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

Supplementary Material

6. Prompts Summary

6.1. Quality Scoring Task Prompts

- **Training:** *From Human:* [image]/[video] *From GPT:* The quality of the image/video is [level].
- **Inference:** *From Human:* [image]/[video] The quality of the image/video is *From GPT:* [Predicted Quality Level]

The difference between the training and inference prompts is designed to better locate the **quality token** and ensure the accuracy of the scoring prediction.

6.2. Text Generation Task Prompts

- **Distortion Recognition Subtask**
  - **Training:** *From Human:* [image]/[video] *From GPT:* [Distortion Severity]/[Distortion Category]
- **Quality Interpreting Subtask**
  - **Training:** *From Human:* [image]/[video] *From GPT:* [Corresponding Statement]
  - **Annotator Instruction Prompts** Describe the visual quality of the video/image in detail. /Elaborate on the quality of the video/image in detail.
  - **Description Paragraph Processing Prompts:**
    - Please reformulate the paragraph into several concise sentence-level statements, split by periods. The requirements are as follows:
      - First, discard all statements about vague descriptions of visual quality (not on explicit quality attributes).
      - Simultaneously, disregard any statements that provide a conclusion assessment of the video's visual quality (e.g., "thus, the visual quality of this video is high").
    - Directly output the revised description without any prefix or suffix.
    - You should simulate as if you have derived the summary from the video itself, so do not reveal any trace of the provided description paragraph.
  - **Rejection Sampling Prompts**
    - (Each round provides identical instructions to every judge. In practice, videos and images are processed sequentially, but for convenience, all are written in a single prompt here.) Please carefully observe the given video/image and assess whether you agree with the provided evaluation statement. Rate your assessment according to the following criteria:
      - 2 points: The evaluation statement is largely consistent with the given video/image, with only minor inaccuracies or non-standard descriptions.
      - 1 point: The evaluation statement shows some deviation from the key frame sequence, primarily due to inconsistencies in temporal quality or degree of distortions or noticeable inaccurate descriptions of quality factors.
      - 0 points: The evaluation statement is largely or completely inconsistent with the observed key frame sequence; the described elements do not appear in the given video/image or exhibit significant discrepancies. Please provide your reason if you rate 0 marks.
      - If you give a score of 1 point, please give your correction of the original statement, aiming to preserve a similar sentence structure and style, while making it more accurate and academic.
- **Self-Judge Instruction Prompts**
  - (Each round of instructions has a different structure, but the meaning remains consistent. The question is designed in a multiple-choice format to ensure that the output is verifiable and standardized.)
    - Given one video/image, describe whether the visual quality assessment statement [statement] is correct.
      - A. Yes
      - B. No

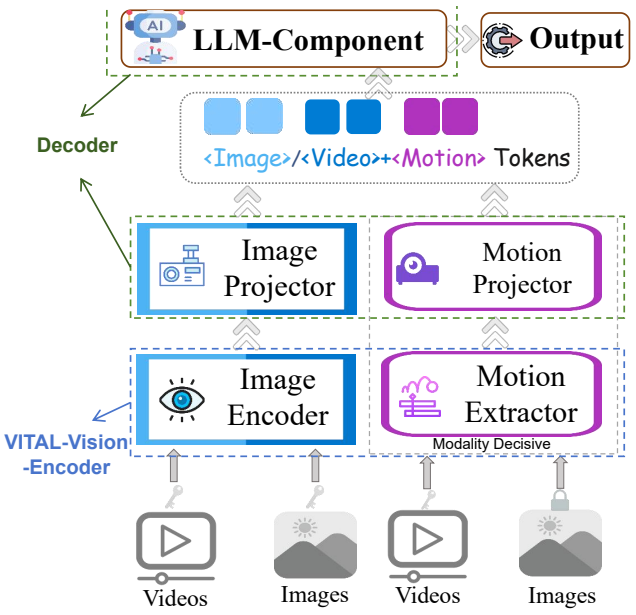

Figure 5. The model structure of the *VITAL-Series*.

7. Model Structure/Hyper-parameters

The detailed model structure and hyperparameters are shown in Tab. 8. The model structure (with encoder/decoder split) and the "prompt disentangled" input strategy are depicted in Fig. 5.

8. Additional Analysis for Main Methodology

8.1. Inference Details

Quality scoring inference details. We adopt the following procedure to assess the quality during evaluation:

$$\mathcal{Q} = \sum_{i=1}^{5} \omega_i \frac{e^{\mathcal{P}_{quality\_levels[i]}}}{\sum_{i=1}^{5} e^{\mathcal{P}_{quality\_levels[i]}}},$$

where *quality_levels* refers to a list of predefined quality levels: *[High, Good, Fair, Poor, Low]*, and $\mathcal{P}$ denotes the model's **logit outputs** for each quality level. Specifically, the vector corresponding to the quality description word in the model's output sequence is first extracted, where its dimension matches the tokenizer's vocabulary size (located at the -3 index in our model). The logit values at the specific indices of this vector, which correspond to the 5 quality level in the tokenizer's vocabulary (indices 1550, 1661, 6624, 7852, and 3347 in our model), are then selected. These logits are subsequently normalized using the softmax function.

The values ω represent the weight factors assigned to the normalized probabilities of each quality level, given by $[1, 0.75, 0.5, 0.25, 0]$. The resulting weighted sum of these probabilities produces the predicted quality score $\mathcal{Q}$, which is confined within the range of $[0, 1]$.

Quality interpreting inference details. For the quality understanding task, we use *model.generate()* with *greedy search* to ensure the reproducibility of the results. For multiple-choice questions in the benchmarks, we compare the first letter of the output (usually the selected option) with the correct answer and report the accuracy. For open-ended questions and multiple-choice questions where the first letter is not an option, we use *GPT-5-nano* for judgment (except for *GPT-4o* itself) since this is actually a textual analysis task with no need for powerful LMMs. For multiple-choice questions, we directly assess whether the answer is correct (scoring 0 or 1). For open-ended questions, we evaluate them based on three criteria: *completeness*, *accuracy*, and *relevance*, with a score of 0, 1, or 2. The specific evaluation standards are as follows:

"Given the [question], evaluate whether the response [answer] completely matches the [correct answer]. First, check the response, and please rate the score 0 if the response is not a valid answer. Please rate score 2 if the response completely or almost completely matches the correct answer on completeness, accuracy, and relevance. Please rate score 1 if the response partly matches the correct answer on completeness, accuracy, and relevance. Please rate score 0 if the response doesn't match the correct answer on completeness, accuracy, and relevance at all. Please provide the result in the following format: Score:"

We set up 5 rounds of *GPT* scoring for each question. The final score for the question is determined by "majority voting", selecting the most frequently occurring score. Based on our experiments, there has been no instance where the score distribution resulted in a "2/2/1" split.

8.2. Supplementary Experiments

The Rationality on the Vision-Encoder-centered Setting We conduct experiments to rationalize the choice of vision-encoder-centered training. First, we replace the entire pretraining process with full-parameter fine-tuning, applying the same structural transfer operations to obtain the *Zero* and *Warm-up* series (*Full-finetuning Reference*). We then compare these models with the corresponding series trained using the vision-encoder-centered approach (our setting in the main paper) across 15 datasets. The results are shown in Tabs. 9 and 10. Experimental results show that during pretraining, fine-tuning only the vision encoder or performing full-parameter fine-tuning results in nearly **identical** performance on the scoring task (comparing the **Base** models). However, after full fine-tuning, the transfer performance of the corresponding *Zero* and *Warm-up* series exhibits a noticeable decline. This highlights that training centered around the vision encoder is crucial for the pretraining task, especially for scoring tasks.

Further Proof of Model Transferability We use the same warm-up data to fine-tune the base models (*InternVL-1B, InternVL-2B, and InternVL-14B*) from scratch, which is recorded as *warm-up (reference)*. The results are also documented in Tabs. 9 and 10. Additionally, we apply more heterogeneous LLM decoders (including *Qwen2.5-7B*, *Qwen2-7B*, and *Internlm-2.5-7B*), without training on general LMMs to further test the model transferability, obtaining corresponding *Zero* models (*Additional VITAL-Zero Models*). The results are recorded in Tabs. 9 and 10. Experimental results demonstrate that the *VITAL-Warm-up* series shows significant advantages over the results obtained by fine-tuning the base model with the same warm-up data (the *warm-up reference*), particularly in video quality scoring. This further emphasizes the importance of pre-training in enhancing model structural transferability and data efficiency. The additional *VITAL-Zero* series has also shown good adaptability across a broader range of heterogeneous decoders (not only the *Qwen2.5* series). This demonstrates the model structural transferability and adaptability gained from pretraining.

Linear-Probe Experiment Details We extract features from the 6th, 12th, and 18th layers (1024 dimensions each) of the image encoder (the *InternViT*), and concatenate these

Table 8. Details of the model structure and hyperparameters for the model training. Entries without re-definition indicate that the hyperparameter remains consistent.

Model Structure/Training Hyper-Parameters	Name/Value	More Information
VITAL-Base-8B / VITAL-Assistant-8B		
Vision encoder: Image encoder *init.*	*InternViT-300M-448px*	*Parameter size*=304.01M
Vision encoder: Motion Extractor *init.*	*SlowFast-R50*	*Parameter size*=33.64M,*Use the fast-path feature*
Decoder: Image projector *init.*	*2-layers MLP+GeLU*	*Parameter size*=27.54M *(Layernorm+Linear(1024,3584)+GELU+Linear(3584,3584))*
Decoder: Motion Projector *init.*	*2-layers MLP+GeLU*	*Parameter size*=13.77M, *(Layernorm+Linear(256,3584)+GELU+Linear(3584,3584))*
Decoder: LLM *init.*	*Qwen-2.5-7B*	*parameter size*=7612.82M,Decoder-only model
Keyframes Sampling Interval (For video)	1 second	/
Video Keyframes / Image Patch Resolution	448×448	/
Token Feature Dimension (hidden size)	3584	/
Frames (for motion extraction) Resolution	448×448	/
Batch Size	16 (8 for pair-wise)	*Per device train batch size*=2 *(for pair-wise training, this is set to* 1*)*
LR Max	2e-5 / 1e-5	2e-5 in pre-traning, 1e-5 in post-training.
Gradient Accumulation Steps	2	/
Numerical Precision	`bfloat16`	/
Epoch	1	/
Eval Steps	None (no eval)	/
Optimizer	AdamW	/
Activation Checkpointing	✓	/
Deepspeed Stage	2	/
VITAL-Warm-up-1B / VITAL-Zero-1B		
Decoder: LLM *init.*	*Qwen2.5-0.5B*	/
Token Feature Dimension (hidden size)	896	/
VITAL-Warm-up-2B / VITAL-Zero-2B		
Decoder: LLM *init.*	*Qwen2.5-1.5B*	/
Token Feature Dimension (hidden size)	1586	/
VITAL-Warm-up-14B / VITAL-Zero-14B		
Decoder: LLM *init.*	*Qwen2.5-14B*	/
Token Feature Dimension (hidden size)	5120	/

features to form the image feature tokens. Next, we extract motion feature tokens (256 dimensions) from *SlowFast*. All tokens were mapped to 3584 dimensions using their corresponding projectors, after which we averaged the tokens at each position to obtain four feature tokens (the 6th, 12th, 18th layer features from *InternViT* and the *SlowFast* features). These feature tokens were concatenated and passed through a linear layer without non-linear activation to produce logits for the five quality levels. *CE loss* was used for training. During training, no validation was performed, and the model was trained for 5 full epochs using *LSVQ (train)*. The trained model was then tested directly. For the *Simple-VQA* reference model, we used its open-source weights (with *Swin-B* and *SlowFast* backbone) trained on *LSVQ (train)* for testing.

Comparison with the *VQA^2-Assistant-Enhanced* We compare the performance of *VITAL-Assistant-8B* with the original in-domain annotator (*VQA^2-Assistant-Enhanced*) for the quality interpreting task on *QBench-Video-Test-Single* (see Tab. 11). The experimental results show that *VITAL-Assistant* outperforms its original annotator.

Supplemented Experiments for Efficiency Concerns We select 10%, 30%, and 50% of the full data(with equal propotion of all tasks) for pretraining and conduct warm-up training with 1B model. The experimental results(Tab.12) show that pretraining with a further smaller scale of data still significantly aids in improving model transfer performance, which is beneficial for lower-cost, small-scale training replication for small groups.

Supplemented Experiments for Further Substantiate the Validity of PMOD In *PMOD*, we use multi-teacher, each with **different parameter sizes and model structures**. This provides **diverse perceptual perspectives**, **avoiding systematic, shared biases across all teachers** when scoring the same data. Therefore, the role of *PMOD* is: Compared to single-model(only the *Q-Align*)annotation or using point-wise training objectives(we conduct an experiment for this shown in Tab. 13), *PMOD* reduces the occurrence of systemic errors and the direct impact of potentially biased annotations when used as training objectives (through "softening" the point-wise label with proxy distribution).

Supplemented Experiments for Comparison with Enhanced Baselines We first train an enhanced version of *Qualiclip (OU)* using the full *Kadis700K* (instead of the previous 140K subset) and conduct full fine-tuning of the *Q-Align* on our complete pretraining set. Since Qualiclip is only effective for images, this model is tested solely on the IQA task. The results, shown in Tab.14), indicate that compared to the stronger OU model and the in-domain LMM

Table 9. Performance of *VITAL-Seris* models and their reference counterparts (directly "warm-up" from base models (denoted as *Warm-up Reference*) or through full-parameter pretraining (denoted as *Full-finetuning Reference*)) on the video quality scoring task. Moreover, there are additional VITAL-Zero models (with other heterogeneous decoders). Datasets marked in *italics* denote *OOD* datasets. *Mini-VQA* refers to *Minimalistic-VQA*.

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↑	
Full-finetuning Reference																	
BASE-8B-FULL	**0.794**	**0.824**	**0.878**	**0.875**	**0.805**	**0.851**	**0.874**	**0.878**	**0.851**	**0.858**	0.702	**0.758**	**0.828**	**0.857**	**0.742**	**0.758**	**0.821**
WARM-UP-1B-FULL	0.722	0.750	0.805	0.802	0.712	0.758	0.802	0.806	0.754	0.756	0.675	0.722	0.755	0.769	0.681	0.693	0.748
WARM-UP-2B-FULL	0.715	0.741	0.801	0.799	0.691	0.742	0.803	0.808	0.765	0.766	0.651	0.699	0.756	0.768	0.624	0.642	0.735
WARM-UP-14B-FULL	0.715	0.747	0.811	0.803	0.690	0.758	0.813	0.814	0.755	0.753	0.646	0.677	0.754	0.765	0.659	0.686	0.740
ZERO-1B-FULL	0.611	0.540	0.639	0.645	0.641	0.646	0.657	0.592	0.579	0.563	0.469	0.406	0.594	0.645	0.402	0.303	0.558
ZERO-2B-FULL	0.594	0.559	0.698	0.686	0.589	0.641	0.760	0.659	0.609	0.599	0.508	0.529	0.640	0.607	0.493	0.501	0.604
ZERO-14B-FULL	0.600	0.594	0.760	0.672	0.670	0.626	0.740	0.628	0.737	0.675	0.509	0.572	0.725	0.721	0.478	0.463	0.636
Warm-up Reference																	
WARM-UP-1B-REFERENCE	0.691	0.736	0.76	0.757	0.65	0.702	0.746	0.75	0.762	0.764	0.582	0.641	0.726	0.765	0.56	0.587	0.699
WARM-UP-2B-REFERENCE	0.715	0.75	0.771	0.77	0.708	0.749	0.786	0.787	0.771	0.768	0.572	0.638	0.721	0.745	0.578	0.602	0.714
WARM-UP-14B-REFERENCE	0.506	0.567	0.604	0.584	0.374	0.411	0.585	0.592	0.592	0.599	0.246	0.27	0.301	0.369	0.364	0.336	0.456
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
Additional VITAL-Zero Models																	
VITAL-ZERO(QWEN2-7B)	0.699	0.730	0.784	0.764	0.677	0.701	0.812	0.805	0.465	0.538	0.542	0.565	0.538	0.529	0.337	0.332	0.614
VITAL-ZERO(QWEN2.5-7B)	0.781	0.793	0.864	0.841	0.718	0.772	0.867	0.845	0.826	0.817	**0.708**	0.75	0.8	0.797	0.627	0.667	0.780
VITAL-ZERO(INTERNLM-7B)	0.679	0.722	0.765	0.744	0.639	0.667	0.767	0.779	0.455	0.501	0.538	0.551	0.498	0.502	0.308	0.308	0.589

Table 10. Performance of *VITAL-Series* models and their reference counterparts (directly "warm-up" from base model (denoted as *Warm-up Reference*) or through full-parameter funetuning pretraining (denoted as *Full-finetuning Reference*)) on the image quality scoring task. Moreover, there are additional VITAL-Zero models (with other heterogeneous decoders).

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↑	
Full-finetuning Reference															
BASE-8B-FULL	**0.928**	**0.933**	0.872	0.875	**0.855**	**0.871**	**0.740**	**0.815**	**0.773**	**0.719**	0.652	0.69	**0.817**	**0.838**	**0.813**
WARM-UP-1B-FULL	0.846	0.844	0.787	0.792	0.764	0.773	**0.749**	0.793	0.613	0.601	0.613	0.602	0.731	0.737	0.732
WARM-UP-2B-FULL	0.849	0.845	0.803	0.808	0.780	0.787	0.723	0.785	0.607	0.589	0.595	0.604	0.736	0.745	0.733
WARM-UP-14B-FULL	0.846	0.844	0.807	0.813	0.791	0.789	0.683	0.750	0.627	0.605	0.598	0.612	0.721	0.730	0.730
ZERO-1B-FULL	0.725	0.675	0.695	0.693	0.590	0.560	0.594	0.616	0.579	0.590	0.611	0.602	0.602	0.605	0.624
ZERO-2B-FULL	0.774	0.747	0.752	0.746	0.711	0.683	0.626	0.638	0.554	0.559	0.593	0.578	0.684	0.697	0.667
ZERO-14B-FULL	0.786	0.788	0.780	0.756	0.719	0.688	0.626	0.635	0.624	0.592	0.542	0.577	0.692	0.688	0.678
Warm-up-Reference															
WARM-UP-1B-REFERENCE	0.828	0.832	0.844	0.842	0.69	0.751	0.73	**0.811**	0.632	0.647	**0.679**	**0.703**	0.676	0.763	0.745
WARM-UP-2B-REFERENCE	0.875	0.87	0.856	0.847	0.76	0.802	**0.757**	**0.827**	0.621	0.631	0.665	0.681	0.713	0.789	0.764
WARM-UP-14B-REFERENCE	0.737	0.701	0.56	0.652	0.618	0.586	0.726	0.719	0.536	0.541	0.478	0.498	0.509	0.496	0.597
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
Additional VITAL-Zero Models															
VITAL-ZERO(QWEN2-7B)	0.876	0.889	0.831	0.838	0.811	0.814	0.733	0.794	0.675	0.622	0.613	0.66	0.769	0.786	0.765
VITAL-ZERO(QWEN2.5-7B)	**0.924**	**0.921**	**0.874**	0.875	0.851	**0.859**	0.73	0.776	**0.708**	**0.696**	0.62	0.668	**0.82**	**0.845**	**0.798**
VITAL-ZERO(INTERNLM-7B)	0.870	0.879	0.825	0.804	0.792	0.786	0.687	0.778	0.652	0.583	0.598	0.652	0.751	0.774	0.745

trained on an equivalent data scale, VITAL still demonstrates superior performance.

Ablation Study for Prompt Disentanglement(PD) *PD* is to preserve the model's instruction-following ability, facilitating post-training. We conduct an additional ablation study. If without PD, We use semantically complete sentences and images as input (e.g., in the scoring task, "How is the quality of this video/image" + image). On the QBench-video-test, the overall accuracy of the referenced **VITAL-Base** is only 25.1%. This is because the model severely overfits on fixed input patterns.

8.3. Justification on Annotator's Selection

For the scoring task, the VQA and IQA models we selected include some of the most renowned SOTA models from recent years. These models vary in parameter size and design approaches, and importantly, they are all with open-source full weights and runnable code. While other models could have been selected, we believe the models chosen are the most representative, as others are either older or structurally or conceptually similar to the current models.

For the quality description task, *VQA2-Assistant-Enhanced* is **the only** open-sourced in-domain LMM that possesses both image and video quality annotation capabilities, making the model selection reasonable.

Table 11. Evaluation results of *VITAL-Assistant* and *VQA^2-Assistant-Enhanced* on the *Qbench-video-test-single*.

Sub-categories	Question Types			Quality Concerns				Overall↑
Models	*Binary* ↑	*Multi.* ↑	*Open-ended* ↑	*Technical* ↑	*Aesthetic* ↑	*Temporal* ↑	*AIGC* ↑	
VQA^2-ASSISTANT-ENHANCED (the annotator)	69.70%	70.73%	40.66%	59.64%	58.53%	56.80%	55.59%	59.83%
VITAL-ASSISTANT-8B	**72.05%**	**72.13%**	**46.52%**	**61.75%**	**64.45%**	**61.56%**	**60.25%**	**63.11%**

Table 12. Warm-up performance with different pre-training data amounts.

Versions	LIVE-VQC	KoNViD-1k	YT-Gaming	SPAQ	KADID	CSIQ
Internvl-1B (0%)	0.422 / 0.476	0.516 / 0.538	0.276 / 0.282	0.750 / 0.757	0.410 / 0.404	0.452 / 0.511
VITAL-Warmup-1B (10%)	0.571 / 0.613	0.685 / 0.707	0.575 / 0.591	0.803 / 0.808	0.557 / 0.560	0.702 / 0.695
VITAL-Warmup-1B (30%)	0.702 / 0.745	0.812 / 0.826	0.643 / 0.682	0.814 / 0.825	0.621 / 0.616	0.726 / 0.734
VITAL-Warmup-1B (50%)	0.731 / 0.760	0.835 / 0.847	0.671 / 0.710	0.828 / 0.843	0.630 / 0.633	0.750 / 0.753
VITAL-Warmup-1B (full)	0.771 / 0.812	0.873 / 0.875	0.705 / 0.754	0.871 / 0.877	0.674 / 0.659	0.787 / 0.793

Table 13. Ablation study of different *VITAL-Base* pre-training strategies.

Strategies	LIVE-VQC	KoNViD-1k	YT-Gaming	SPAQ	KADID	CSIQ
Single-teacher (Q-Align)	0.667 / 0.702	0.772 / 0.781	0.502 / 0.577	0.781 / 0.810	0.501 / 0.556	0.645 / 0.663
Pointwise (Teacher's Mean)	0.725 / 0.773	0.835 / 0.840	0.598 / 0.657	0.865 / 0.872	0.602 / 0.668	0.743 / 0.785
PMOD (ours)	0.800 / 0.843	0.878 / 0.881	0.710 / 0.764	0.884 / 0.886	0.759 / 0.708	0.823 / 0.851

Table 14. Comparison with enhanced OU and in-domain LMM models.

Versions	LIVE-VQC	KoNViD-1k	YT-Gaming	SPAQ	KADID	CSIQ
VITAL-8B (Base)	0.800 / 0.843	0.878 / 0.881	0.710 / 0.764	0.884 / 0.886	0.759 / 0.708	0.823 / 0.851
QualiCLIP (Enhanced)	/	/	/	0.860 / 0.868	0.703 / 0.695	0.792 / 0.821
Q-Align (Enhanced)	0.780 / 0.835	0.873 / 0.876	0.673 / 0.722	0.885 / 0.880	0.712 / 0.701	0.756 / 0.810

8.4. Data Overlapping Check

Given that the pretraining process utilizes a vast amount of *in-the-wild* image/video, it is crucial to ensure that there is no serious overlap between the training and test datasets. We compare the names of each image/video in the training dataset with those in the test dataset to ensure that there are no duplicate samples.

8.5. Justification on PMOD Construction

In the machine opinion list collection, we first apply the following formula to use a nonlinear regression method to map the predicted results from each machine annotation method to the ground truth values on the *LSVQ (test)* (video) and *KonIQ (test)* (image) datasets, respectively:

$$Q'_m = \gamma_3 \,\text{Sigmoid}\left(\gamma_1 Q_m + \gamma_2\right) + \gamma_4, \tag{6}$$

where Q_m denotes the annotated raw score of a single machine annotator and Q'_m represents the mapped value. This yields the mapping parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ for each machine scorer. Then, we apply the obtained mapping parameters to the machine opinions of our own dataset. Finally, for both images and videos (separated), we combine the scores obtained from all six methods and uniformly scale them to the $[0-1]$ range.

For *pairwise PMOD* construction:

$$p^{pred}(I > II) = \Phi\left(\frac{\mu_I^{pred} - \mu_{II}^{pred}}{\sqrt{\left(\sigma_I^{pred}\right)^2 + \left(\sigma_{II}^{pred}\right)^2}}\right), \tag{7}$$

where Φ is the ***cumulative distribution function of the standard Gaussian distribution***, and the two variables μ_I^{pred} and μ_{II}^{pred} can be directly derived from the discrete probability distribution of the five quality levels in the quality token.

Specifically, $\mu_{\text{pred}} = \sum_{i=1}^{5} \omega_i \mathcal{P}_{quality\_levels[i]}$ where *quality_levels* refers to *[High, Good, Fair, Poor, Low]* and ω_i denotes the mid-value of each quality interval($[0.1, 0.3, 0.5, 0.7, 0.9]$). $\sigma_{\text{pred}} = \sum_{i=1}^{5} (\omega_i - \mu_{\text{pred}})^2 \mathcal{P}_{quality\_levels[i]}$. In Eq.2, we set σ_{pred} to 1 because when directly using the σ_{pred}, it is generally very small in the early stages of training (since the distribution has not been learned yet and the 5 level probabilities are nearly uniform). According to Eq.2, this often leads to numerical explosion and training crashes. When setting σ_{pred} to 1, Eq.2 approximates the **margin rank-loss with infinite target margin**, focusing solely on training the ranking objective. When combined with the distribution prediction in Eq.1 (to improve numerical precision), this combination is still effective.

8.6. Justification on Other Key Issues

Why is the "quality-scoring" task primarily used to validate the pretraining effect?

In our experiments validating the generalization and scalability of the *VITAL-Series*, we chose the "quality-scoring" task as the primary focus because the scoring task data comprises a significant portion of the training data (analogous to the vanilla *CLIP*, which also focuses on classification tasks, similar to its training task). Additionally, the probability distribution-based testing approach is well-suited for evaluating OOD data or the zero-shot scoring capability of the *Zero-series* models. For other quality evaluation tasks, such as text generation, since pretraining does not involve complex instruction following and the inevitable overfitting from supervised fine-tuning (SFT), directly using the *VITAL-Series* models in quality interpreting benchmark tests does not yield prominent results (though still outperforming the base model). Instead, we use data efficiency experiments to demonstrate the effectiveness of pretraining on the text generation task (in Fig. 4 in the main paper). We believe large-scale pre-training significantly enhances the model's upper bound for visual quality evaluation, but **justifying this enhancement requires designing appropriate testing methods**, such as the quality scoring task used here. For other visual quality evaluation tasks, we believe specialized downstream task training or structural transfer is required to fully demonstrate the effect of pretraining.

Why use the *InternVL-Instruct* as the pretraining base rather than training from scratch?

The *InternVL-Instruct* already possesses extensive visual question answering priors, including foundational knowl-

edge for visual quality assessment tasks. This allows us to quickly transfer model functionality without the need to "teach the model how to speak". Furthermore, the *VITAL-Zero* series benefits from the shared vision-language priors in general LLM decoders, such as *InternVL*, which is essential for enhancing the model's structural transferability. Using a general LMM as a base model is therefore both reasonable and necessary.

Where does the performance advantage of *VITAL-Series* in the quality scoring task lie?

The performance advantage of the *VITAL-Series* in the quality scoring task is primarily demonstrated in its more accurate scoring ability for OOD visual content types. For in-domain datasets like *LSVQ (test)*, *KoNViD-1K*, and *SPAQ* (which primarily consist of in-the-wild UGC visual content and authentic distortions), the model performance has nearly saturated, and the differences between different models are minimal. However, for OOD data (including synthetic distortion content, various PGC content such as CG, AIGC, HD, HFR, etc.), the *VITAL-Series* models demonstrate a clear advantage. We believe ensuring strong generalization performance is critical in current visual quality assessment tasks.

Why is there no strict feature-based filtering for the candidate pool data selection?

Our goal was to **achieve data expansion without relying on prior feature-statistic knowledge for filtering** (e.g., sampling according to widely used datasets based on statistical feature distributions). This approach ensures the successful application of large-scale in-the-wild data. While specific prior standards may improve performance on certain test sets, they can potentially harm the model's **real-world generalization** ability.

Why not use synthetic distortion data for the scoring task, and instead only include it in the text generation task?

Although synthetic distortion data is easy to generate, datasets like *KADID*, *TID*, and *CSIQ* are based on synthetic distortions, and including synthetic data in the scoring task could compromise the fairness of comparative experiments. Therefore, we avoided including synthetic data in the scoring task.

Why are the image and motion projector modules included in the decoder part rather than the *VITAL vision encoder*?

The reason for including the image and motion projectors in the decoder part rather than the vision encoder is that different sizes of LLMs correspond to different token hidden sizes (see Tab 8). This mismatch means that unfreezing the projectors (typically used for dimensional adaptation) during pretraining would result in incompatibility with LLM decoders of different parameter sizes, making direct adaptation difficult.

Why is the warm-up fine-tuning designed as decoder-only? The design of the *VITAL-Vision-Encoder* aims for zero-shot transfer to other heterogeneous decoder structures. Therefore, during the transfer process, the *VITAL-Vision-Encoder* itself remains unchanged.

Why use *QBench-Video* as the evaluation set for the text generation task instead of the earlier *QBench*? *QBench-Video* is a comprehensive evaluation set for video visual quality, which includes spatial visual quality for images and temporal-related quality for video features. It contains the two text generation pretraining subtasks: distortion identification and quality interpretation-related test questions. This makes it a more suitable benchmark for evaluating text generation tasks. In contrast, the *QBench* series, which focuses on images and was released at a much earlier date, is too simple in structure and does not provide a satisfactory level of complexity. As a result, we exclude this earlier benchmark from our evaluation.

How is *VITAL*'s generalization ability on high level tasks? Although our work primarily focuses on VQualA, due to the retention of instruction-following and the inherent ability of the base model, the model still maintains a certain-level capability for other low-level or high-level image and video analysis tasks, even without adaptation. Generalization on these tasks has not been entirely lost.

9. Additional Dataset Materials

9.1. Summary of Existing MIDBS

We conduct a summary of existing MIDBs for VQA and IQA in Tabs. 15 and 16. This clearly demonstrates the scale advantage of our dataset.

9.2. Statistic Information

The word clouds of the VL pairs in the text generation task are shown in Fig. 6. The proportion of samples for each distortion type of images and videos is illustrated in Fig. 7. The sentence lengths of the VL pairs in the quality interpreting subtask are shown in Fig. 8. The attribute metrics for the video data in the pre-training dataset are shown in Fig. 9.

9.3. Dataset Construction Details

For image data, we build our dataset through a large-scale web-scraping pipeline applied to *Baidu Image Search*. To achieve broad coverage of real-world visual scenes, we construct a query list containing more than 30 high-level keywords—such as *animals*, *landscapes*, *transportation*, etc. Each keyword is further expanded into multiple subtopics, enabling the crawler to retrieve diverse and representative visual content across various domains. All scraping procedures strictly follow ethical and legal constraints: only publicly available images with permissive usage licenses

Table 15. Summary of existing VQA MIDBs.

MIDBs for VQA	# Videos	MOS	# MOS	VL-pair	# VL-pair	Description
LIVE-VQA	160	✓	160	✗	/	Full-reference video quality rating
CVD2014	234	✓	234	✗	/	Quality assessment of video captured by cameras
LIVE-Qualcomm	208	✓	208	✗	/	Mobile in-capture video quality rating
KoNViD-1K	1,200	✓	1,200	✗	/	Unified UGC video quality rating
LIVE-VQC	585	✓	585	✗	/	Quality rating of real world UGC videos
YouTube-UGC	1,380	✓	1,380	✗	/	Quality rating of UGC videos
LSVQ	39,075	✓	39,075	✗	/	Large-scale quality rating of UGC videos
LIVE-NFLX-I	558	✓	558	✗	/	Quality-of-experience (QoE) rating of hand-craft streaming videos
LIVE-NFLX-II	420	✓	420	✗	/	QoE rating of real-world streaming videos
WaterlooSQoE-III	450	✓	450	✗	/	QoE rating of hand-craft streaming videos
LBVD	1,013	✓	1,013	✗	/	QoE assessment of in-the-wild streaming videos
WaterlooSQoE-IV	1,350	✓	1,350	✗	/	Large-scale QoE assessment of hand-craft streaming videos
TaoLive	3,762	✓	3,762	✗	/	Quality rating of live streaming (compresqsed) videos
Maxwell	4,543	✓	9,086	✗	/	Fine-grained (technical/aesthetic) quality rating of UGC videos
VQA^2-Stage-1	12,385	✗	/	✓	12,385	Pre-training MIDB for distortion recognition
VQA^2-Stage-2	30,156	✓	30,156	✓	30,156	Large-scale MIDB specially for video quality rating.
VQA^2-Stage-3	15,500	✗	/	✓	115,214	Human-annotated MIDB for video quality understanding.
OmniVQA-Chat-20K	20,000	✓	20,000	✓	20,000	Large-scale MIDB for quality rating for in-the-wild UGC videos
OmniVQA-MOS-400K	86,716	✗	/	✓	402,987	Machine-vision-dominated MIDB for video quality understanding

Table 16. Summary of existing IQA MIDBs.

MIDBs for IQA	# Images	MOS	# MOS	VL-pair	# VL-pair	Description
TID2013	3000	✓	3000	✗	/	Full-reference, synthetic distortion
KonIQ-10K	10,073	✓	10,073	✗	/	No-reference, authentic distortion
KADID-10K	10,125	✓	10,125	✗	/	No-reference, synthetic distortion, weakly-supervised
SPAQ	11,125	✓	11,125	✗	/	No-reference, authentic distortion,smartphone photos
AGIQA-3K	2,982	✓	2,982	✗	/	AIGC Images
LIVE-C	1,169	✓	1,169	✗	/	Authentic distortions, in-the-wild images, no-reference.
CSIQ	866	✓	866	✗	/	Full-reference, synthetic distortion with DMOS.
Q-Align-DB	15,800	✓	15,800	✓	15,800	Reformulating existing IQA datasets (KonIQ+SPAQ)
Q-Pathway-200K	18,973	✗	/	✓	200K	Human annotated image technical quality interpreting VL pairs.
AesMMIT	21,904	✗	/	✓	409K	Human annotated image aesthetic quality interpreting VL pairs.
DepictQA-V1	197K	✗	120,500	✓	/	Human-in-the-loop annotated image technical quality.
DepictQA-V2	140K	✗	495K	✓	/	Human-in-the-loop annotated(mainly with synthetic distortions).

are collected, and no content requiring authentication or explicit authorization is included.

For each video, we compute nine low-level quality metrics that capture fundamental visual characteristics. As illustrated in Fig. 9, the distributions of these metrics demonstrate that our dataset spans a wide and diverse range of visual quality conditions. We briefly interpret each metric as follows. *Blockiness* refers to blocking artifacts caused by aggressive compression, which is assessed by comparing luminance differences within and across block boundaries. *Blur* quantifies perceptual sharpness using the Cumulative Probability of Blur Detection, which evaluates edge-width statistics relative to human blur sensitivity. *Contrast* measures the dispersion of pixel intensities and indicates the overall dynamic range of the frame. *Noise* captures random high-frequency fluctuations by estimating the residual energy remaining after low-pass filtering. *Flickering* denotes unstable temporal variations identified by counting significant frame-to-frame luminance changes. *Colourfulness* evaluates the distribution and balance of chromatic components based on opponent-color statistics. *Luminance* measures the overall brightness level of each frame. *Spatial information (SI)* characterizes spatial complexity using the standard deviation of Sobel-filtered edge responses. *Temporal information (TI)* captures motion intensity by evaluating the variability of inter-frame differences.

We select 25 types of spatial distortions (following the methodology in *KADIS-700K*) with five severity levels and four types of video-specific distortions with three severity levels. Here we provide a brief description of each **spatial distortion** type:

Gaussian Blur. Gaussian blur smooths local image structures by convolving the image with a Gaussian kernel of standard deviation σ:

$$\tilde{I} = I * G_\sigma.$$

As σ increases, edges become softer and fine textures gradually disappear, producing a natural smoothing effect.

Lens Blur. Lens blur simulates optical defocus by filtering the image with a circular point spread function (PSF), represented by a normalized disk kernel of radius r:

$$\tilde{I} = I * D_r.$$

Figure 6. Wordclouds of the VL pairs in the text generation task.

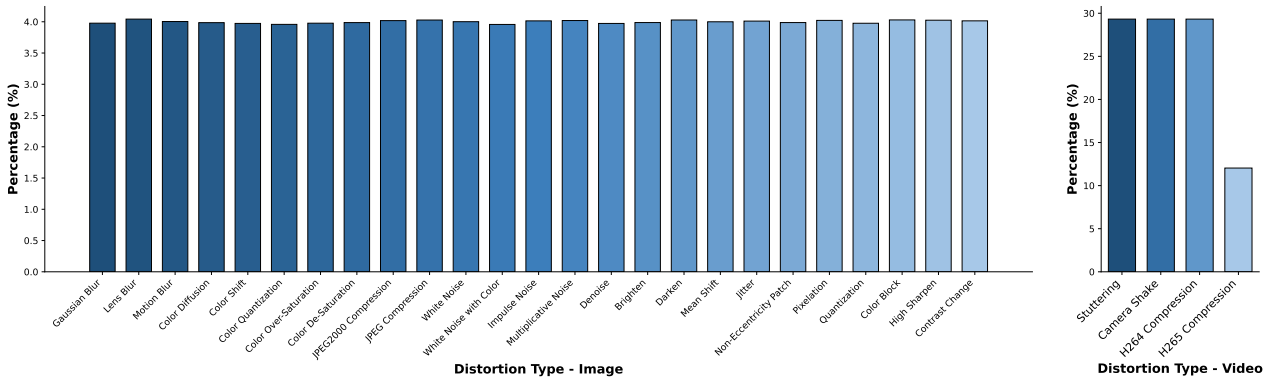

Figure 7. Statistics of distortion types.

This distortion produces blurred highlights and rounded defocus patterns similar to bokeh observed in real camera lenses.

Motion Blur. Motion blur imitates camera or object motion during exposure by convolving the image with a linear PSF of length r and direction θ:

$$\tilde{I} = I * M_{r,\theta}.$$

The resulting image exhibits directional streaks whose strength depends on the blur length and orientation.

Color Diffusion. Color diffusion selectively smooths chromatic information while preserving luminance. In Lab space, only the a and b channels are blurred:

$$\tilde{a} = a * G_\sigma, \qquad \tilde{b} = b * G_\sigma.$$

This produces a soft watercolor-like appearance with reduced color contrast but intact structural edges.

Color Shift. Color shift introduces subtle chromatic misalignment by spatially translating a single color channel (typically green) by a displacement Δ. Using a spatial blending mask M, the shifted channel becomes:

$$\tilde{G} = G(\cdot - \Delta)\, M + G(1 - M).$$

This produces mild color fringing reminiscent of chromatic aberration or sensor alignment errors.

Color Quantization. Color quantization compresses color richness by mapping each pixel to one of K representative palette entries:

$$\tilde{I}(x, y) = \mathcal{C}[X(x, y)],$$

where X is an indexed image and $\mathcal{C}$ is the learned palette. Reducing K introduces banding artifacts and coarse color transitions.

Color Over-Saturation. Over-saturation increases chromatic intensity by scaling the saturation channel in HSV space:

$$\tilde{S} = \gamma S, \quad \gamma > 1.$$

This results in vivid, exaggerated colors commonly seen in overly processed photographs.

Color De-Saturation. De-saturation reduces color strength by attenuating the chromatic channels in Lab space:

$$\tilde{a} = \beta a, \qquad \tilde{b} = \beta b, \quad 0 < \beta < 1.$$

The output appears washed-out or faded, with weakened chromatic contrast.

JPEG2000 Compression. JPEG2000 distortion arises from wavelet-based compression. Applying a compression–decompression cycle yields:

$$\tilde{I} = \mathrm{Decode}\big(\mathrm{Encode}_{\mathrm{JP2}}(I)\big).$$

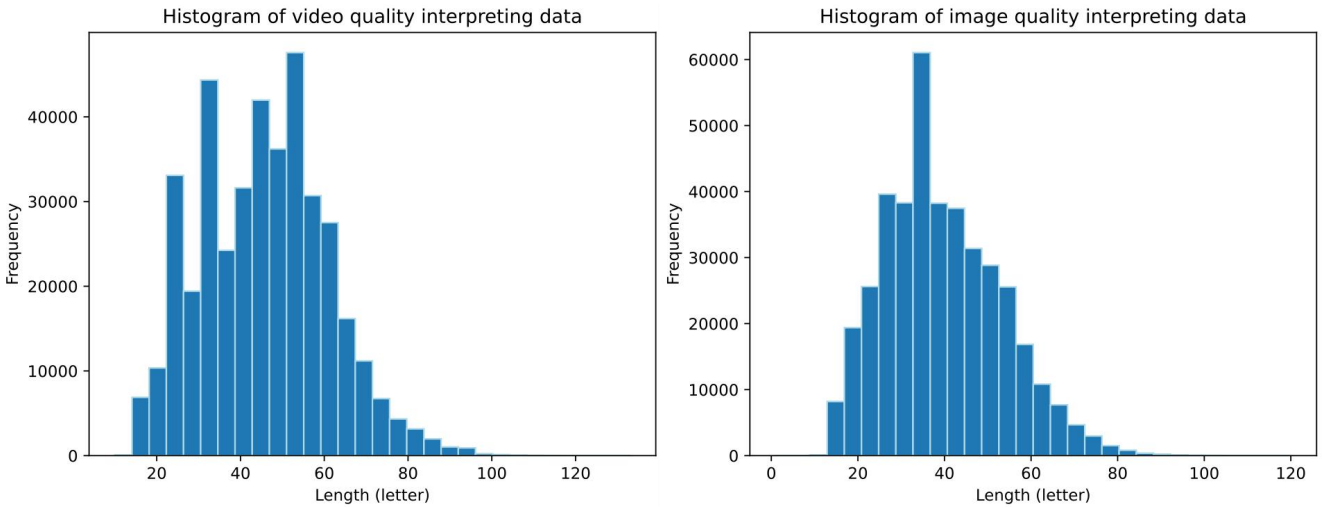

Figure 8. Sentence lengths of the VL pairs in the quality interpreting subtask.

Artifacts include edge ringing, wavelet granularity, and smooth but degraded textures.

JPEG Compression. JPEG compression produces DCT-based artifacts after quantizing blockwise frequency coefficients:

$$\tilde{I} = \mathrm{Decode}\big(\mathrm{Encode}_{\mathrm{JPEG}}(I)\big).$$

Depending on the quality level, the result contains blocking, ringing, and loss of fine detail.

White Noise. White noise adds pixelwise Gaussian noise to each color channel:

$$\tilde{I} = I + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma^2).$$

This simulates sensor noise or low-light imaging degradation.

White Noise with Color. This variant injects Gaussian noise into YCbCr channels before converting back to RGB:

$$(Y', Cb', Cr') = (Y, Cb, Cr) + (\eta_Y, \eta_{Cb}, \eta_{Cr}).$$

Because chrominance is also corrupted, the noise exhibits visible color tints.

Impulse Noise. Impulse noise replaces random pixels with extreme values:

$$\tilde{I}(x,y) = \begin{cases} 0 & \text{with probability } p_i/2, \\ 1 & \text{with probability } p_i/2, \\ I(x,y) & \text{otherwise.} \end{cases}$$

This produces classic salt-and-pepper artifacts.

Multiplicative Noise. Multiplicative (speckle) noise modulates intensity by a multiplicative Gaussian factor:

$$\tilde{I} = I(1+\epsilon), \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

This results in grainy, signal-dependent variations typical in coherent imaging systems.

Denoise. The image is first corrupted with Gaussian noise and then passed through a CNN denoiser $\mathcal{D}$:

$$\tilde{I} = \mathcal{D}(I + \eta).$$

Residual smoothing and faint artifacts remain due to imperfect restoration.

Brighten. Brightening enhances mid-tone luminance using a nonlinear tone curve:

$$\tilde{L} = \Gamma(L; \alpha_b).$$

This increases brightness while maintaining overall dynamic range.

Darken. Darkening applies the inverse tone curve:

$$\tilde{L} = \Gamma(L; \alpha_d).$$

It reduces luminance in the mid-range while preserving highlights and shadows.

Mean Shift. Mean shift adds a constant offset to all pixels:

$$\tilde{I} = \mathrm{clip}(I + \delta_m).$$

This produces a uniform global change in brightness.

Jitter. Jitter displaces pixels by random offsets and resamples the image:

$$\tilde{I}(x,y) = I(x + \Delta_x,\ y + \Delta_y).$$

The result exhibits irregular spatial jittering and local distortions.

Non-Eccentricity Patch. Small patches are relocated to nearby positions to create local inconsistencies:

$$\tilde{I}(\Omega'_k) = I(\Omega_k).$$

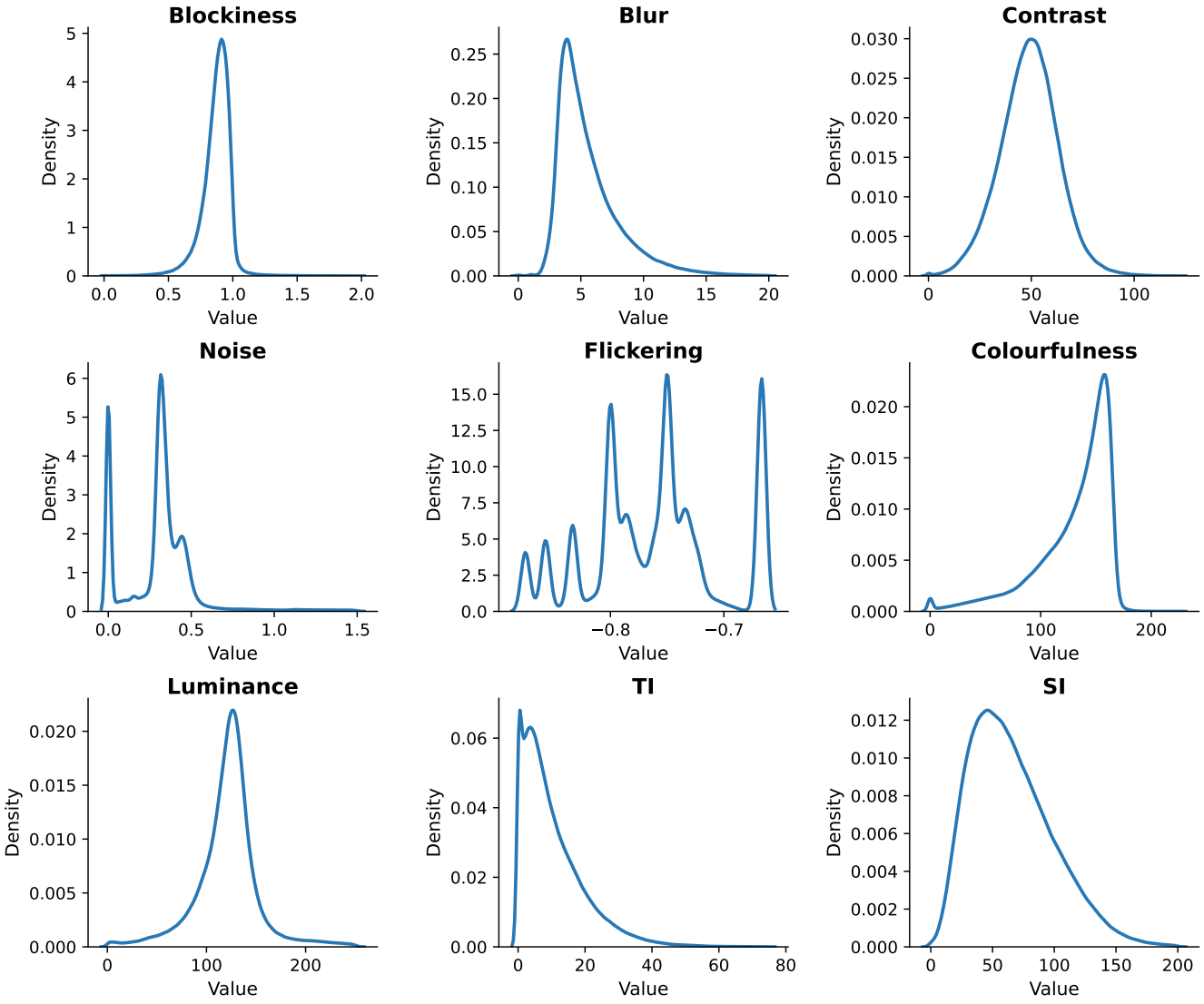

Figure 9. Statistics metrics for the video data in the pre-training dataset.

Although local geometry is altered, global structure remains mostly preserved.

Pixelation. Pixelation downscales the image and then upsamples it via nearest-neighbor interpolation:

$$\tilde{I} = \text{resize}_{nn}(I_{\downarrow}).$$

This produces blocky, coarse patterns indicative of low-resolution displays.

Quantization. Pixel intensities are mapped to discrete levels using multi-threshold partitioning:

$$\tilde{I}(x, y) = q_j \quad \text{if } I(x, y) \in [t_j, t_{j+1}).$$

This introduces tonal banding in smooth gradient regions.

Color Block. Random solid-color rectangles are inserted into the image:

$$\tilde{I}(x, y) = v_k \quad \text{for } (x, y) \in \Omega_k.$$

This simulates occlusion or block-based corruption.

High Sharpen. High sharpening applies an intensified unsharp mask:

$$\tilde{I} = \text{clip}(I + \alpha(I - B)).$$

Large α produces pronounced ringing and edge overshoots.

Contrast Change. Contrast is modified using a nonlinear tone curve:

$$\tilde{p} = \Gamma(p; \beta).$$

Increasing or decreasing β adjusts the steepness of mid-tone transitions, altering global image contrast.

Here we provide a brief description of each **video-specific distortion** type:

Stuttering. Stuttering simulates temporal freezing in videos by intermittently dropping and repeating frames. The distortion level determines the probability q_s that the current frame is substituted by a previous one. For each frame F_t, a random variable $u \sim \mathcal{U}(0,1)$ is drawn, and the output frame is computed as

$$\tilde{F}_t = \begin{cases} F_{t-1}, & \text{if } u \leq q_s, \\ F_t, & \text{otherwise.} \end{cases}$$

This mechanism introduces irregular temporal progression and discontinuities, producing a visual effect similar to playback stuttering in low-quality or unstable video streams.

Camera Shake. Camera shake simulates unintended hand-held camera motion by applying small random spatial perturbations to consecutive frames, with the shake intensity controlling the displacement magnitude. For each frame, slight horizontal and vertical offsets are introduced via an affine transformation:

$$T = \begin{bmatrix} 1 & 0 & \Delta_x \\ 0 & 1 & \Delta_y \end{bmatrix},$$

where Δ_x and Δ_y are random shifts sampled according to the shake level. This produces a frame-wise jitter characteristic of unstable hand-held video capture.

H264 Compression. H264 compression applies the widely used AVC encoding standard to simulate streaming-related video degradation. Each video is encoded using three constant rate factor (CRF) values—24, 36, and 48—via the `fast` preset in `ffmpeg`. This configuration introduces typical AVC artifacts, including blocking, blurring, texture loss, and reduced edge fidelity.

H265 Compression. H265 compression employs the HEVC standard to generate distortions representative of modern high-efficiency codecs. We apply the same CRF values (24, 36, and 48) as in H264 compression and encode all videos using the `very slow` preset in `ffmpeg`. This setting produces characteristic HEVC artifacts such as smoother textures, finer detail suppression, and mild blocking or ringing effects.

10. Limitations

As shown in Fig. 4 in the main paper, the data-scaling effects from both pretraining and post-training have not yet reached the peak, and there remains room for exploration. However, due to resource limitations, we are unable to present the results of larger-scale data training in this work. Nevertheless, this is easily extendable in future studies and is one of the primary goals of our upcoming research.

11. Applications

11.1. Potential Future Applications

The *VITAL-Vision-Encoder* can be applied to downstream tasks in more specialized visual quality evaluation scenarios, while the *VITAL-Zero* (including linear probe) and *VITAL-Warm-up* series can be applied to specific scenarios with varying model parameter size requirements.

11.2. Case Studies

Here we use *Gradio Demo* and conduct six case studies with various image and video contents and types (shown in Figs. 10, 11, 12, 13, 14, and 15).

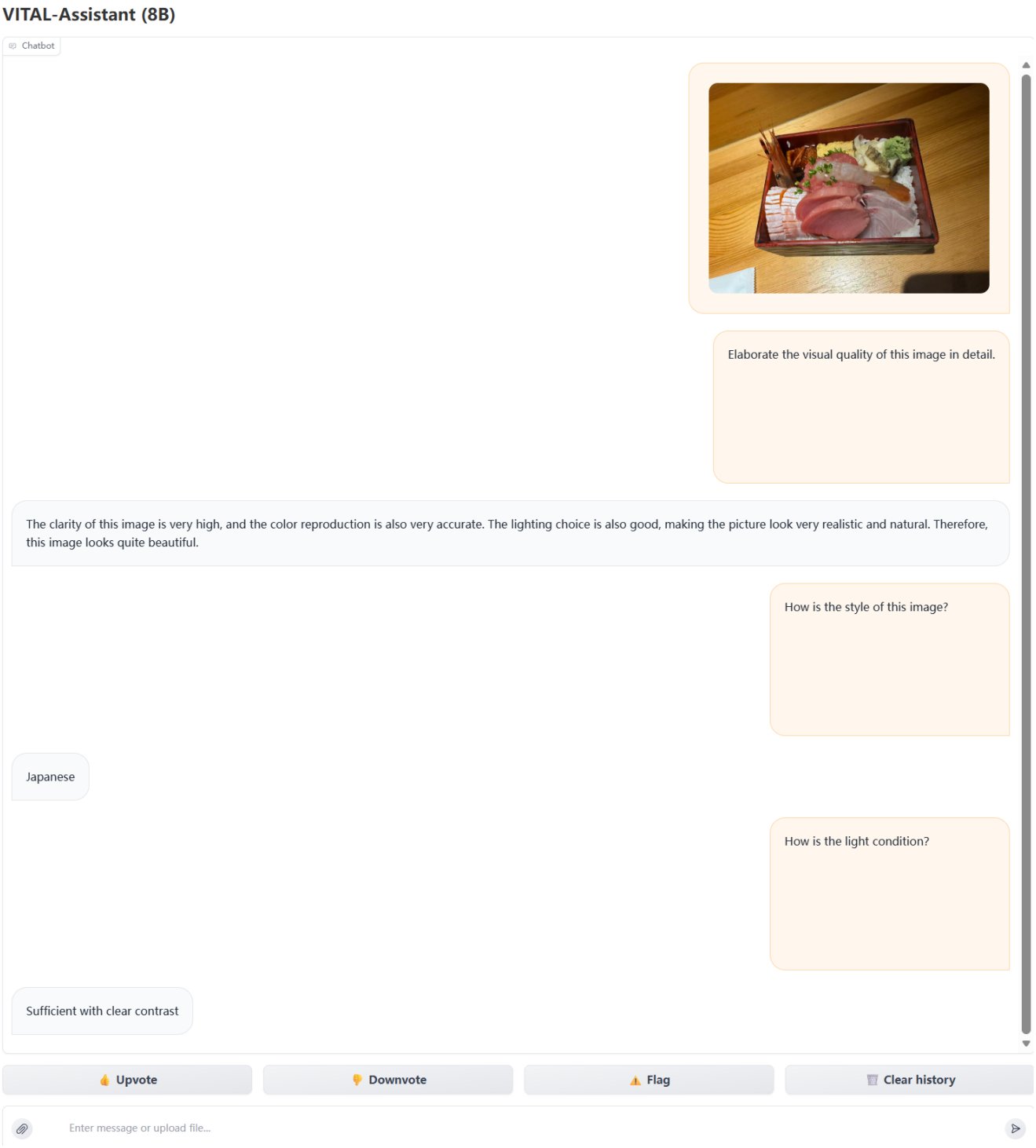

Figure 10. A photo taken by a smartphone in a Japanese restaurant.

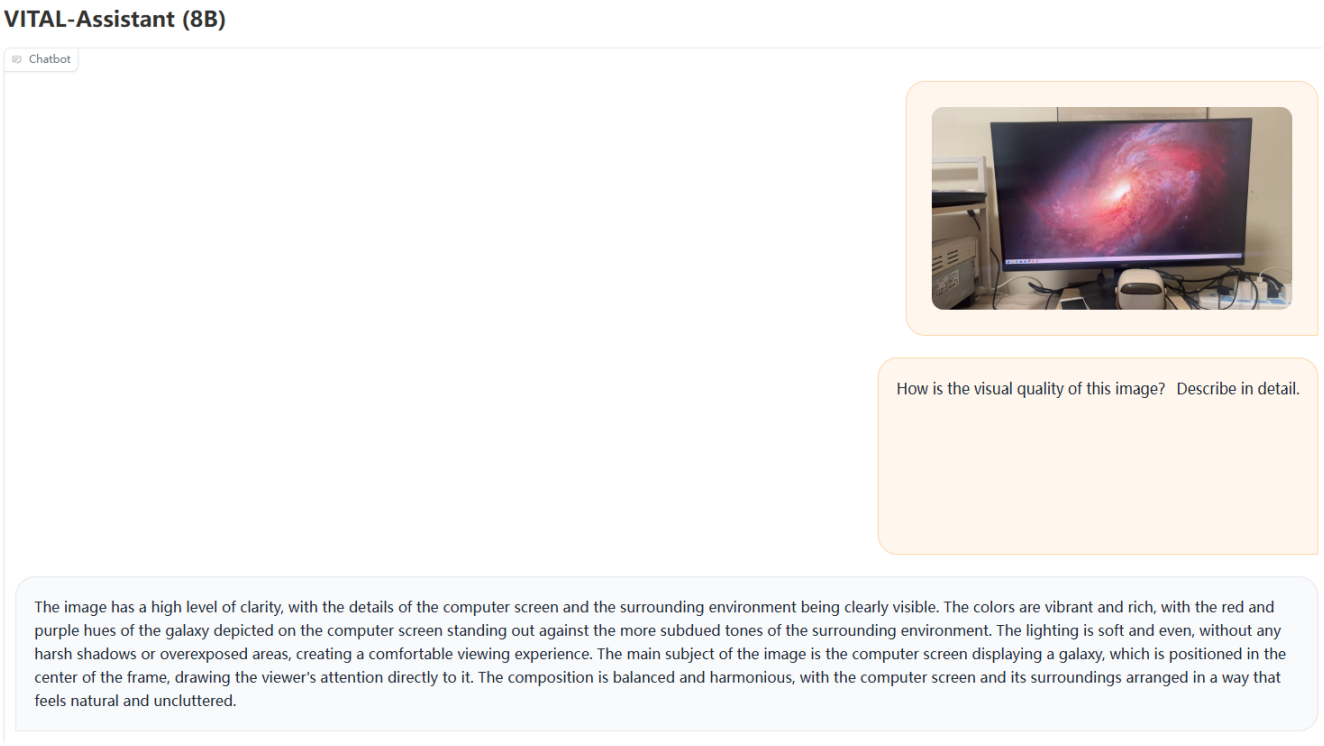

Figure 11. A photo taken by a camera in an office.

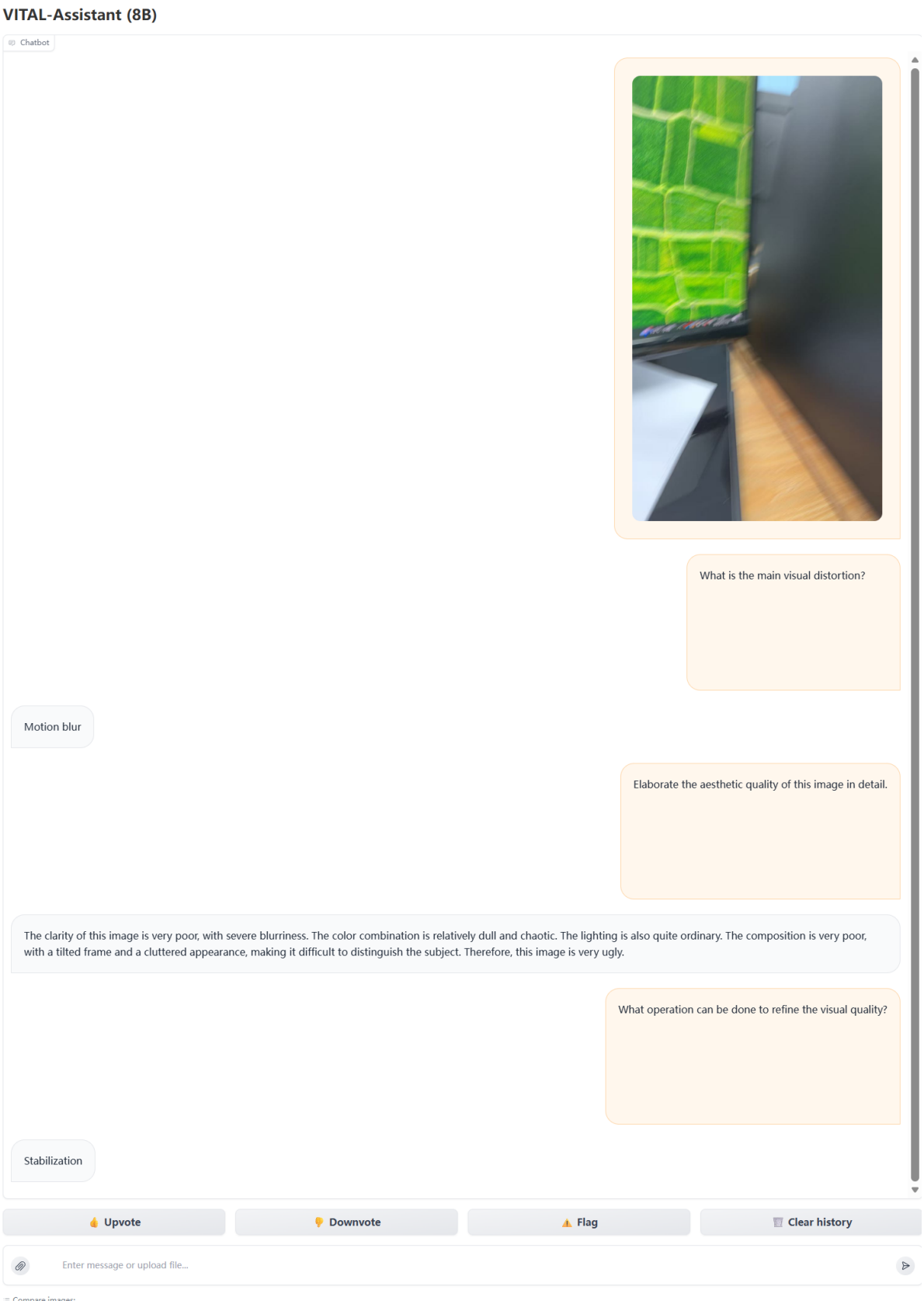

Figure 12. A photo taken by a smartphone, but without the necessary control of the visual quality.

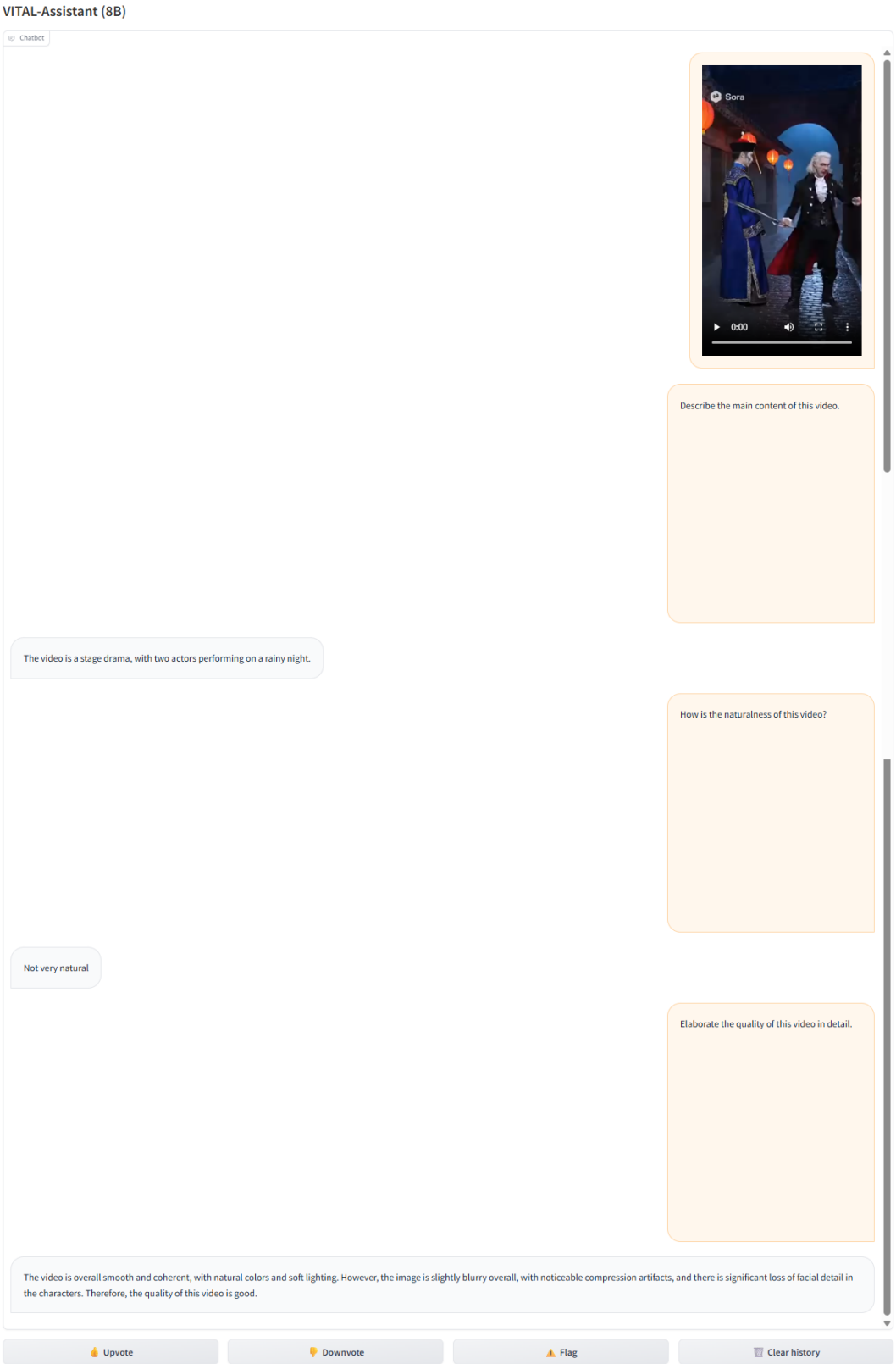

Figure 13. A video generated by *Sora2* with the prompt: Generate a video of two actors, one playing a Chinese ancient zombie and the other playing a European vampire, battling each other.

VITAL-Assistant (8B)

Chatbot

Describe the quality of this video in detail.

The video is relatively clear, with smooth playback and no stuttering. The colors are rich and the lighting is sufficient, but there is a slight blur in the picture. The video is composed of multiple scenes stitched together, resulting in a lack of coherence in the footage. Therefore, this video looks quite beautiful.

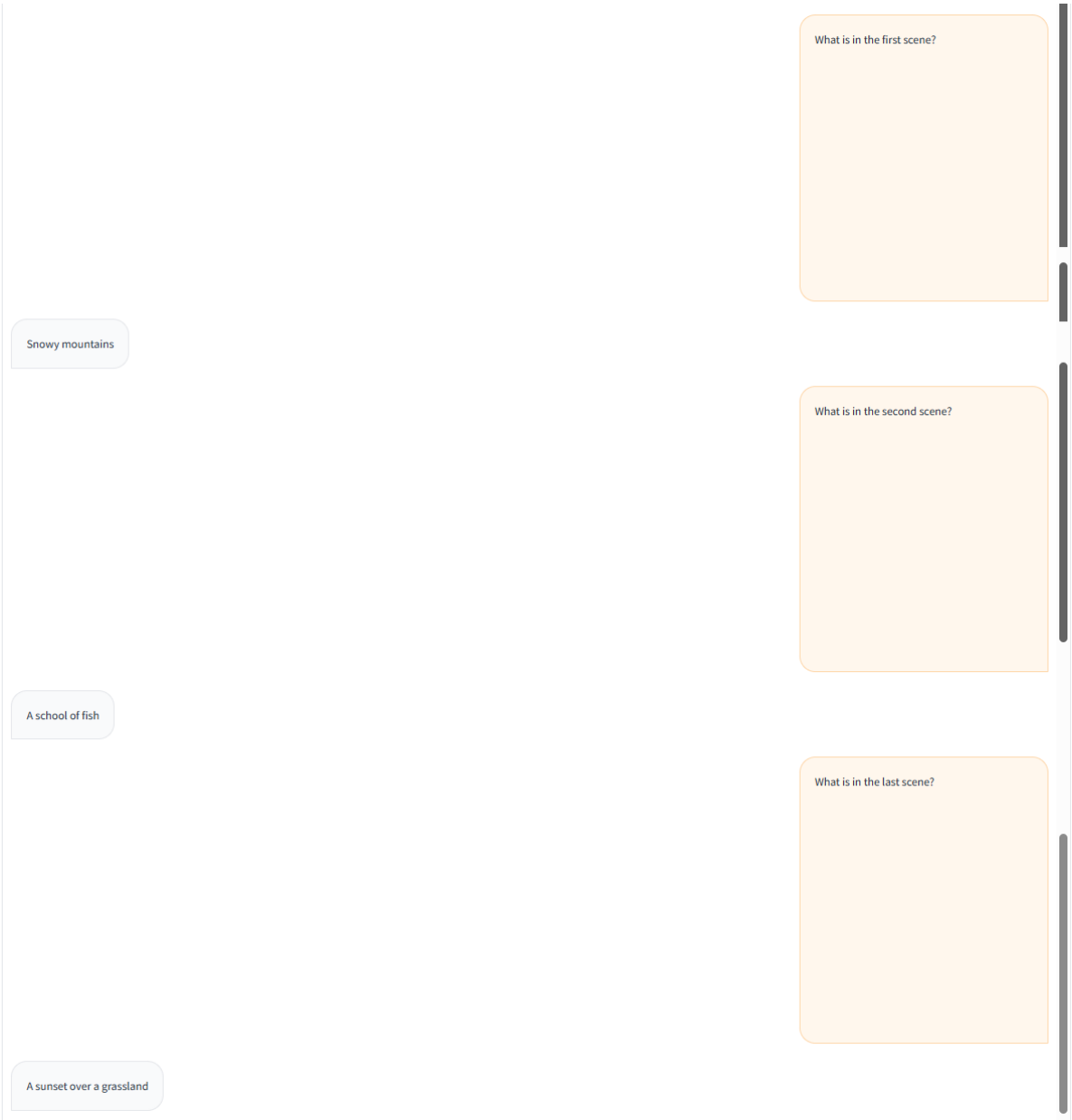

Figure 14. A video generated by *Sora2* with the prompt: Present a video consisting of four scenes: snow mountain, underwater, desert, and grassland. Each scene should be highly realistic and appear to be filmed with a handheld camera.

VITAL-Assistant (8B)

Chatbot

Elaborate the visual quality of this video in detail.

The video has low sharpness, with noticeable motion blur and severe ghosting. The color is relatively monotonous, lacking aesthetic appeal in terms of composition. Therefore, this video is not visually appealing.

Figure 15. A video captured with a shaking camera in a staircase.