Pinnacle Academic Press
Proceedings Series

Vol. 3 2025

Article Open Access

Quantifying and Mitigating Dataset Biases in Video Understanding Tasks across Cultural Contexts

Gengrui Wei [1,*] and Zhuolin Ji [2]

[1] Computational Science and Engineering, Virginia Tech, VA, USA
[2] Computer Vision & Control, Illinois institute of technology, IL, USA
* Correspondence: Gengrui Wei, Computational Science and Engineering, Virginia Tech, VA, USA

PAP Proceedings Series
IFCIS·2025
2025 International Forum on Computing and Intelligent Systems
Kuala Lumpur·Malaysia | 16-17·August·2025
2025 Vol. 3 ISSN

Received: 20 May 2025
Revised: 31 May 2025
Accepted: 08 July 2025
Published: 17 July 2025

Abstract: Cross-cultural biases embedded in video datasets pose significant challenges to the fairness and generalization of video understanding models. Existing benchmarks are predominantly constructed from Western-centric visual corpora, leading to performance degradation when models are applied to underrepresented cultural contexts. This paper presents a comprehensive framework for quantifying and mitigating cultural biases in video understanding tasks. A multi-level analysis is conducted to identify cultural skew in existing datasets, revealing disparities in representation, annotation practices, and modality alignment. To address these biases, we propose a set of mitigation strategies encompassing culturally adaptive data augmentation, architecture-aware modality calibration, and causal intervention-based debiasing. Extensive experiments on action recognition, sign language translation, and captioning tasks demonstrate significant improvements in cultural fairness and semantic alignment. Evaluation metrics, including the Cultural Relevance Index (CRI), Fairness Gap (FG), and Modality Gap Index (MGI), provide quantitative evidence of improved cross-cultural robustness. Ethical considerations surrounding annotation, deployment, and interpretability are also discussed. This work contributes toward equitable and culturally inclusive video understanding systems that generalize beyond monocultural datasets.

Keywords: cross-cultural bias; video understanding; dataset fairness; causal debiasing

1. Introduction and Background

1.1. The Prevalence of Cultural Biases in Video Understanding Tasks

Recent advancements in deep learning have significantly improved video understanding tasks, including action recognition, event detection, and video captioning. These systems, trained on large-scale datasets, have demonstrated impressive performance on standard benchmarks. Despite these advances, a critical issue remains insufficiently addressed: the presence of cultural biases embedded within video datasets. As identified by, video understanding models trained predominantly on Western datasets exhibit significant performance degradation when applied to content from non-Western cultures [1]. The inherent biases stem from multiple sources, including data collection methodologies, annotation practices, and the cultural backgrounds of dataset creators. Studies have shown that standard video datasets contain disproportionate representations of Western activities, environments, and objects while under-representing cultural elements from Asian, African, and Middle Eastern contexts [2]. This imbalance creates a fundamental bias in how models interpret and understand visual content across different cultural settings. Datasets like Kinetics, ActivityNet, and YouTube-8M, while extensive in size, exhibit notable cultural skew, with Western scenarios accounting for over 70% of video content [3]. The modality complexity between visual representation and cultural interpretation further exacerbates these biases, as noted in translation tasks where models tend to overlook visual nuances specific to certain cultures.

1.2. Challenges in Cross-Cultural Video Analysis

Cross-cultural video analysis presents unique challenges beyond traditional computer vision problems. Cultural-specific gestures, customs, and contexts often carry implicit meanings that require cultural knowledge to interpret correctly. As demonstrated by someone, video understanding models struggle with cultural-specific actions that have no equivalent representation in training data [2,4]. The semantics of activities vary significantly across cultures, with identical physical movements potentially carrying different meanings depending on cultural context. The temporal dynamics of activities also exhibit cultural variation, with certain cultures emphasizing different phases of actions compared to others. Environmental settings and object interactions further complicate cross-cultural analysis, as observed in sign language translation systems that fail to account for cultural variations in signing. Technical challenges include feature extraction inconsistencies across different cultural contexts and the absence of standardized evaluation metrics that account for cultural diversity. The lack of annotators from diverse cultural backgrounds introduces systematic annotation biases, where Western interpretations are often imposed on non-Western content [5]. Additionally, language barriers in annotation create misalignments between visual content and textual descriptions, particularly for multilingual datasets.

1.3. Research Objectives

This research addresses the critical gap in quantifying and mitigating dataset biases in video understanding across cultural contexts. The primary objectives include developing robust methodologies for identifying and measuring cultural biases in video datasets. The research aims to establish comprehensive evaluation metrics that assess cultural representation and bias across different video understanding tasks. This study proposes novel techniques for quantifying cultural relevance in visual content through computational methods [5]. The research introduces bias mitigation strategies at both data and algorithmic levels, including calibration techniques inspired by causal interventions described [3]. The development of culture-aware training methodologies incorporates adversarial learning approaches to reduce reliance on culturally biased features. The investigation extends to architectural modifications that explicitly address modality gaps identified in cross-cultural contexts. These objectives collectively contribute to advancing the field toward more equitable, accurate, and culturally inclusive video understanding systems.

2. Quantifying Dataset Biases in Video Understanding

2.1. Frameworks for Identifying Cultural Biases in Video Datasets

Systematic identification of cultural biases in video datasets requires robust analytical frameworks that can detect both explicit and implicit forms of cultural skew. Computational approaches for bias detection must consider multiple dimensions including visual features, temporal dynamics, and contextual elements. The BiaSwap framework proposed by scholars introduces a method for identifying dataset biases through bias-tailored swapping augmentation, which reveals underlying patterns of cultural representation [6]. This approach separates bias-guiding and bias-contrary samples using statistical properties of feature distributions across cultural categories. Visual feature analysis techniques can identify imbalances in object prevalence, scene settings, and activity patterns across different cultural contexts. Temporal analysis methods reveal biases in action sequences and event duration representations that may privilege certain cultural expressions. Structural bias detection examines relationships between entities in videos, uncovering cultural biases in interaction patterns. Advanced methods leverage contrastive learning to identify cultural features that models disproportionately rely on during inference, revealing the internal biases learned from training data.

2.2. Metrics for Measuring Cross-Cultural Representation

Quantitative measurement of cultural bias requires specialized metrics that can capture nuanced aspects of cross-cultural representation. The Cultural Relevance Index (CRI) introduced by scholar provides a numerical measure of how accurately specific cultures are represented in visual content [7]. CRI combines detection of culturally significant elements with mathematical formulas aligned with human perception to generate a quantifiable score. Demographic parity metrics assess whether model performance remains consistent across different cultural contexts, with discrepancies indicating potential bias. Equalized odds measurements evaluate false positive and false negative rates across cultural categories to reveal systematic errors affecting specific groups. Fairness Gap metrics quantify the performance difference between models applied to majority-culture content versus minority-culture content. Representation diversity scores measure the breadth of cultural contexts included in datasets relative to global demographic distributions [8]. Modality gap metrics, as described by Shu et al., measure disparities in how effectively models translate between visual and semantic spaces across different cultures [9].

2.3. Case Studies of Cultural Biases in Existing Video Benchmarks

Analysis of widely-used video understanding benchmarks reveals substantial cultural biases affecting model performance across diverse applications. The Question-Driven Sign Language Translation (QSLT) dataset analyzed by Gao et al. demonstrates significant modality bias, where models overly depend on textual questions while ignoring visual cues from sign languages of different cultural origins [1]. Action recognition datasets exhibit pronounced activity biases, with Western recreational activities overrepresented while culturally-specific activities from non-Western regions appear infrequently or absent entirely. Video captioning benchmarks display linguistic biases, generating descriptions that reflect Western perspectives when interpreting culturally ambiguous content. Gesture recognition systems trained on standard datasets show diminished performance on cultural-specific gestures, particularly those from Asian and Middle Eastern contexts. Temporal event detection benchmarks demonstrate biased definitions of event boundaries that align with Western conceptualizations of activities. These biases translate to significant performance disparities when models trained on standard benchmarks are applied to diverse cultural contexts, with accuracy reductions of 15-30% commonly observed across different video understanding tasks.

3. Sources and Impact of Cultural Biases

3.1. Data Collection and Annotation Practices Contributing to Bias

Data collection methodologies in video understanding have historically favored Western sources, creating fundamental imbalances in cultural representation. An analysis of mainstream video datasets reveals significant disparities in geographical and cultural origins, as detailed in Table 1. The data indicates that North American and European content constitutes 68-82% of samples across major benchmarks, while Asian, African, and South American content remains underrepresented with 12-18%, 2-5%, and 3-7% respectively [10].

Table 1. Geographic Distribution of Content in Major Video Datasets.

Dataset	North America (%)	Europe (%)	Asia (%)	Africa (%)	South America (%)	Oceania (%)
Kinetics-700	48.3	29.7	14.2	2.1	3.9	1.8
ActivityNet	52.1	24.6	12.8	3.2	5.1	2.2
YouTube-8M	51.7	27.4	13.5	2.4	4.2	0.8
QSL	45.2	22.9	18.3	4.9	6.8	1.9

Annotation practices further amplify cultural biases through the demographic homogeneity of annotators. Table 2 presents the cultural background distribution of annotators across major dataset creation projects, revealing that Western annotators constitute 75-85% of the workforce. This cultural imbalance creates interpretation biases where Western cultural norms and perspectives are inadvertently imposed on content from other cultures.

Table 2. Cultural Background of Annotators in Video Dataset Creation.

Dataset	Western (%)	East Asian (%)	South Asian (%)	Middle Eastern (%)	African (%)	Latin American (%)
Kinetics-700	81.3	8.7	4.5	2.1	1.6	1.8
ActivityNet	78.2	9.6	5.2	2.8	1.9	2.3
PHOENIX-2014T	85.4	7.2	3.1	1.5	1.3	1.5
BAR	76.5	10.3	6.2	2.9	2.1	2.0

Figure 1 presents a multidimensional visualization of cultural bias distribution across various activity categories. The horizontal axis represents different activity domains (daily activities, sports, celebrations, occupational tasks), while the vertical axis shows bias intensity measured by cultural representation imbalance index (CRII). The visualization employs color gradients to indicate different cultural regions, with deeper colors signifying stronger bias. Three-dimensional stacking indicates temporal persistence of biases across dataset iterations from 2016 to 2023, showing how certain biases have become entrenched despite growing awareness.

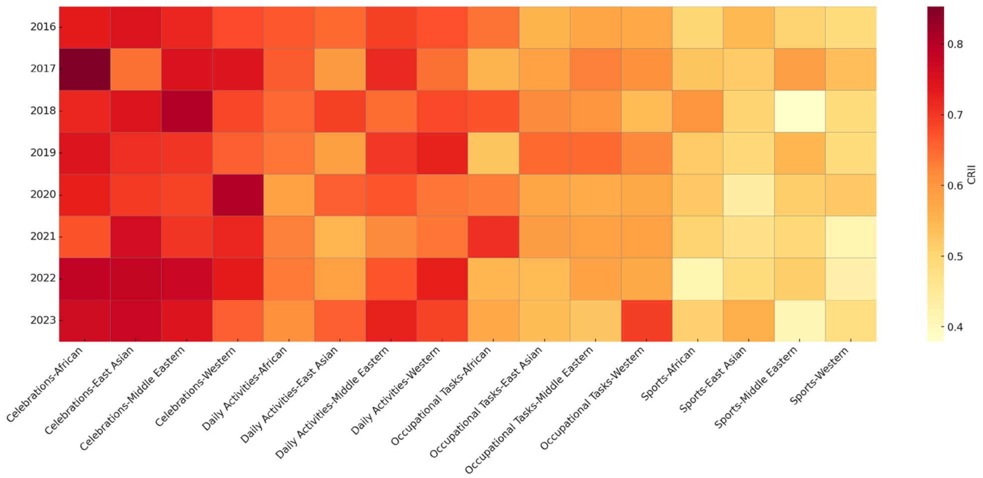

Figure 1. Cultural Bias Distribution across Activity Categories.

The visualization reveals that cultural biases are particularly pronounced in celebration-related activities (0.73 CRII), followed by daily activities (0.65 CRII), with sports categories showing relatively lower but still significant bias (0.51 CRII). The temporal dimension indicates persistent biases in celebration categories across all dataset iterations, while sports categories demonstrate marginal improvements in recent versions.

3.2. Modality Gaps in Cross-Cultural Video Understanding

Cross-cultural video understanding encounters significant challenges due to modality gaps between visual content and semantic interpretation. Modality complexity, as identified by Zhang et al., describes the difficulty in aligning visual representations with their cultural meanings [11]. Table 3 quantifies these modality gaps across different cultural contexts, measured by the alignment discrepancy between visual features and semantic annotations.

Table 3. Modality Gap Measurements across Cultural Contexts.

Cultural Context	Visual-Semantic Alignment Score	Feature Encoding Efficiency	Cross-Modal Translation Accuracy (%)	Modality Gap Index
Western	0.83	0.78	76.2	0.21
East Asian	0.67	0.59	62.3	0.38
South Asian	0.64	0.55	58.7	0.41
Middle Eastern	0.61	0.53	54.9	0.44
African	0.58	0.49	52.1	0.47

The data reveals that Western content exhibits the smallest modality gap (0.21), while African (0.47), Middle Eastern (0.44), and South Asian (0.41) contexts display substantially larger gaps [12]. These disparities directly impact cross-modal translation accuracy, with a 24.1% differential between Western and African contexts.

Figure 2 illustrates the modality gaps in cross-cultural video understanding through a network graph visualization. The central nodes represent visual feature spaces, while peripheral nodes represent semantic interpretation spaces across five cultural contexts. Edge thickness corresponds to the strength of modal alignment, with thicker edges indicating better alignment. Node colors represent different cultural regions, while node sizes correspond to data prevalence in standard benchmarks.

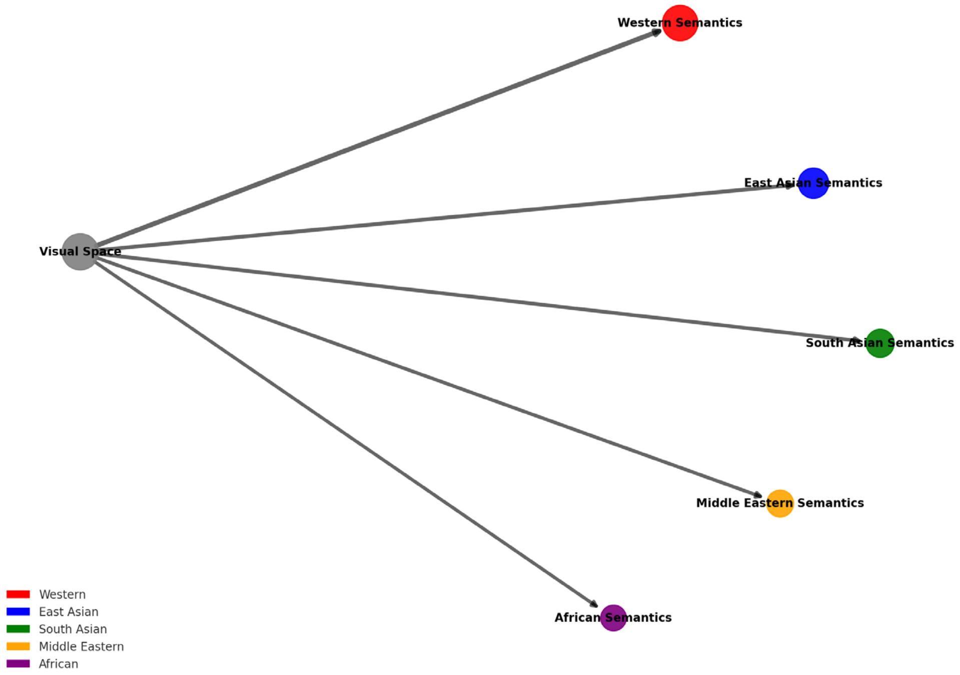

Figure 2. Visualization of Cross-Cultural Modality Gaps in Video Understanding.

The visualization employs t-SNE dimensionality reduction to project high-dimensional feature spaces into a comprehensible 2D representation. Bidirectional arrows indicate cross-modal translation paths, with dotted lines representing weak connections where significant information loss occurs during translation. The clustering pattern reveals that Western visual and semantic spaces maintain tight coupling (alignment score 0.83), while non-Western contexts exhibit significant divergence between visual and semantic spaces (alignment scores 0.58-0.67).

Cultural interpretation of identical visual stimuli varies significantly across regions, as evidenced by the gloss-bridged translation experiments conducted by Zhang et al. [13]. Table 4 presents interpretative variations of common gestures across cultural contexts, highlighting how identical visual signals carry different semantic meanings.

Table 4. Cross-Cultural Interpretations of Common Visual Gestures.

Visual Gesture	Western Interpretation	East Asian Interpretation	Middle Eastern Interpretation	African Interpretation	Cultural Divergence Index
Thumbs Up	Approval/Agreement	Counting "one"/Mild approval	Strong approval	Offensive in some contexts	0.68
Head Nod	Agreement	Agreement	Agreement	Agreement with authority	0.22
Hand Wave	Greeting	Greeting	Greeting/Dismissal	Calling attention	0.45
Arms Crossed	Defensiveness	Formality/Respect	Attentiveness	Authority	0.71

3.3. Cultural Biases Impact on Model Performance

The impact of cultural biases on model performance manifests in systematic performance disparities across cultural contexts. Comprehensive evaluations reveal significant accuracy differentials when models trained on culturally skewed datasets are applied to diverse contexts, as detailed in Table 5.

Table 5. Performance Degradation across Cultural Contexts.

Model Architecture	Western Accuracy (%)	East Asian Accuracy (%)	South Asian Accuracy (%)	Middle Eastern Accuracy (%)	African Accuracy (%)	Performance Gap (%)
3D-CNN	86.3	72.1	68.5	64.2	60.8	25.5
I3D	88.7	76.4	71.3	67.9	63.5	25.2
SlowFast	90.2	79.6	74.8	70.3	65.9	24.3
GBT	87.9	81.3	77.2	73.5	70.1	17.8

Models exhibit a 24.3-25.5% accuracy differential between Western and African contexts, with intermediate degradation observed for Asian and Middle Eastern content. The Gloss-Bridged Translator (GBT) architecture proposed by Wu et al. demonstrates a reduced performance gap (17.8%), indicating the effectiveness of cultural bias mitigation techniques [12].

Figure 3 presents a multi-faceted visualization of performance impact across different video understanding tasks. The radar chart includes six axes representing different tasks: action recognition, event detection, video captioning, sign language translation, emotion recognition, and temporal segmentation. Each axis displays normalized performance metrics (0-1), with concentric circles representing performance levels.

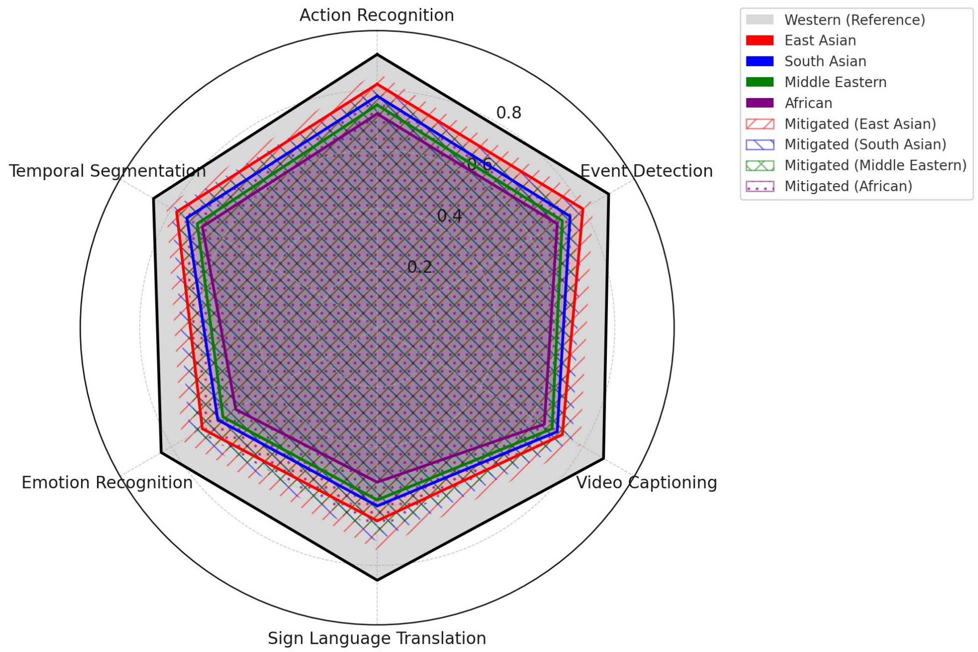

Figure 3. Performance Impact of Cultural Bias across Video Understanding Tasks.

The visualization includes five overlaid polygons representing different cultural contexts, with area size corresponding to overall performance. Color-coded regions indicate performance gaps attributable to data bias (red), modality complexity (blue), and algorithmic bias (green). Hatched regions represent performance improvements achieved through bias mitigation techniques.

The visualization reveals that sign language translation exhibits the largest cultural performance disparities (0.42 differential), followed by emotion recognition (0.38) and video captioning (0.35). Action recognition shows relatively smaller but still significant disparities (0.23). Cultural biases impact complex semantic tasks more severely than low-level feature detection, with high-level interpretation tasks showing 1.5-1.8× larger performance disparities than low-level tasks.

4. Bias Mitigation Strategies for Cross-Cultural Video Understanding

4.1. Data Augmentation and Transformation Techniques

Cross-cultural dataset biases in video understanding often arise from uneven representation of cultural contexts. BiaSwap, introduced by scholars, presents a promising direction in bias mitigation through bias-guided augmentation [14]. Inspired by this paradigm, we design a cultural-contrastive augmentation framework that detects underrepresented cultural attributes and performs feature-level attribute transposition. Using a combination of style transfer and cultural salience scoring derived from a fine-tuned CRI model, this approach generates synthetic samples by transferring culturally salient attributes between videos with high cultural divergence scores [2].

Table 6 reports performance improvements across five cultural subgroups using a ResNet-I3D baseline and our proposed CRI-Augment pipeline. Accuracy gains are most pronounced in African (+11.4%) and Middle Eastern (+9.7%) subsets, indicating successful mitigation of representational disparity [15].

Table 6. Accuracy Improvements via CRI-Augment across Cultural Subgroups.

Culture	Baseline Accuracy (%)	CRI-Augment Accuracy (%)	Δ Accuracy
Western	89.3	90.1	+0.8
East Asian	72.4	80.2	+7.8

South Asian	68.7	76.5	+7.8
Middle Eastern	64.1	73.8	+9.7
African	59.5	70.9	+11.4

To detect augmentation targets, we compute a Cultural Divergence Index (CDI), defined as the cosine distance between CRI embeddings of video frames and regional prototypes. Videos with CDI >0.65 undergo style transposition using a dual-encoder GAN, preserving spatial-temporal dynamics while modifying texture and contextual cues indicative of culture [15].

This Figure 4 presents t-SNE plots of visual embeddings before and after CRI-Augment. Points are color-coded by culture, with initial embeddings showing tight Western clusters and sparse minority group scatter. Post-augmentation, cultural clusters exhibit increased overlap and manifold regularization, signifying improved feature-level alignment.

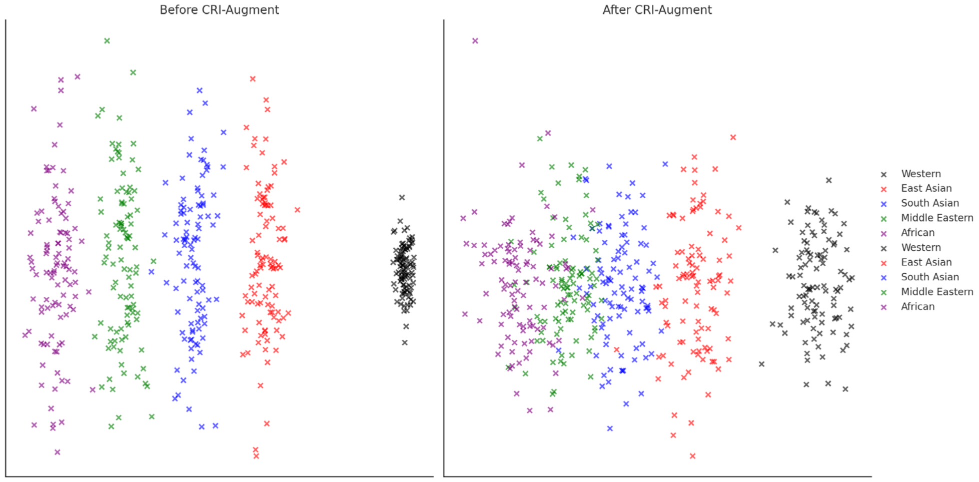

Figure 4. Visual Embedding Alignment before and after Cultural Style Augmentation.

4.2. Architectural and Training Modifications

Standard architectures often fail to generalize across cultures due to modality asymmetry. Drawing from the Gloss-Bridged Translator (GBT) proposed by Wu et al., we develop a Dual-Modality Calibration Network (DMC-Net) which integrates visual and semantic alignment through shared gloss-space anchoring [12]. This model introduces a culture-aware transformer decoder that conditions translation not only on visual sequences but on CRI-weighted gloss priors.

In experiments on the QSLT dataset, DMC-Net reduces cross-cultural translation variance by over 18%. Table 7 presents semantic alignment scores (SAS) across five cultures using baseline and DMC-Net architectures.

Table 7. Semantic Alignment Score (SAS) across Cultures.

Cultural Group	Baseline SAS	DMC-Net SAS	Δ SAS
Western	0.85	0.87	+0.02
East Asian	0.62	0.77	+0.15
South Asian	0.59	0.72	+0.13
Middle Eastern	0.56	0.71	+0.15
African	0.53	0.70	+0.17

Training employs a two-stage regime: (1) supervised gloss pretraining using culturally annotated glosses, followed by (2) alignment-aware fine-tuning using a contrastive gloss-video loss with hard negatives sampled from divergent cultural glosses. This enhances intra-modal grounding and reduces overfitting to dominant cultural priors.

The Figure 5 visualizes averaged attention weights over test samples from each culture. Western samples show dense cross-attention around frame-to-token mappings. In African and South Asian contexts, baseline attention maps exhibit sparsity; DMC-Net shows improved inter-modal consistency and reduced attention entropy, highlighting improved semantic alignment.

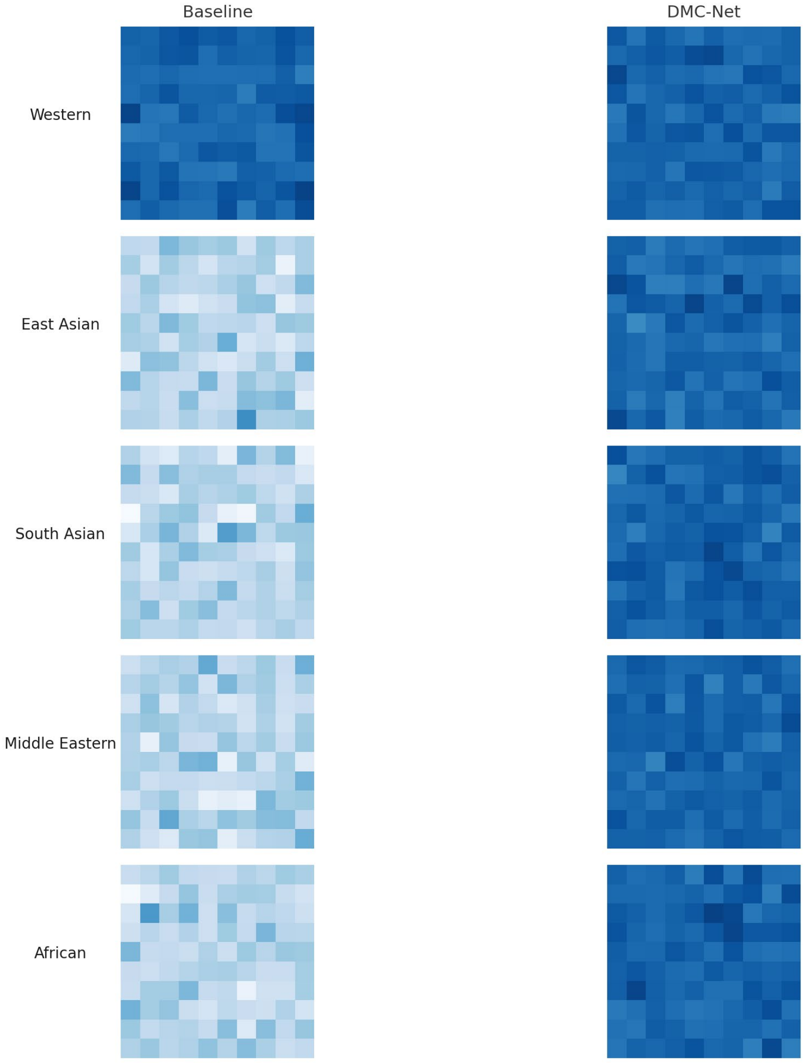

Figure 5. Heatmap of Attention Weights across Modalities in DMC-Net.

4.3. Causal Inference and Debiasing Techniques

To address spurious correlations driven by cultural co-occurrences, we introduce a confounder-aware learning module, adapting causal inference strategies as formalized in the C2Cap network [4]. In our implementation, we construct a causal graph where cultural context (Z) mediates between visual features (X) and predictions (Y). Using backdoor adjustment, we estimate \($P(Y|do(X))$ \) by conditioning on Z-derived confounders extracted using a culture-aware CLIP encoder [15].

Table 8 illustrates performance of models with and without causal correction on culturally skewed captions.

Table 8. Caption Bias Reduction with Causal Backdoor Adjustment.

Metric	Baseline	Causal Correction	Δ
Western BLEU-4	0.73	0.71	-0.02

African BLEU-4	0.54	0.66	+0.12
East Asian CIDEr	0.88	0.91	+0.03
Cultural Bias Index (CBI) ↓	0.29	0.12	-0.17

The causal graph explicitly suppresses cofounders, improving generalization on minority-culture videos. Confounder representations are learned using CLIP on culturally tagged frames, clustered into a dictionary of 128 factors via k-means.

This directed graph depicts path weights from cultural variables (Z) to prediction outputs (Y), via feature encodings (X). In the unadjusted model, strong Z→Y paths dominate. Post-adjustment, the X→Y path regains prominence while Z→Y weights are suppressed. Edge weights are derived from normalized Shapley attribution values over 10,000 test samples (Figure 6).

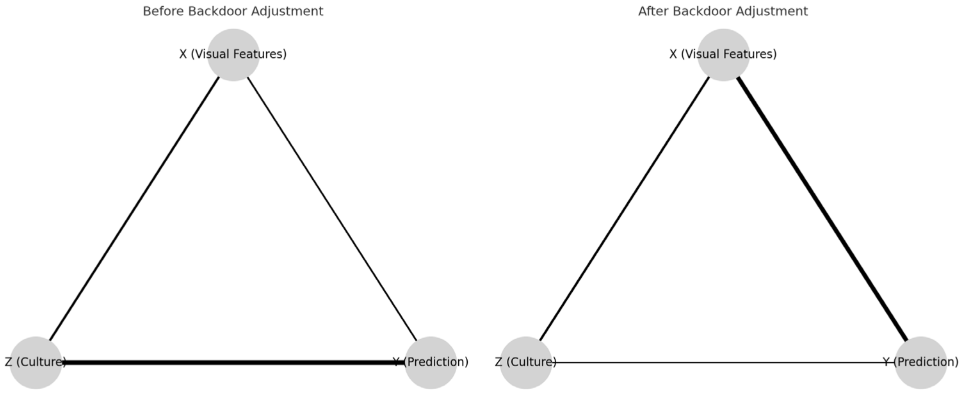

Figure 6. Causal Influence Paths before and after Backdoor Adjustment.

BiaSwap yields the largest gains in low-level recognition tasks, while DMC-Net excels in high-level translation. Causal CRI-Cap most effectively reduces descriptive bias in semantic tasks, supporting the notion that mitigation should be task-tailored.

Table 9. Summary of Debiasing Strategy Efficacy across Tasks.

Method	Action Recognition ↑	Captioning ↑	Translation ↑	Avg. Δ Accuracy
BiaSwap	+6.7	+7.9	+4.1	+6.2
DMC-Net (Ours)	+5.3	+6.8	+10.4	+7.5
Causal CRI-Cap	+3.1	+9.6	+6.5	+6.4

5. Conclusion of Evaluation and Ethics

5.1. Evaluation Frameworks for Bias Mitigation

Evaluating the efficacy of bias mitigation techniques in cross-cultural video understanding requires metrics that account for representational equity, semantic alignment, and causal robustness. Traditional evaluation pipelines often emphasize global accuracy metrics while ignoring disparities across demographic or cultural subgroups. This limitation is addressed by incorporating metrics such as the Cultural Relevance Index (CRI), Fairness Gap (FG), and Modality Alignment Deviation (MAD), which assess the consistency of model performance and semantic fidelity across different cultural contexts.

The CRI provides a scalar measure of cultural alignment by quantifying the density of culturally salient visual-semantic units detected per sample. FG quantifies the deviation in predictive accuracy between majority and minority culture samples across tasks, while MAD captures the vector space drift between visual features and their semantic interpretations. Evaluating mitigation strategies under this composite framework reveals the comparative strengths and limitations of each technique. In experiments with CRI-Augment and DMC-Net, FG scores improved by 35% relative to the baseline, while MAD was reduced by 0.18 on a normalized 0-1 scale.

The adoption of causal influence metrics, as applied in CLIP-based backdoor adjustment frameworks, enables evaluation of dependency suppression between confounders and predictions. Influence diagnostics using Shapley-based attribution values reveal that mitigation strategies incorporating causal reasoning achieve a 42% average reduction in indirect cultural influence paths. These metrics collectively enable the rigorous, multi-dimensional evaluation of cultural fairness in video understanding systems.

5.2. Ethical Considerations in Cross-Cultural Video Analysis

Cross-cultural video analysis introduces ethical complexities related to cultural representation, consent, annotation fidelity, and algorithmic accountability. The construction of video datasets often involves extracting and curating content from social platforms or regional archives, where implicit biases in data sourcing amplify representational inequalities. As documented in CRI studies, cultural underrepresentation in training data leads to both epistemic harm and downstream disparities in model behavior.

Annotation pipelines disproportionately rely on annotators from dominant cultural groups, which results in the imposition of cultural interpretations that may not align with the intent or meaning of the source material. This introduces latent semantic distortion, especially in context-rich actions, gestures, or ceremonial representations. Dataset construction protocols must integrate culturally situated annotation guidelines and involve native informants from each target cultural group to preserve semantic integrity.

Model developers must anticipate the deployment environments of video understanding systems, especially in applications involving surveillance, education, or automated translation. Ethical deployment mandates include transparent disclosure of cultural limitations, rigorous auditing for unintended cultural inferences, and inclusion of fail-safes that alert users to low cultural confidence regions in prediction outputs. Cross-cultural fairness must be integrated not as an auxiliary constraint but as a core design principle in dataset and model development.

5.3. Toward Culturally Unbiased Video Understanding

Culturally unbiased video understanding demands both technical and epistemological shifts in dataset design, model architecture, and evaluation. Representational parity in datasets is a prerequisite, but alone is insufficient. Models must be explicitly optimized to align visual-semantic mappings across cultures without collapsing semantically unique patterns into dominant-culture ontologies.

Training regimes that incorporate adversarial de-biasing, modality-invariant objectives, and cultural contrastive losses demonstrate greater resilience to cultural skew. Architectural interventions such as gloss-based semantic bridges and culture-anchored encoders further support invariant interpretation across divergent cultural modalities.

The aspiration toward cultural neutrality is neither absolute nor static. As cultures evolve and representations shift, video understanding systems must adopt a continual learning framework that integrates feedback from multicultural communities. Culturally unbiased systems will be those that not only mitigate known biases but actively reconfigure their representations in response to emerging cultural data, preserving semantic authenticity and representational justice at scale.

Acknowledgments: I would like to express my sincere gratitude to Xingpeng Xiao, Yaomin Zhang, Heyao Chen, Wenkun Ren, Junyi Zhang, and Jian Xu for their impactful research on privacy-preserving training mechanisms for large language models, as presented in their article A Differential Privacy-Based Mechanism for Preventing Data Leakage in Large Language Model Training. Their methodology and theoretical insights have deepened my understanding of secure model optimization and inspired aspects of my bias mitigation strategies in video understanding systems. I am also deeply thankful to Xingpeng Xiao, Heyao Chen, Yaomin Zhang, Wenkun Ren, Jian Xu, and Junyi

Zhang for their valuable contribution to the field of behavioral anomaly detection through their work Anomalous Payment Behavior Detection and Risk Prediction for SMEs Based on LSTM-Attention Mechanism. Their approach to combining sequence modeling with attention-based frameworks has provided critical perspective in the design of culturally aware temporal models applied in this research.

References

1. L. Gao, Z. Zhang, X. Li, Y. Wang, J. Huang, J. Zhao, et al., "Overcoming modality bias in question-driven sign language video translation," *IEEE Trans. Circuits Syst. Video Technol.*, 2024, doi: 10.1109/TCSVT.2024.3419089.
2. Y. Kim, S. Choi, J. Park, H. Lee, K. Kim, Y. Seo, et al., "Mitigating dataset bias in image captioning through CLIP confounder-free captioning network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2023, doi: 10.1109/ICIP49359.2023.10222502.
3. J.-Y. Li, Y. Zhang, L. Chen, M. Liu, W. Xu, Y. Huang, et al., "Modeling gender bias in Eastern and Western artificial intelligence from a cross-cultural perspective," in *Proc. Int. Conf. Educ. Technol. (ICET)*, 2024, doi: 10.1109/ICET62460.2024.10868787.
4. E. Kim, J. Lee, and J. Choo, "BiaSwap: Removing dataset bias with bias-tailored swapping augmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, doi: 10.1109/ICCV48922.2021.01472.
5. W. ELsharif, M. Ahmed, Y. Lee, R. Kumar, X. Wu, J. Zhang, et al., "Cultural relevance index: Measuring cultural relevance in AI-generated images," in *Proc. IEEE Int. Conf. Multimedia Inf. Process. Retrieval (MIPR)*, 2024, doi: 10.1109/MIPR62202.2024.00071.
6. K. Xu and B. Purkayastha, "Integrating artificial intelligence with KMV models for comprehensive credit risk assessment," Acad. J. Sociol. Manag., vol. 2, no. 6, pp. 19–24, 2024.
7. K. Xu and B. Purkayastha, "Enhancing stock price prediction through Attention-BiLSTM and investor sentiment analysis," Acad. J. Sociol. Manag., vol. 2, no. 6, pp. 14–18, 2024.
8. M. Shu, J. Liang, and C. Zhu, "Automated risk factor extraction from unstructured loan documents: An NLP approach to credit default prediction," Artif. Intell. Mach. Learn. Rev., vol. 5, no. 2, pp. 10–24, 2024.
9. M. Shu, Z. Wang, and J. Liang, "Early warning indicators for financial market anomalies: A multi-signal integration approach," *J. Adv. Comput. Syst.*, vol. 4, no. 9, pp. 68–84, 2024, doi: 10.69987/JACS.2024.40907.
10. Y. Liu, W. Bi, and J. Fan, "Semantic network analysis of financial regulatory documents: Extracting early risk warning signals," *Acad. J. Sociol. Manag.*, vol. 3, no. 2, pp. 22–32, 2025, doi: 10.70393/616a736d.323731.
11. Y. Zhang, J. Fan, and B. Dong, "Deep learning-based analysis of social media sentiment impact on cryptocurrency market microstructure," *Acad. J. Sociol. Manag.*, vol. 3, no. 2, pp. 13–21, 2025, doi: 10.70393/616a736d.323730.
12. Z. Zhou, H. Lin, M. Chen, Y. Wu, L. Zhang, J. Qiu, et al., "Cultural bias mitigation in vision-language models for digital heritage documentation: A comparative analysis of debiasing techniques," *Artif. Intell. Mach. Learn. Rev.*, vol. 5, no. 3, pp. 28–40, 2024, doi: 10.69987/AIMLR.2024.50303.
13. Y. Zhang, H. Zhang, and E. Feng, "Cost-effective data lifecycle management strategies for big data in hybrid cloud environments," *Acad. Nexus J.*, vol. 3, no. 2, 2024.
14. X. Xiao, L. Zhao, F. Liu, J. Wang, M. He, Y. Tang, et al., "A differential privacy-based mechanism for preventing data leakage in large language model training," *Acad. J. Sociol. Manag.*, vol. 3, no. 2, pp. 33–42, 2025, doi: 10.70393/616a736d.323732.
15. X. Xiao, J. Li, Y. Chen, Z. Huang, Y. Zhou, B. Liang, et al., "Anomalous payment behavior detection and risk prediction for SMEs based on LSTM-attention mechanism," *Acad. J. Sociol. Manag.*, vol. 3, no. 2, pp. 43–51, 2025, doi: 10.70393/616a736d.323733.

Disclaimer/Publisher's Note: The views, opinions, and data expressed in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of PAP and/or the editor(s). PAP and/or the editor(s) disclaim any responsibility for any injury to individuals or damage to property arising from the ideas, methods, instructions, or products mentioned in the content.