WACV 2026

Language Integration in Fine-Tuning Multimodal Large Language Models for Image-Based Regression

RvTC: Regression via Transformer-Based Classification

Roy H. Jennings Samsung R&D Center, Israel

github.com/royhj/rvtc

The Problem: Are MLLMs Actually Using Language?

Current MLLM Methods

Preset vocabulary
"Excellent", "Good", "Fair", "Poor"

Generic task prompts
"How would you rate this image?"

Assumption:
This mimics human rating behavior
and leverages language understanding

Our Key Finding

Neither preset vocabularies nor generic prompts improve over image-only training!

Current methods fail to leverage semantic understanding from text.

Language is being added for no benefit.

RvTC: A Simpler, Better Approach

Previous Methods

Preset vocabulary
"Excellent" → 5
"Good" → 4
Manual crafting
Rigid, limited

RvTC

Bin-based classification
Image only
Flexible bin count
No vocabulary crafting

RvTC+

Data-specific prompts
"Outdoor Macro Shot"
Semantic info per image
Cross-modal reasoning

How It Works

Replace vocabulary-constrained output with flexible bins → **Simply increase bin count to reduce discretization error**

No complex distributional modeling needed. Just more bins = better accuracy.

What Kind of Language Actually Helps?

Generic Task Prompts

"How would you rate this image?"
"Rate the aesthetic quality."

No improvement over image-only.
Equivalent to not using text at all.

vs

Data-Specific Prompts

"Outdoor Macro Shot"
"Portrait — Natural Lighting"

Substantial improvement!
Unlocks cross-modal reasoning.

Verified on AVA & AGIQA-3k: Semantic understanding, not statistical bias

We systematically compare all training/evaluation combinations to isolate true semantic understanding from potential statistical shortcuts. Challenge titles provide genuine image-level information that MLLMs leverage for improved predictions.

Results: State-of-the-Art Performance

0.83
SRCC on AVA
Image-only RvTC

→

0.90
SRCC on AVA
RvTC+ (with prompts)

0.82
Q-Align
(prev. SOTA w/ text)

Method	Text Input	AVA SRCC	AVA PLCC	Architecture
Q-Align (2024)	Vocab + Prompt	0.820	0.822	mPLUG-Owl2
RvTC (Ours)	None (image-only)	0.833	0.831	mPLUG-Owl2
RvTC+ (Ours)	Data-specific	0.899	0.901	mPLUG-Owl2
RvTC (Ours)	None (image-only)	0.843	0.842	Qwen2-VL-2B
RvTC+ (Ours)	Data-specific	0.906	0.908	Qwen2-VL-2B

State-of-the-art across 4 image assessment benchmarks, validated on 2 MLLM architectures.

Evidence: The Model Is Using Cross-Modal Features

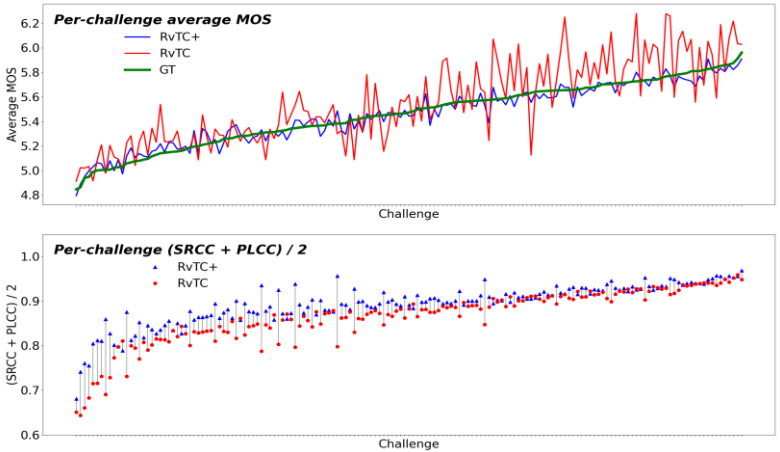
Shuffled Titles Ablation

We decompose gains into inter-challenge (statistical bias) and intra-challenge (semantic understanding).

Method	SRCC	PLCC
RvTC (image-only)	0.833	0.831
RvTC + Challenge ID	0.851	0.843
RvTC + Shuffled Titles	0.860	0.851
RvTC+ (real titles)	**0.899**	**0.901**

Challenge ID & Shuffled Titles
preserve grouping bias but remove semantics.
Real titles significantly outperform both →
Semantic content drives the gains.

Intra vs. Inter Challenge Analysis



Top: Inter-challenge avg MOS — could be statistical bias
Bottom: Intra-challenge correlation — cannot be explained by per-challenge statistics

RvTC+ (blue) consistently outperforms RvTC (red) in intra-challenge correlation → cross-modal reasoning.

Full details in Table 5 and Figure 3 of the paper.

Key Takeaways

1

RvTC: Simpler & Better

Bin-based regression replaces rigid vocabularies.
Image-only achieves new SOTA (0.83 on AVA).

2

Language Matters — If Done Right

Data-specific prompts unlock cross-modal reasoning.
Generic prompts provide zero benefit.

3

Generalizable

Consistent gains across 4 datasets,
2 architectures (mPLUG-Owl2, Qwen2-VL).

github.com/royhj/rvtc

arxiv.org/abs/2507.14997

Thank you!