Modification instructions for the NTIRE-122 paper

Note: This paper is based on the second-place solution of CVPR NTIRE 2025 XGC Quality Assessment - Track 2: AI Generated Video. When this paper was submitted for the first time, it was considered "Conditional Accept". The following are the reviewers' revision suggestions and corresponding explanations for the revisions.

@Reviewer #1

Modification suggestions: This paper lacks validation experiments. The model is only tested on the NTIRE 2025 XGC Quality Assessment - Track 2 AI Generated Video test dataset. Further validation of the model's generalization performance on other AIGV datasets is needed.

We tested the AIGVEval method proposed in this paper on the open source dataset T2VQA-DB and analyzed the results in the section **4.4. Experimental Results.**

Table 4. Performance Comparison on T2VQA-DB[20]. **Red**: the best, **Bold**: ours

Model	PLCC ↑	SROCC ↑	KRCC ↑	RMSE ↓
CLIPSim [44]	0.1277	0.1047	0.0702	21.683
BLIP [22]	0.1860	0.1659	0.1112	18.373
ImageReward [52]	0.2121	0.1875	0.1266	18.243
ViCLIP [42]	0.1449	0.1162	0.0781	21.655
UMTScore [27]	0.0721	0.0676	0.0453	22.559
SimpleVQA [38]	0.6338	0.6275	0.4466	11.163
BVQA [21]	0.7486	0.7390	0.5487	15.645
FAST-VQA [45]	0.7295	0.7173	0.5303	10.595
DOVER [46]	0.7693	0.7609	0.5704	9.8072
T2VQA [20]	0.8066	0.7965	0.6058	9.0221
T2VEval [34]	**0.8175**	**0.8049**	**0.6159**	**8.6133**
AIGVEval	**0.7494**	**0.7636**	**0.5608**	**10.034**

@Reviewer #3

Modification suggestions: It is recommended to add som latest LLM-based VQA methods, such as [1,2] in Section 2.2, to make the related works section more comprehensive.

We have added these two references in the **Related Work** section, as shown in the figure below:

respectively, and predicts five-class scores with Vicuna v1.5 [57] along with language explanations. Wang et al. proposed the AIGV-Assessor [40], which extracts spatial and temporal features of video content using InternViT and SlowFast, respectively. It employs InternVL2-8B to map vi-

tion of low-level distortions. LMM-VQA [7] performs end-to-end score regression by feeding visual features into a Llama-3-8b-Instruct decoder [8]. Duan et al. proposed Fine-VQ [5], which extracts spatial and temporal features of video content using InternViT and SlowFast, while utilizing InternVL2-8B for quality prediction. Additionally,