- [44] Wei Sun, Wen Wen, Xiongkuo Min, Long Lan, Guangtao Zhai, and Kede Ma. Analysis of video quality datasets via design of minimalistic video quality models. *IEEE TPAMI*, 2024. 3, 6
- [45] Wangtao Sun, Xiang Cheng, Xing Yu, Haotian Xu, Zhao Yang, Shizhu He, Jun Zhao, and Kang Liu. Probabilistic uncertain reward model. *arXiv preprint arXiv:2503.22480*, 2025. 3
- [46] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE TIP*, 2018. 6
- [47] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 4
- [48] Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017. 5
- [49] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE TIP*, 30:4449–4464, 2021. 2
- [50] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, pages 2555–2563, 2023. 6, 8
- [51] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internv1.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 7
- [52] Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. Llava-critic-r1: Your critic model is secretly a strong policy model. *arXiv preprint arXiv:2509.00676*, 2025. 4
- [53] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *IEEE MMSP*, pages 1–5, 2019. 6
- [54] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, pages 396–416. Springer, 2024. 1
- [55] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE TPAMI*, 2023. 3, 6
- [56] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, pages 20144–20154, 2023. 3, 6

- [57] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *CVPR*, pages 25490–25500, 2024. 2
- [58] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: teaching Imms for visual scoring via discrete text-defined levels. In *ICML*, pages 54015–54029, 2024. 2, 3, 6
- [59] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *ECCV*, pages 360–377, 2024. 2
- [60] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1, 7
- [61] Li Yang, Mai Xu, Shengxi Li, Yichen Guo, and Zulin Wang. Blind vqa on 360° video via progressively learning from pixels, frames, and video. *IEEE TIP*, 32:128–143, 2022. 2
- [62] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq: patching up the video quality problem. In *CVPR*, pages 14019–14029, 2021. 3, 6
- [63] Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Chao Dong, and Tianfan Xue. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*, 2024. 2
- [64] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *ECCV*, pages 259–276. Springer, 2024. 2
- [65] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *CVPR*, pages 14483–14494, 2025. 2, 4, 5, 6
- [66] Xiangxu Yu, Zhengzhong Tu, Zhenqiang Ying, Alan C Bovik, Neil Birkbeck, Yilin Wang, and Balu Adsumilli. Subjective quality assessment of user-generated content gaming videos. In *WACV*, pages 74–83, 2022. 6
- [67] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *SCIS*, 63(11):211301, 2020. 2, 3
- [68] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE TCSVT*, 30(1):36–47, 2020. 6
- [69] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*, pages 14071–14081, 2023. 3, 6
- [70] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1
- [71] Zhichao Zhang, Wei Sun, Xinyue Li, Yunhao Li, Qihang Ge, Jun Jia, Zicheng Zhang, Zhongpeng Ji, Fengyu Sun, Shangling Jui, et al. Human-activity agv quality assessment: A benchmark dataset and an objective evaluation metric. *arXiv preprint arXiv:2411.16619*, 2024. 2
- [72] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. Q-bench-video: Benchmarking the video quality understanding of Imms. In *CVPR*, 2025. 7
- [73] Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, et al. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In *CVPR*, pages 10621–10631, 2025. 2
- [74] Zicheng Zhang, Junying Wang, Yijin Guo, Farong Wen, Zijian Chen, Hanqing Wang, Wenzhe Li, Lu Sun, Yingjie Zhou, Jianbo Zhang, Bowen Yan, Ziheng Jia, Jiahao Xiao, Yuan Tian, Xiangyang Zhu, Kaiwei Zhang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Qi Jia, and Guangtao Zhai. Aibench: Towards trustworthy evaluation under the 45° law. *Displays*, page 103255, 2025. 1
- [75] Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, Xiangyu Zhao, Xinyu Fang, Shengyuan Ding, Ziheng Jia, Jiahao Xiao, Ye Shen, Yushuo Zheng, Xiaorong Zhu, Yalun Wu, Ziheng Jiao, Wei Sun, Zijian Chen, Kaiwei Zhang, Kang Fu, Yuqin Cao, Ming Hu, Yue Zhou, Xuemei Zhou, Juntao Cao, Wei Zhou, Jinyu Cao, Ronghui Li, Donghao Zhou, Yuan Tian, Xiangyang Zhu, Chunyi Li, Haoning Wu, Xiaohong Liu, Junjun He, Yu Zhou, Hui Liu, Lin Zhang, Zesheng Wang, Huiyu Duan, Yingjie Zhou, Xiongkuo Min, Qi Jia, Dongzhan Zhou, Wenlong Zhang, Jiezhong Cao, Xue Yang, Junzhi Yu, Songyang Zhang, Haodong Duan, and Guangtao Zhai. Large multimodal models evaluation: A survey. *SCIS*, 2025.
- [76] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 1
- [77] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *NIPS*, 37:32611–32629, 2024. 2
- [78] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 4, 6, 7