ICML
International Conference On Machine Learning

ICML 2026

GeoReward: Mitigating Contextual Variable Overestimation in Vision-Language Models for Cross-Market Preference Prediction

A Presentation for the International Conference on Machine Learning (ICML) 2026

Shuo Liu • Huixiang Cai • Weiru Zhang • Xiaoyi Zeng

Alibaba International Digital Commerce Group

Alibaba Group 阿里巴巴集团

Background & Real-World Motivation

The Power of VLMs

Vision-Language Models (VLMs) like Qwen2-VL demonstrate remarkable capabilities in multimodal reasoning, enabling tasks from visual QA to complex scene understanding.

A key application is using VLMs as **multimodal reward models** to guide image generation training and rank user preferences.

The Real-World Pain Point

In cross-border e-commerce, the same product can have drastically different Click-Through Rate (CTR) preferences across countries.

However, existing VLMs often **fail to adapt**, producing **market-invariant predictions** that ignore the target country, leading to catastrophic decision-making collapse.

Input: Product Image (e.g., Sneakers)

USA Market

User Preference:**Image A**

Korea Market

User Preference:**Image B**

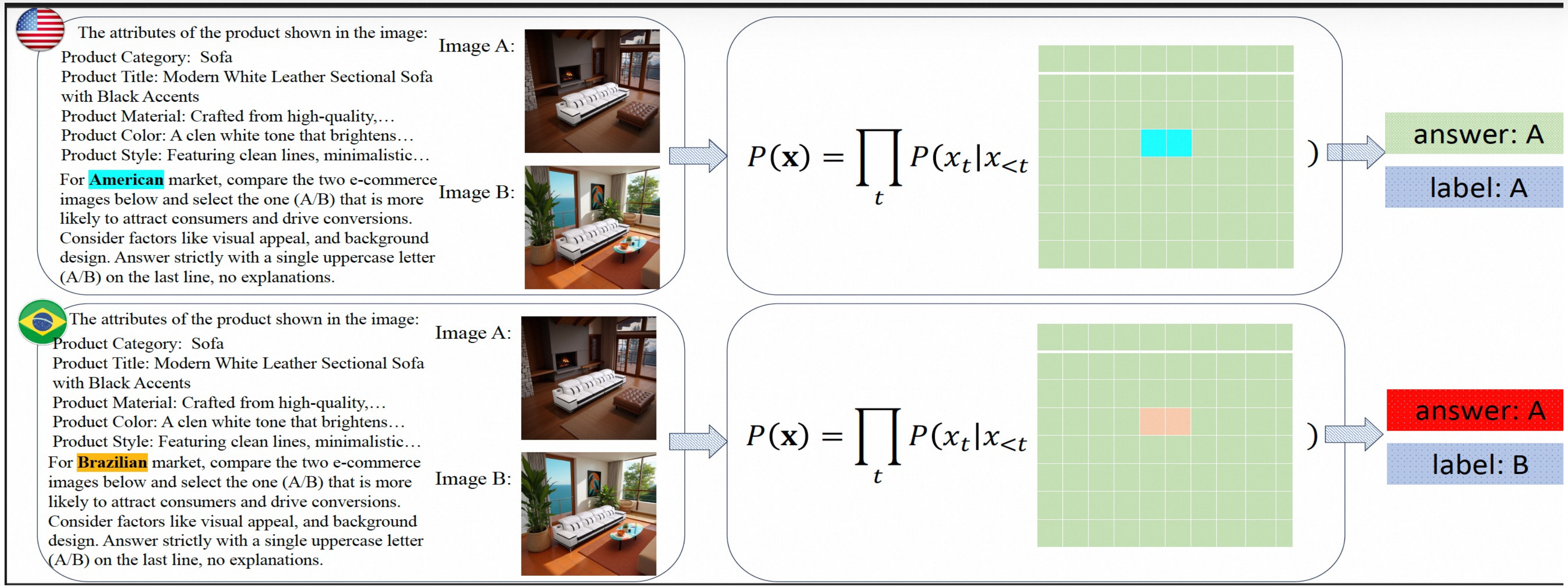
Standard VLM Prediction (Flaw)

Output: Always chooses Image A (Invariant)

Problem Definition: Contextual Variable Overestimation (CVE)

A Fundamental Architectural Bias

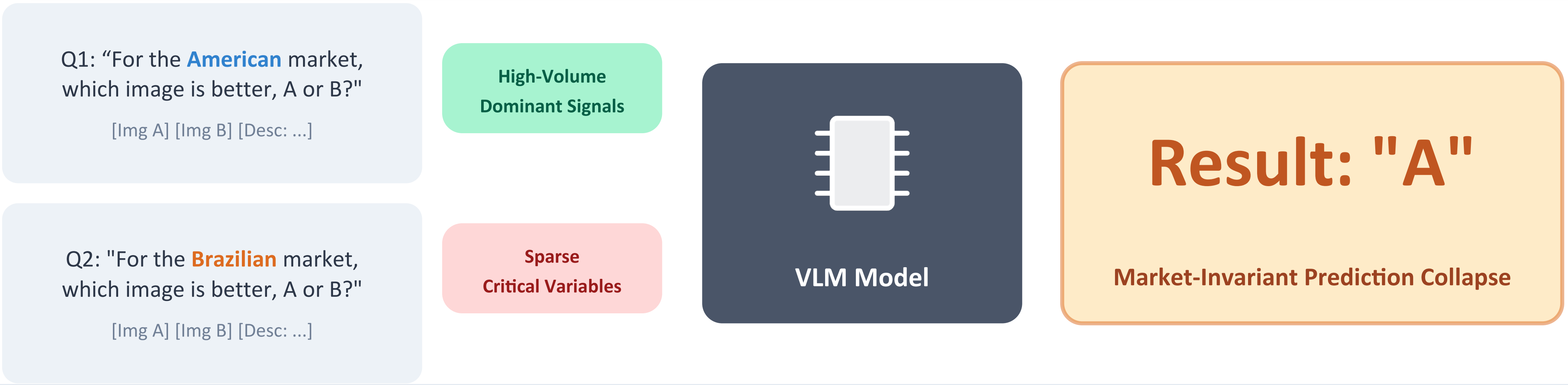
We formally define a new failure mode in VLMs: Contextual Variable Overestimation (CVE). This occurs when models systematically **overweight dense, high-volume visual or textual features**(like image patches or product descriptions) and **underweight sparse, but instruction-critical variables**(e.g., a single country name). The root cause lies in the model's autoregressive nature and attention mechanism, which statistically drown out the influence of sparse tokens.



Problem Definition: Contextual Variable Overestimation (CVE)

A Fundamental Architectural Bias

We formally define a new failure mode in VLMs: Contextual Variable Overestimation (CVE). This occurs when models systematically **overweight dense, high-volume visual or textual features**(like image patches or product descriptions) and **underweight sparse, but instruction-critical variables**(e.g., a single country name). The root cause lies in the model's autoregressive nature and attention mechanism, which statistically drown out the influence of sparse tokens.



Key Challenges & Research Gap

Pure SFT

Standard fine-tuning leads to severe prediction collapse. The model simply memorizes dominant patterns and ignores sparse context.

Prompt Engineering

Rearranging or repeating instructions cannot fix the **representation-level imbalance**. The model's internal attention is still biased.

Reward Models

Most multimodal reward models focus on general alignment and do not explicitly account for sparse, instruction-critical variables like CVE.

Core Technical Challenge: How to make VLMs actively attend to sparse critical variables without being overwhelmed by dominant features?

MACP Dataset: A New Benchmark(Core Contribution)

Multi-Country Ad Click Preference Dataset

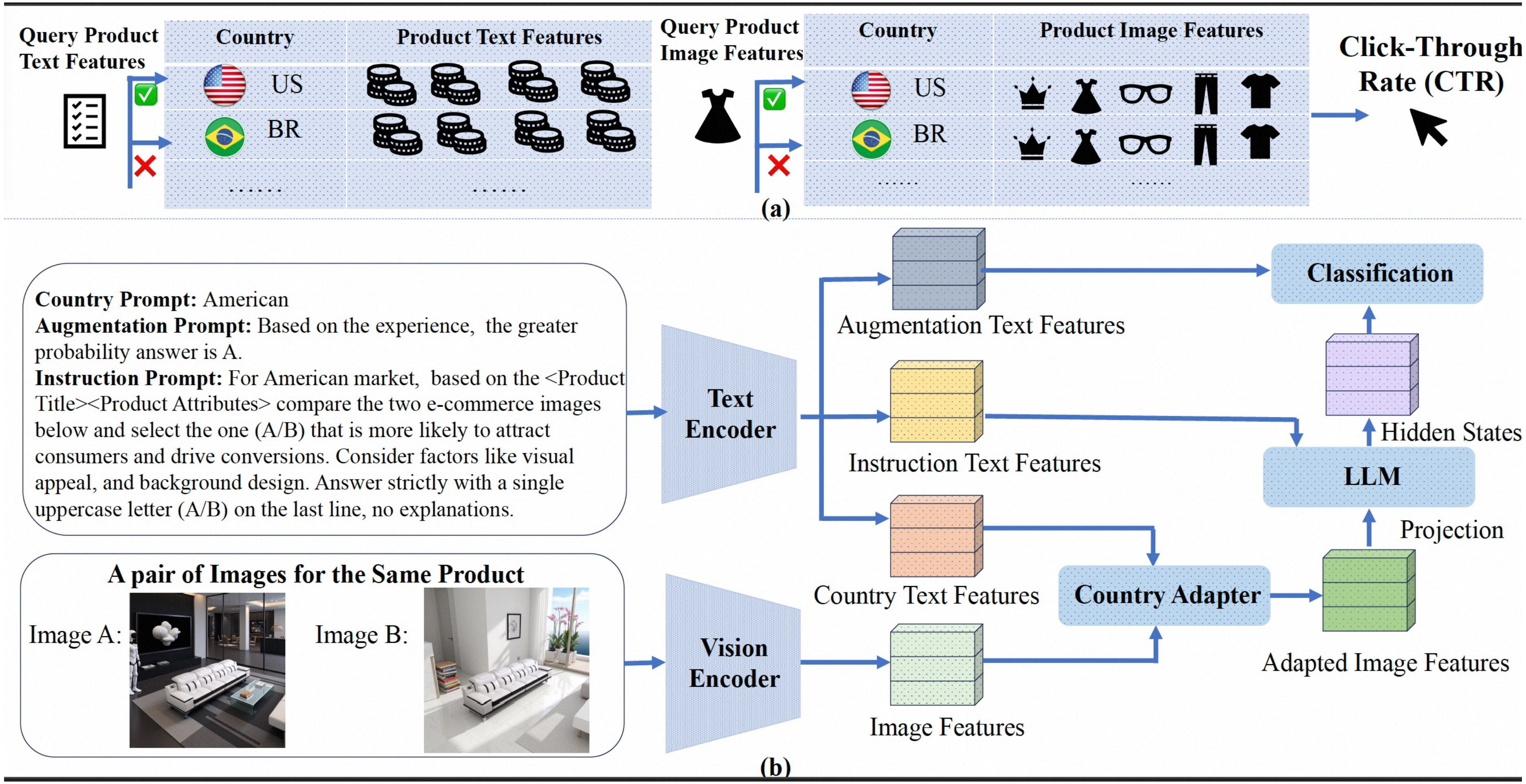
823K Training Samples	**180K** Test Samples	**10** Balanced Global Markets
67K Unique Products	**250K** Unique Ad Images	**30-day** Real CTR Preference as Ground Truth

The dataset is perfectly balanced across countries and product categories, providing a rigorous and unbiased testbed for studying Cross-Market Variation Effects (CVE).

Global Market Coverage

- **US** United States
- **FR** France
- **MX** Mexico
- **ES** Spain
- **BR** Brazil
- **SA** Saudi Arabia
- **JP** Japan
- **KR** Korea
- **CL** Chile
- **AU** Australia

Overall Framework: GeoReward Triple-Gate Pipeline



Overall Framework: GeoReward Triple-Gate Pipeline

1.Retrieval Gate (MA-RAG)

Market-Aware Retrieval-Augmented Generation

Injects external, market-specific knowledge to amplify the sparse context signal.

2.Consolidation Gate (CGVM)

Context-Guided Visual Modulation

Aligns internal visual representations with the country context to eliminate feature dominance.

3.Sensitivity Gate (SSL)

Selective Sensitivity Loss

Calibrates the learning process to penalize inattention to critical variables.

A synergistic solution working at the **data**, **feature**, and **optimization** levels.

Module 1: MA-RAG (Market-Aware Retrieval Augmentation)

Amplifying Sparse Context Signals

🔧 Core Mechanism

1.**Two-Stage Retrieval:**First semantic text retrieval, then visual similarity retrieval within results.
2.**Similarity Filtering:**A relevance threshold prevents data leakage and ensures distinct neighbors.
3.**Preference Aggregation:**Position-weighted voting on neighbors' CTRs generates a market-aware prior.

🎯 Purpose

Explicitly injects real-world, country-specific preference knowledge to compensate for weak context signals in the original instruction.

INPUT: User Query
Product Text | Product Image | Target Country

Step 1: Text-based Retrieval (KB)
→ Neighbor Set Nt (Semantic Similarity)

Step 2: Image-based Retrieval (on Nt)
→ Neighbor Sets Ni_A & Ni_B (Visual Similarity)

Step 3: CTR Preference Voting
→ Augmented Market Preference (A or B)

OUTPUT: Augmented Prior ➔ Main Model

Module 2: CGVM (Context-Guided Visual Modulation)

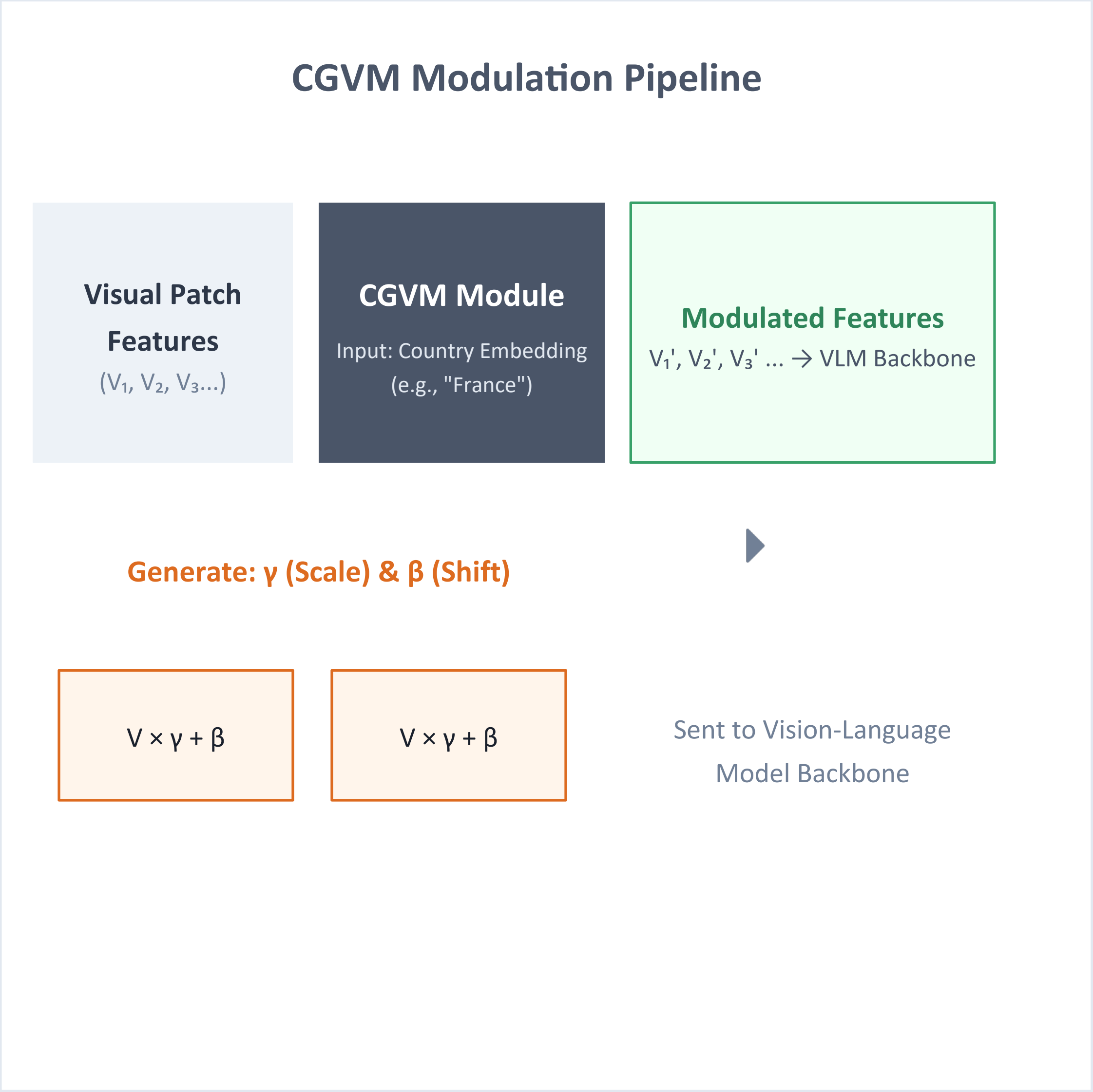
Aligning Visual Representations

Mechanism

1. A small feed-forward network takes the **country text embedding** as input.
2. It generates adaptive affine transformation parameters: scale (γ) and shift (β).
3. These parameters are applied to **all** visual patch features from the image encoder.

Purpose

This lightweight adapter forces the visual representation to be **country-condition-aware**. It ensures that the model "sees" the image based on the target country, eliminating the dominance of generic visual features.



Module 3: SSL (Selective Sensitivity Loss)

Calibrating the Learning Process

Total Loss Function

$$L = L_{BCE} + \lambda \cdot L_{penalty}$$

🔑 Mechanism

1. Focus Calculation: Measures attention (via hidden state norms) to critical tokens: Country, Product, and Image.
2. Conditional Penalty: Activates a penalty term **only when** the model makes a wrong prediction.
3. Adaptive Correction: Penalty is proportional to under-attention. Guides the model to reallocate attention.

⚙️ Execution Logic

Input: Compare Model Prediction vs. Ground Truth Label

Decision: Is the prediction incorrect?

❌ **No**
Proceed without penalty

✅ **Yes**
Calculate focus → Compute penalty → Add to loss

Outcome: Gradually corrects attention allocation

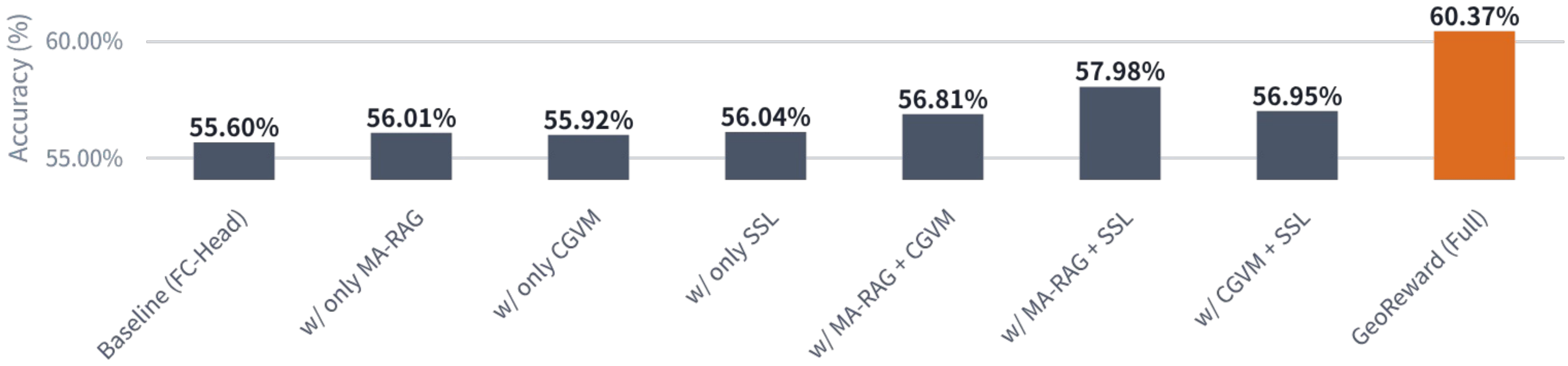
🎯 Purpose

Fixes the attention collapse and gradient vanishing problem for sparse context tokens during training.

Main Results: Effectively Mitigating CVE on MACP

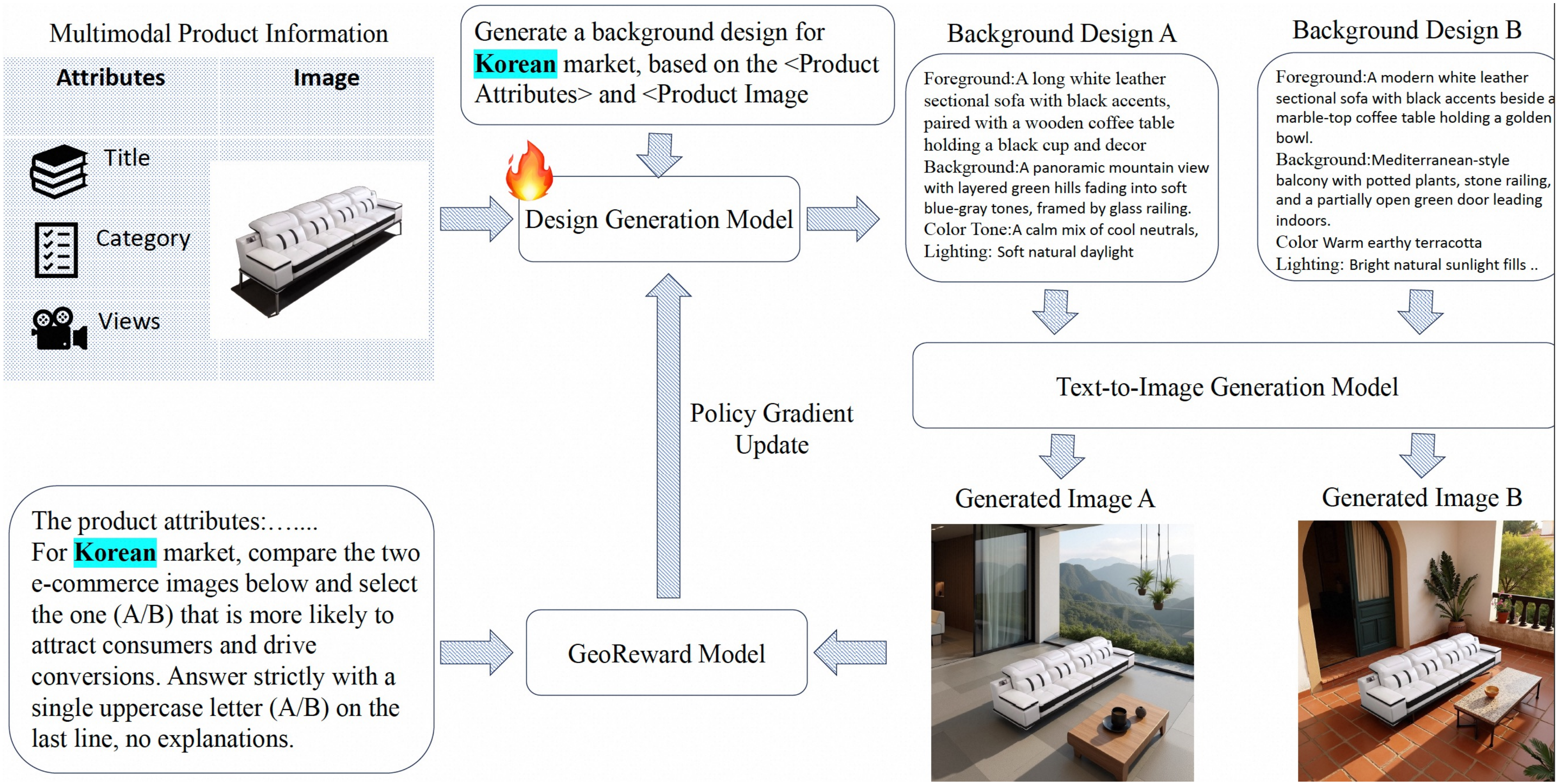
Model	Accuracy	Sensitivity	Note
Qwen2-VL (SFT)	44.61%	20.82%	⚠️ Severe collapse
Qwen2-VL (FC-Head)	55.60%	36.73%	🏛 Strong baseline
Ablation (w/o MA-RAG/CGVM/SSL)	~56% - 58%	~37% - 39%	📈 Partial improvement
GeoReward (Full Model)	**60.37%**	**40.84%**	**+15.75% relative gain (SOTA)**

Ablation Study: The Synergy of Three Gates



No single component is sufficient. The **synergy** of all three gates is crucial for achieving the best performance, forming a closed-loop solution.

Downstream Application: Market-Aware Ad Generation



Downstream Application: Market-Aware Ad Generation

1. DGM (VLM)

Design Generation Model

INPUT
- Product Information
- Target Country ID

OUTPUT
Background Design Prompts
(d_A, d_B)

2. T2I Model

SDXL + ControlNet

INPUT
- Design Prompts
- Original Product Image

OUTPUT
Candidate Ad Creatives
(I_A, I_B)

3. GeoReward (RM)

Geographic Reward Model

INPUT
- Generated Images
- Product + Country Context

OUTPUT
Pairwise Preference Score
(d+ = Preferred, d- = Rejected)

4. DPO Fine-tuning

Direct Preference Optimization

INPUT
- Preference Signal
(High-quality d+ / Low-quality d-)

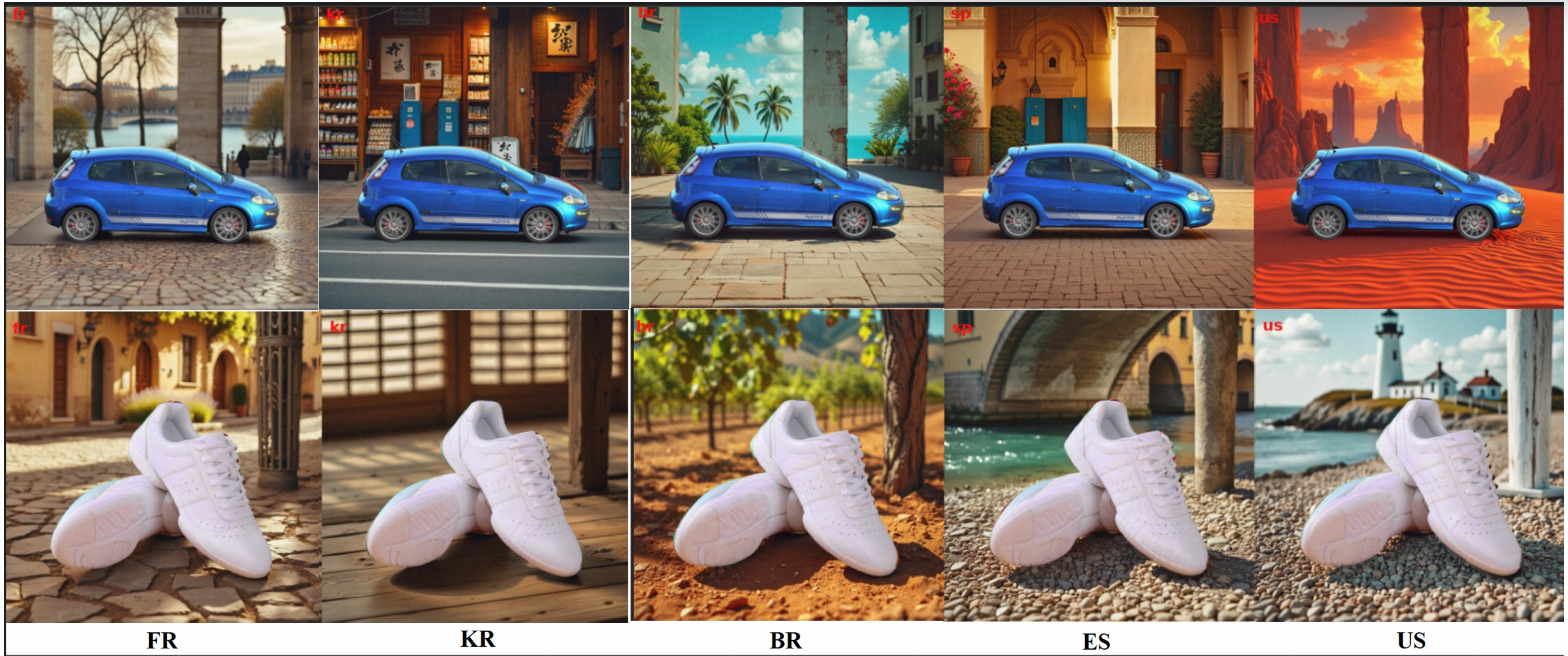
OUTPUT
Optimized DGM
Generates Market-Specific Ads

Downstream Application: Market-Aware Ad Generation

Experiment Setup

Baseline: DGM (w/o RL) — A standard fine-tuned model without reward model optimization.
Ours: DGM (+ GeoReward RL) — Optimized using Direct Preference Optimization (DPO) with GeoReward as the reward signal.

DGM (w/o RL)	DGM (+ GeoReward RL)
56.04%	59.60%



Contributions Summary

01 Problem Identification

We are the first to formally define and characterize **Contextual Variable Overestimation (CVE)**, a fundamental context-insensitivity failure mode in VLMs.

02 Benchmark Creation

We release **MACP**, a large-scale, balanced, real-world benchmark for cross-market preference prediction to facilitate future research on CVE.

03 Method Proposal

We introduce **GeoReward**, a novel triple-gate framework that mitigates CVE through synergistic data-level, feature-level, and optimization-level interventions.

04 Practical Application

We demonstrate how GeoReward enables **market-customized ad generation**, showcasing its tangible industrial value and broader applicability.

Thank You

Shuo Liu| jiexuan.ls@alibaba-inc.com