

STORM: Benchmarking Visual Rating of MLLMs with a Comprehensive Ordinal Regression Dataset

Supplementary Material / Appendix

A APPENDIX OVERVIEW

Our supplementary includes the following sections:

- **Section B: Framework details.** Details for model design, implementation and training data.
- **Section C: More Dataset Details and Visualization.** More Details and Visualization of our dataset and demos.
- **Section D: More experiment results.** Additional performance evaluation and performance analysis.
- **Section E: Prompt design.** Prompt for generating the coarse-to-fine CoT dataset and evaluating the performance.
- **Section F: Limitations.** Discussion of limitations of our work.
- **Section G: Potential negative societal impacts.** Discussion of potential negative societal impacts of our work.
- **Section H: Disclaimer.** Disclaimer for the visual rating dataset and the related model.
- **Section I: Use of LLM.** Describe the usage of LLM.

Reproducibility Statement: We offer the anonymous Datasets and code links below to ensure our framework can be reproduced easily.

Artifcat	Link	License
Code Repository	https://anonymous.4open.science/r/STORM-CDC7/README.md	Apache-2.0 license
Data	https://huggingface.co/datasets/ttlyy/ORD	CC BY 4.0
Model Weights	https://huggingface.co/datasets/ttlyy/ORD	Apache-2.0 license

The authors are committed to ensuring its regular updates.

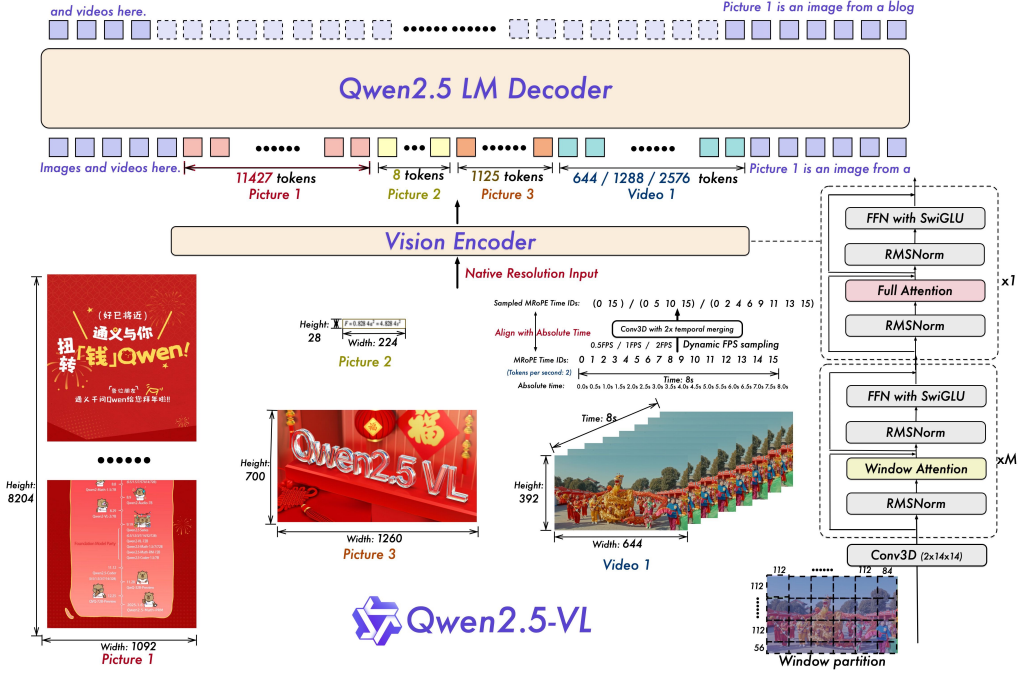


Figure 5: Overview of Qwen-2.5-VL pipeline.

B FRAMEWORK DETAILS

B.1 MODEL DETAILS

For LLaVA-1.5-7B, we choose the pre-trained ViT-L/14 of CLIP [Radford et al. \(2021b\)](#) as the vision encoder and Vicuna-7B [Chiang et al. \(2023\)](#) as our LLM, which has better instruction following capabilities in language tasks compared to LLaMA [Touvron et al. \(2023\)](#). For Qwen2.5-VL-3B, the vision encoder the native dynamic resolution ViT. The overview of Qwen-2.5-VL [Bai et al. \(2023b\)](#) are shown in Fig. 5. Considering an input original image, we take the vision encoder to obtain the visual feature. Our STORM-3B employs Qwen-2.5-VL-3B as the backbone.

B.2 IMPLEMENTATION DETAILS

Our model undergoes a two-stage training process. In the first stage, we pre-train the model for 1 epoch using a learning rate of $2e-3$ and a batch size of 128. For the second stage, we fine-tune the model for 1 epoch on our visual rating dataset, employing a learning rate of $2e-5$ and a batch size of 128. The Adam optimizer with zero weight decay and a cosine learning rate scheduler are utilized. To conserve GPU memory during fine-tuning, we employ FSDP (Full Shard Data Parallel) with ZeRO3-style. All models are trained using $32 \times A100s$. In the case of training the setting with a 7B LLM and a resolution of 224, the first/second pre-training stage completes within 1/16 hours.

C MORE DATASET DETAILS AND VISUALIZATION

C.1 DATASETS TRAINING AND TESTING SPLIT.

In this section, we provide the sample numbers of training and test split of all datasets, as shown in Tab. 6 and Tab. 7.

Table 6: Training and testing split of IQA and IAA domain datasets. Training split includes full version and lite version.

Dataset	SPAQ	CDB	KonIQ	AVA	TAD66K	Aesthetic
Training Full	8900	936	-	229958	52224	-
Training Lite	8900	936	-	25551	13056	-
Testing	2225	233	2014	25550	14076	1370

Table 7: Training and testing split of FAE, MDG and HDE domain datasets. Training split includes full version and lite version.

Dataset	Adience	CACD	Morph	UTK	Eyepacs	DeepDR	APTOS	HCI
Training Full	15589	147102	40012	-	31599	1200	-	-
Training Lite	15589	16345	10003	-	31599	1200	-	-
Testing	1732	16344	10003	2410	3527	400	366	132

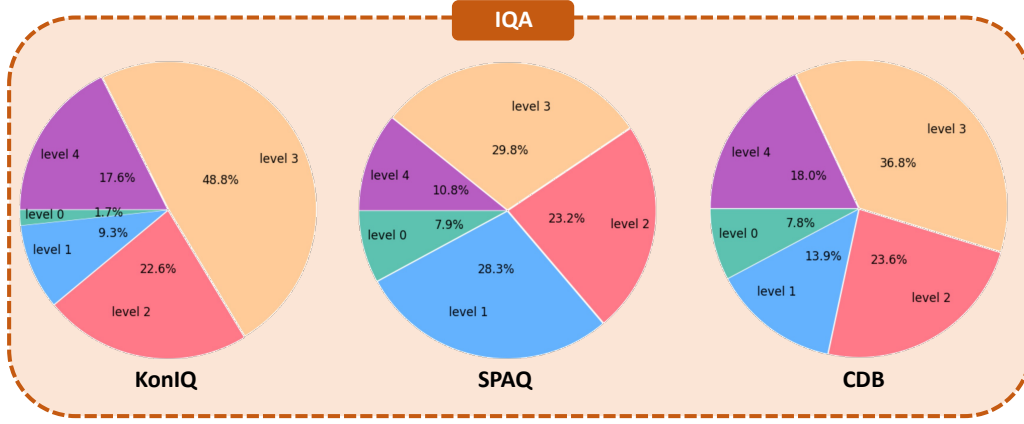


Figure 6: Statistics of the IQA domain datasets.

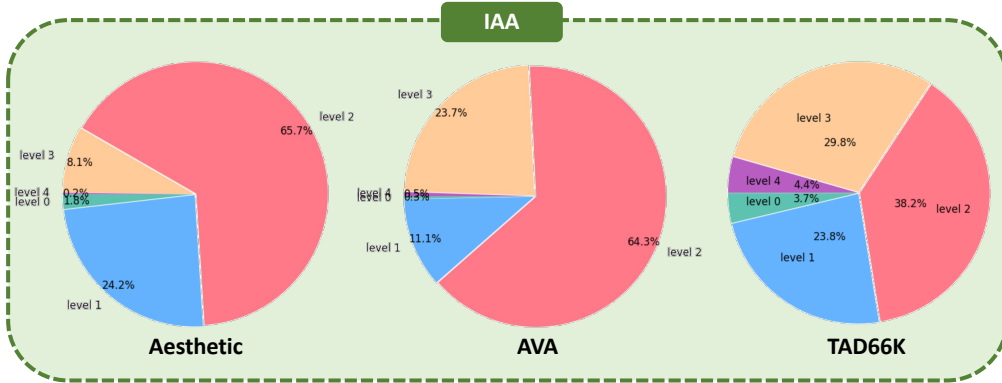


Figure 7: Statistics of the IAA domain datasets.

C.2 DATASETS DISTRIBUTION VISUALIZATION.

In this section, we provide a visualization of the data statistics. We partition the category distribution of each dataset in Fig. 6 Fig. 7 Fig. 8 Fig. 9

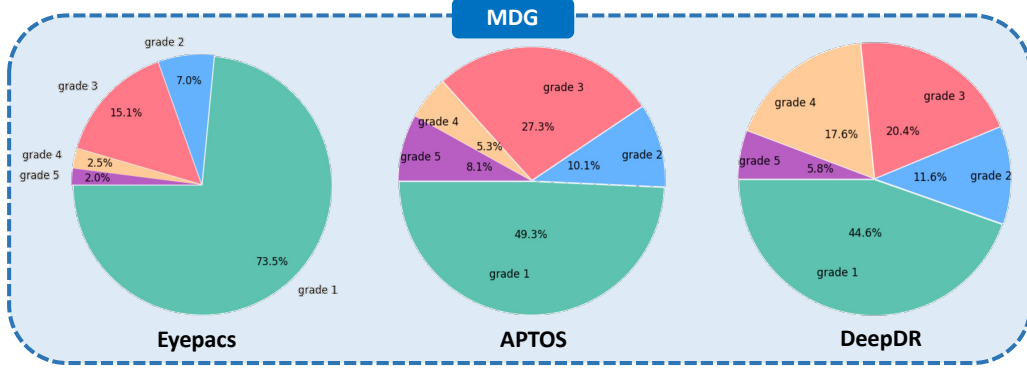


Figure 8: Statistics of the MDG domain datasets.

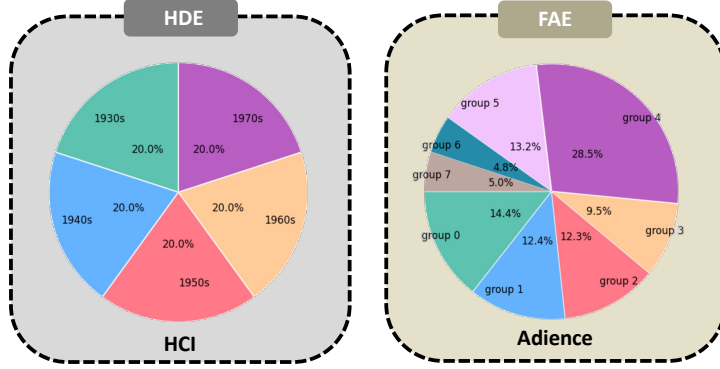


Figure 9: Statistics of the FAE and HDE domain datasets.

D MORE EXPERIMENT RESULTS

D.1 LARGER STORM MODEL

Tab. 8 and Tab. 8 show the performance of STORM-7B using Qwen2.5-VL-7B as the backbone. However, the performance is not much different from the 3B version. Therefore, we choose STORM-3B as the final model.

D.2 PLCC AND SRCC PERFORMANCE IN IQA TASKS.

Tab. 10 show the SRCC and PLCC results. It can be seen that our STORM achieves the best performance both in SRCC and PLCC on all IQA datasets, showing the effectiveness of our method.

D.3 DIFFERENT FINE-TUNING STRATEGIES.

To explore the effect of different parameter fine-tuning methods for LLMs. We compare the commonly used Low-Rank Adaptation (LoRA) (Hu et al., 2022) and Full Fine-Tuning (FFT) methods, and report the results in the lower part of Tab. 11. One can observe that FFT performs better and is more robust. Hence, we adopt FFT for all the fine-tuning experiments.

D.4 CONFUSION MATRIXES ANALYSIS

We provide more visualization results of confusion matrixes of our STORM on zero-shot datasets in Fig. 10 Fig. 11 Fig. 12 Fig. 13 and Fig. 14.

Table 8: ACC performance of the STORM-7B.

		IQA			FAE			
MLLM	Tra.	SPAQ	ChallengeDB	KonIQ	Adience	CACD	Morph	UTK
STORM-3B	Lite	0.583	0.468	0.582	0.534	-	-	-
STORM-7B	Lite	0.514	0.438	0.543	0.503	-	-	-

		IAA			MDG		HDE	Average
MLLM	Tra.	TAD66K	AVA	Aes.	Eyepacs	DeepDR	APTOS	
STORM-3B	Lite	0.370	0.650	0.658	0.734	0.435	0.508	0.341
STORM-7B	Lite	0.367	0.654	0.541	0.177	0.340	0.429	0.250

Table 9: MAE performance of the STORM-7B.

		IQA			FAE			
MLLM	Tra.	SPAQ	ChallengeDB	KonIQ	Adience	CACD	Morph	UTK
STORM-3B	Lite	0.442	0.597	0.431	0.636	8.202	5.975	5.879
STORM-7B	Lite	0.562	0.652	0.496	0.641	7.776	5.405	5.508

		IAA			MDG		HDE	Average
MLLM	Tra.	TAD66K	AVA	Aes.	Eyepacs	DeepDR	APTOS	
STORM-3B	Lite	0.726	0.363	0.360	0.511	1.280	1.098	1.958
STORM-7B	Lite	0.739	0.353	0.514	1.481	1.093	1.003	1.129

Table 10: SRCC and PLCC results of all models in IQA tasks.

		SPAQ		CDB		KonIQ	
MLLM	Tri.	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
LLaVA-1.5-7B	Zero	-0.034	-0.007	0.094	0.130	-0.007	-0.014
LLaVA-1.5-7B	Lite	0.037	0.036	-0.008	-0.005	0.008	0.009
MiniGPT-v2-7B	Zero	0.384	0.376	0.192	0.198	0.279	0.266
BLIP2-opt-2.7B	Zero	-0.067	-0.115	-0.085	-0.073	-0.070	-0.077
BLIP2-opt-2.7B	Lite	0.172	0.173	0.101	0.090	0.038	0.049
InternVL-2B	Zero	0.333	0.327	0.134	0.134	0.095	0.088
InternVL-2B	Lite	0.365	0.354	0.229	0.233	0.107	0.110
Qwen2.5-VL-3B	Zero	0.778	0.790	0.690	0.683	0.597	0.531
Qwen2.5-VL-3B	Lite	0.787	0.803	0.575	0.537	0.677	0.627
STORM-3B	Lite	0.804	0.817	0.675	0.656	0.692	0.650
STORM-3B	Full	0.808	0.822	0.686	0.674	0.727	0.690

Table 11: Ablation study on different training strategies.

Fine-tuning Strategy	IQA		FAE		IAA		MDG		HDE		Average	
	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE
LoRA Hu et al. (2022)	0.171	1.522	0.189	8.301	0.199	1.041	0.553	0.985	0.227	1.311	0.289	3.225
FFT	0.544	0.490	0.534	5.173	0.562	0.483	0.559	0.963	0.341	0.924	0.533	1.958

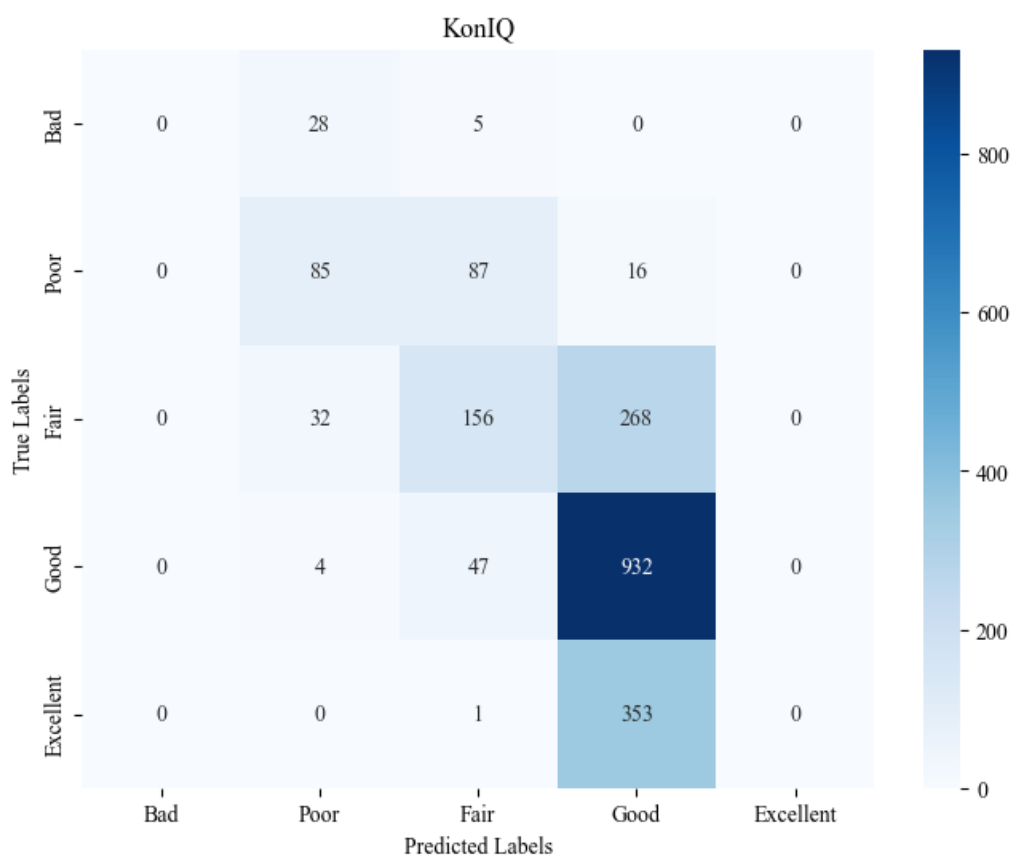


Figure 10: Confusion matrixes visualization results of the STORM on the KonIQ dataset.

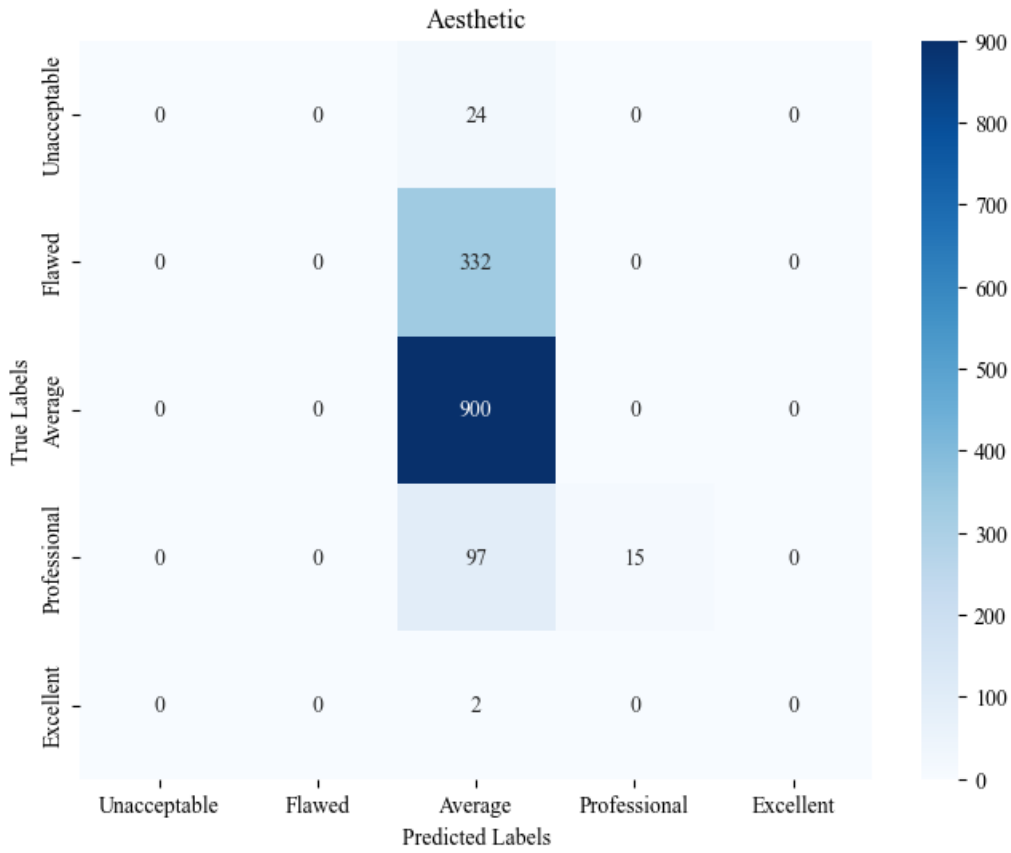


Figure 11: Confusion matrixes visualization results of the STORM on the Aesthetic dataset.

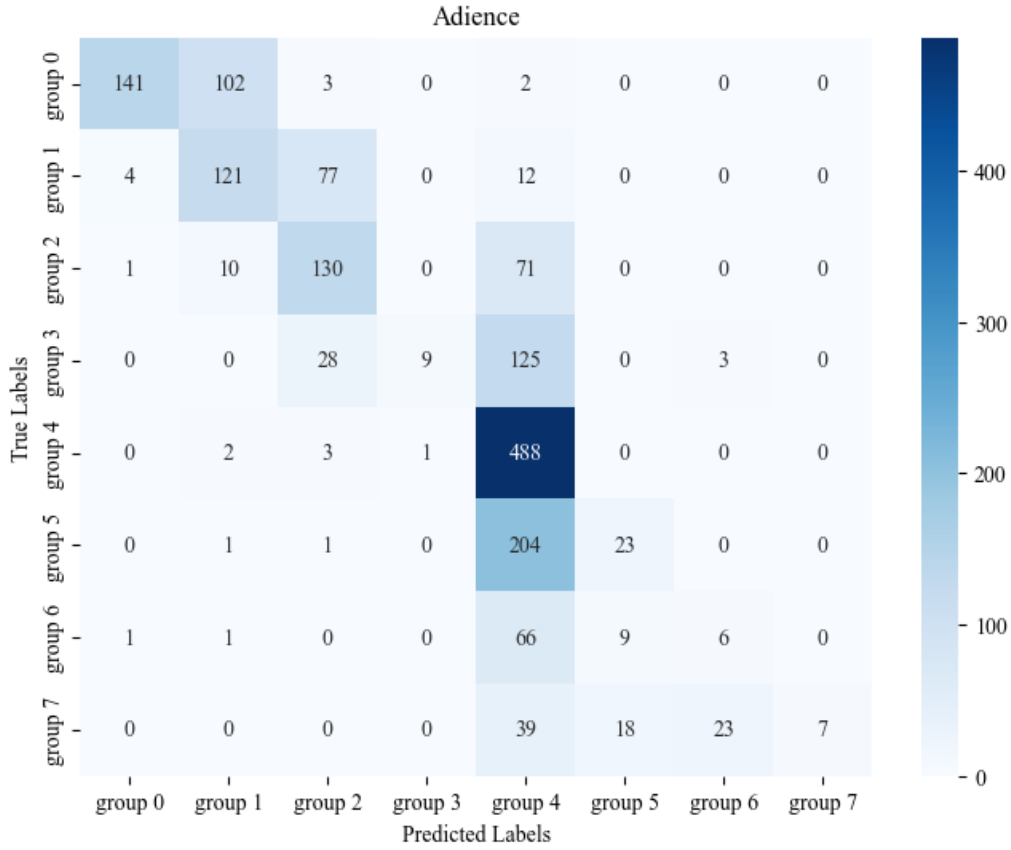


Figure 12: Confusion matrixes visualization results of the STORM on the Adience dataset.

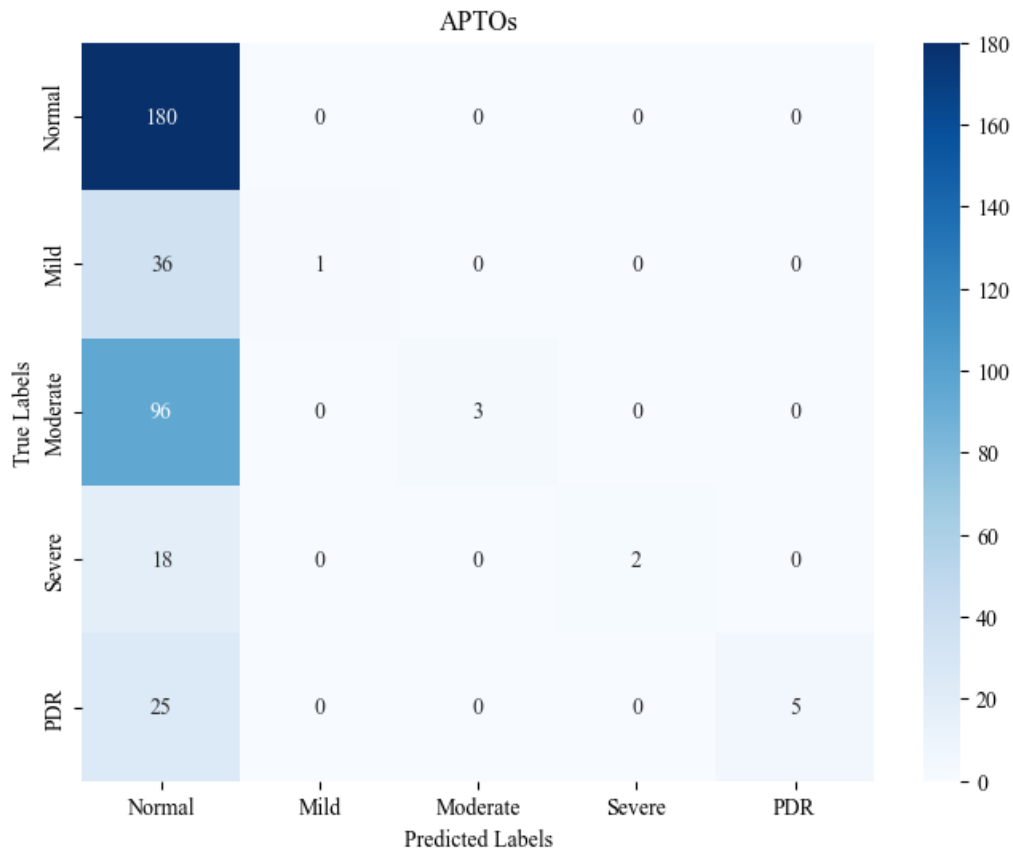


Figure 13: Confusion matrixes visualization results of the STORM on the APTOS dataset.

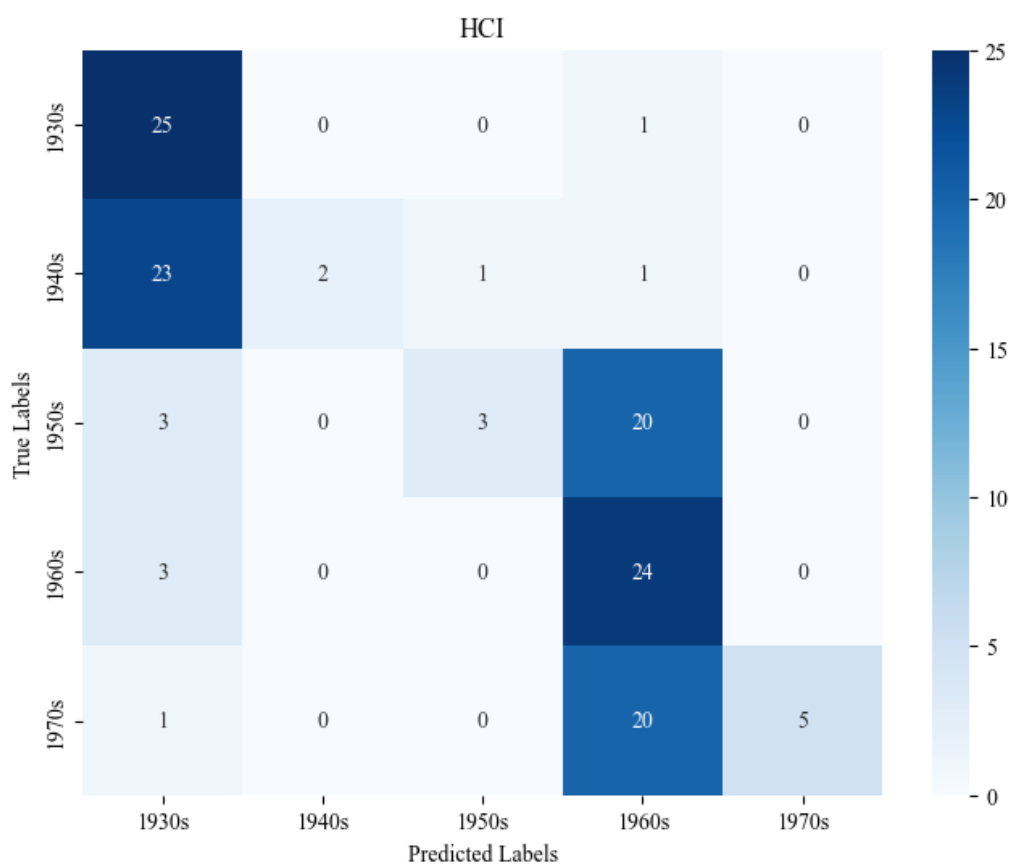


Figure 14: Confusion matrixes visualization results of the STORM on the HCI dataset.

E PROMPT DESIGN

E.1 GENERATING THE DATASET FOR IQA

`image`: You are now an advanced Image Quality Evaluator, and your task is to assess the quality of the provided image. Please evaluate the image's quality based on a 5-rate scale: rate0(Bad), rate1(Poor), rate2(Fair), rate3(Good), rate4(Excellent). Please provide the coarse category that can help you answer the question better. Please first coarsely categorise the image: rate0-1(Below Fair), rate2(Fair), rate3-4(Above Fair). Based on the coarse classification, proceed to make a final rate prediction. The specific steps are as follows:

1. Make the coarse prediction with the candidates:rate0-1(Below Fair), rate2(Fair), rate3-4(Above Fair).
2. Based on the coarse classification, proceed to make a final age prediction with the candidates: rate0(Bad), rate1(Poor), rate2(Fair), rate3(Good), rate4(Excellent).
3. Please note that the coarse thoughts and the final answer should be consistent.

Answer: [Coarse answer], [Final answer]

E.2 GENERATING THE DATASET FOR IAA

`image`: You are now an advanced Aesthetic Evaluation Evaluator, and your task is to assess the aesthetic quality of the provided image. Please evaluate the image's aesthetic quality based on a 5-level scale: level0(Unacceptable), level1(Flawed), level2(Average), level3(Professional), level4(Excellent). Please first coarsely categorise the image: level0-1(Below Average), level2(Average), level3-4(Above Average). Based on the coarse classification, proceed to make a final level prediction. The specific steps are as follows:

1. Make the coarse prediction with the candidates:level0-1(Below Average), level2(Average), level3-4(Above Average).
2. Based on the coarse classification, proceed to make a final age prediction with the candidates: level0(Unacceptable), level1(Flawed), level2(Average), level3(Professional), level4(Excellent).
3. Please note that the coarse thoughts and the final answer should be consistent.

Answer: [Coarse answer], [Final answer]

E.3 GENERATING THE DATASET FOR FAE

`image`: You are an experienced facial analysis expert, and you need to estimate the age group of the person in the provided facial image based on their facial features. The known age range of the image is from 16 to 77 years old. Please first coarsely categorise the image: Teenager(16-24 years old), Adult(25-47 years old), Elder(48+ years old). Based on the coarse classification, proceed to make a final age prediction. The final output should be in the format: Coarse Answer: [result], Predicted Age: [result]. The specific steps are as follows:

1. Make the coarse prediction with the candidates: Teenager(16-24 years old), Adult(25-47 years old), Elder(48+ years old).
2. Based on the coarse classification, proceed to make a final age prediction with the candidates: from 16 to 77 years old.
3. Please note that the coarse thoughts and the final answer should be consistent.

Answer: Coarse answer], [Predicted Age]

E.4 GENERATING THE DATASET FOR MDG

`image`: You are an experienced ophthalmologist, and you need to perform disease grading on the provided fundus image. These are all the candidate stages: stage0(no retinopathy), stage1(mild NPDR), stage2(moderate NPDR), stage3(severe NPDR) and stage4(PDR). Please first coarsely categorise the fundus: Normal(stage0), Early(stage1-2), Late(stage3-4). Based on the coarse classification, proceed to make a final stage prediction. The specific steps are as follows:

1. Make the coarse prediction with the candidates: Normal(stage0), Early(stage1-2), Late(stage3-4).
2. Based on the coarse classification, proceed to make a final age prediction with the candidates: stage0(no retinopathy), stage1(mild NPDR), stage2(moderate NPDR), stage3(severe NPDR) and stage4(PDR).
3. Please note that the coarse thoughts and the final answer should be consistent.

Answer: [Coarse answer], [Predicted grade]

E.5 GENERATING THE DATASET FOR HDE

`image`: You are now an advanced history researcher, and you need to grade the provided images by decade. These are all candidate categories: phase0(1930s), phase1(1940s), phase2(1950s), phase3(1960s), and phase4(1970s). Please first coarsely categorise the image: Early(phase0-phase1), Mid(phase2), Late(phase3-phase4). Based on the coarse classification, proceed to make a final phase prediction. The final output should be in the format: Coarse Classification: [result], Predicted Phase: [result]. The specific steps are as follows:

1. Make the coarse prediction with the candidates: Early(phase0-phase1), Mid(phase2), Late(phase3-phase4).
2. Based on the coarse classification, proceed to make a final age prediction with the candidates: phase0(1930s), phase1(1940s), phase2(1950s), phase3(1960s), and phase4(1970s).
3. Please note that the coarse thoughts and the final answer should be consistent.

Answer: [Coarse answer], [Predicted Phase]

F LIMITATIONS

The definitions of labels for different domain tasks are quite diverse.

In scenarios where the definitions of labels for different domain tasks are quite diverse, STORM may struggle to possess fluctuation in performance according to different text definitions generated of labels. This places a relatively high demand on the user's ability to accurately define corresponding text prompts of rating categories.

Our data pipeline inherits the limitations of utilizing GPT-4 API to generate text definition. (1) Accuracy and Misinformation: Generated content may not always be accurate, which could lead to the spread of misinformation. To mitigate this, we have designed a manual adjustment script as a post-process to improve text prompt quality. (2) Bias and Fairness: Since we do not have access to the training data of GPT-4, the generated instructional data might reflect inherent biases, potentially reinforcing social or cultural inequalities present in the base model training. In terms of data usage, we explicitly state that OpenAI's terms must be adhered to, and the data can only be used for research purposes.

G POTENTIAL NEGATIVE SOCIETAL IMPACTS

The potential negative societal impacts of our work are similar to other MLLMs and LLMs. The development of CoT and MLLMs, while advancing AI, poses societal risks like increased privacy invasion, the perpetuation of biases, the potential for misinformation, job displacement, and ethical concerns regarding accountability and consent.

H DISCLAIMER

This dataset was collected and released solely for research purposes, with the goal of making the MLLMs dynamically focus on visual inputs and provide intermediate interpretable thoughts. The authors are strongly against any potential harmful use of the data or technology to any party.

Intended Use. The data, code, and model checkpoints are intended to be used solely for (I) future research on visual-language processing and (II) reproducibility of the experimental results reported in the reference paper. The data, code, and model checkpoints are not intended to be used in clinical care or for any clinical decision making purposes.

Primary Intended Use. The primary intended use is to support AI researchers reproducing and building on top of this work. STORM and its associated models should be helpful for exploring various vision question answering (VQA) research questions.

Out-of-Scope Use. Any deployed use case of the model — commercial or otherwise — is out of scope. Although we evaluated the models using a broad set of publicly-available research benchmarks, the models and evaluations are intended for research use only and not intended for deployed use cases.

I USE OF LLM

We employed LLMs solely for language polishing and manuscript refinement purposes. The LLM assistance was restricted to improving grammatical accuracy, sentence flow, and overall presentation clarity. All research content, methodology, analysis, and scientific conclusions were developed independently by the authors without LLM contribution. The LLM was not utilized for idea generation, experimental design, data interpretation, or scientific reasoning.