CultureShift: Mapping Temporal Cultural Evolution in Vision-Language Models

Gautam Jajoo[1] Harsh Deshpande[1] Hamna [2] Pranjal Chitale [2]
[1]BITS Pilani, India [2]Microsoft Research
gautamjajoo1729,harshdeshpande2010,hamnaabid@gmail.com

Abstract

As vision-language models (VLMs) become embedded in global technologies, ensuring that they are culturally aware is critical for fairness, representation, and societal relevance. Yet, current benchmarks for evaluating cultural competence treat culture as static, overlooking the fact that cultural norms, aesthetics, and values evolve over time. In this work, we introduce the concept of ***Temporal Cultural Awareness****— the capacity of AI models to recognize and adapt to shifting cultural representations across decades. To operationalize this concept, we present a novel evaluation framework grounded in cinema, leveraging film media as a time-aligned, globally resonant proxy for cultural evolution. We curate* ***CineCulture****, a dataset of annotated movie screenshots from Hollywood and Bollywood films spanning multiple decades, capturing fine-grained, visually evident cultural attributes across themes like clothing, architecture, gender roles, and leisure. This dataset enables systematic assessment of how well VLMs reflect evolving cultural signals, both geographically and temporally. Our contributions include a new benchmark task, proposed evaluation metrics, and an empirical analysis revealing that popular VLMs often fail to track temporal cultural shifts. Our work calls for a new dimension of evaluation in culturally competent AI: not only geographic inclusivity but* ***temporal inclusivity*** *as well. The link to the code and dataset can be found https://github.com/gautamjajoo/TemporalCultureShift*

1. Introduction

As Vision-Language models (VLMs) become increasingly integrated into global applications—from content generation to educational tools and recommendation systems — their ability to operate across diverse cultural contexts has never been more critical. Culturally aware models can foster inclusivity, build user trust, and ensure relevance across diverse environments [10]. In contrast, the failure to account for cultural nuances risks perpetuating bias, marginalizing underrepresented communities, and ultimately undermining both the fairness and efficacy of these technologies [4, 9, 14, 18, 20].

Although recent studies recognize this and to a certain extent assess cultural awareness in VLMs, these efforts largely approach culture as a static phenomenon, evaluated at a single point in time or in a time-agnostic manner using aggregated datasets that obscure temporal nuance [7, 17]. However, culture is inherently dynamic. Social values, norms, aesthetics, and roles shift across decades in response to political movements, economic developments, technological change, and intergenerational ideologies. Moreover, a biased or outdated understanding of a culture can amplify harmful stereotypes. A culturally aware model must therefore recognize not only cultural diversity across geographies but also its evolution across time. We introduce the concept of **Temporal Cultural Awareness** - the ability of AI model to recognize, interpret, and adapt to cultural representations as they evolve over time. For instance, transformations in how themes such as family structure, gender roles, fashion, or leisure are visually and narratively represented over decades provide rich signals about cultural evolution. Yet, most VLMs, trained on temporally unaligned or aggregated visual data, are not equipped to detect or adapt to such longitudinal shifts.

To address this gap, we propose a novel framework for evaluating temporal cultural awareness in VLMs by leveraging film media as a proxy for capturing the evolving cultural expression. *Cinema offers a compelling medium for this purpose: it is both globally influential and temporally rich, reflecting and shaping societal values across generations* [2, 22]. By analyzing visual content from films across different eras, we create a test bed to systematically evaluate how well can VLMs capture temporal shifts in cultural representation.

This study investigages the central research question of whether VLMs can distinguish about how cultural representations have evolved over time?

Our contributions are as follows:

- We formalize Temporal Cultural Awareness as a critical and previously underexplored dimension in evaluating AI cultural competence.

- We demonstrate the viability of using cinema-derived visual data as a scalable, time-aligned resource for studying cultural evolution in VLMs.
- We intend to publicly release the CineCulture Dataset and our evaluation framework, facilitating future research to this end.

In doing so, our work calls attention to an essential frontier in culturally responsive AI: temporal inclusivity. Just as AI systems must understand and respect cultural diversity across geographies, they must also remain sensitive to the evolving nature of cultural expression across generations. Without this capacity, models risk reinforcing outdated assumptions, misrepresenting communities, and failing to serve the needs of an ever-changing world.

2. Related Work

Recent advances in VLMs have spurred a growing interest in evaluating their cultural awareness. This is highighted by the development of benchmarks such as CulturalVQA [16], CVQA [19], All Languages Matter (ALM) Bench [21], and GlobalRG [5], which primarily employ the visual question answering (VQA) tasks to evaluate models on their ability to recognize and reason about culturally grounded elements such as traditional clothing, rituals, food, and everyday practices.

Beyond question-answering, recent efforts have explored generative evaluations of cultural competence. Benchmarks like CUBE [11] test text-to-image generation across domains such as cuisine, landmarks, and art from eight countries. While, DALL-E Street [15] uses culturally diverse household scenes to assess visual representation. Comprehensive efforts like CultureVLM [13] expand these evaluations to more than 100 countries, and benchmarks like K-ViScuit [3] integrate human-in-the-loop evaluation to assess cultural appropriateness in visual scenes.

Complementing these datasets, multiple metrics like Cultural Awareness Score (CAS) [6], diversity@k in GlobalRG and LAVE[16] assess cultural awareness in captions, retrieval, and VQA tasks. However, these benchmarks adopt a static view of culture, capturing representations at a single time point and overlooking temporal shifts. This limits their ability to evaluate AI performance in dynamic, time-sensitive cultural contexts.

In parallel, the VLM and video understanding communities have introduced several movie-based benchmarks, aimed primarily at long-form narrative comprehension. Datasets like MoVQA [24] evaluate models on long-form narrative comprehension through the visual question answering task, while SF20K [8] extends this effort with a larger-scale dataset focused on story-level video QA. MovieBench [23] offers hierarchical annotations across full-length films — at the summary, scene, and shot levels—supporting structured understanding and character-consistent generation. Among efforts intersecting with temporal reasoning, VITATECS [12] introduces a diagnostic benchmark for understanding temporal concepts in video-language models using movie data. However, these movie-based benchmarks largely focus on temporal aspects like coherence, story flow, and character tracking, rather than the cultural implications embedded in visual content.

Crucially, none of these existing benchmarks whether culturally or temporally oriented—explicitly examine how cultural representations evolve over time. This gap leaves unaddressed a key dimension of AI cultural intelligence: Temporal Cultural Awareness. Our work addresses this gap by introducing a new benchmark situated at the intersection of cultural understanding and temporal analysis. Leveraging cinema as a rich, longitudinal record of societal values and norms, we curate a dataset of film imagery spanning multiple decades to study cultural evolution through visual narratives. Unlike prior efforts, our benchmark is designed to evaluate how well VLMs can detect, interpret, and adapt to the temporal shifts in cultural expression. This enables a new class of evaluations that move beyond a static snapshot of cultural representation to assess AI's capacity to understand culture as a dynamic, evolving phenomenon—an essential step toward building AI systems that are both temporally inclusive and globally competent.

Benchmark name	Focus	Culture	Temporal
CulturalVQA	VQA	Yes	No
GlobalRG	RAG	Yes	No
AK-ViScuit	Interpretation	Yes	No
ALM Bench	Multimodal	Yes	No
CVQA	Multimodal	Yes	No
CUBE	Image	Yes	No
MaRVL	Multimodal	Yes	No
CineCulture (Ours)	VQA	**Yes**	**Yes**

Table 1. Existing Benchmarks for Cultural Awareness in VLMs

3. Methodology

3.1. Creating the CineCulture Dataset

Art has long served as a mirror of societal norms and values, with film acting as a particularly rich medium for capturing cultural transformations over time. To enable the quantitative study of such shifts in visual media, we construct a curated dataset comprising carefully selected movie screenshots. These images span a wide range of historical periods, geographic locations, and cultural settings, offering a diverse and representative set of ground-truth visual data.

To structure this dataset for systematic analysis, we develop a cultural taxonomy encompassing key visual categories indicative of cultural identity. These are organised

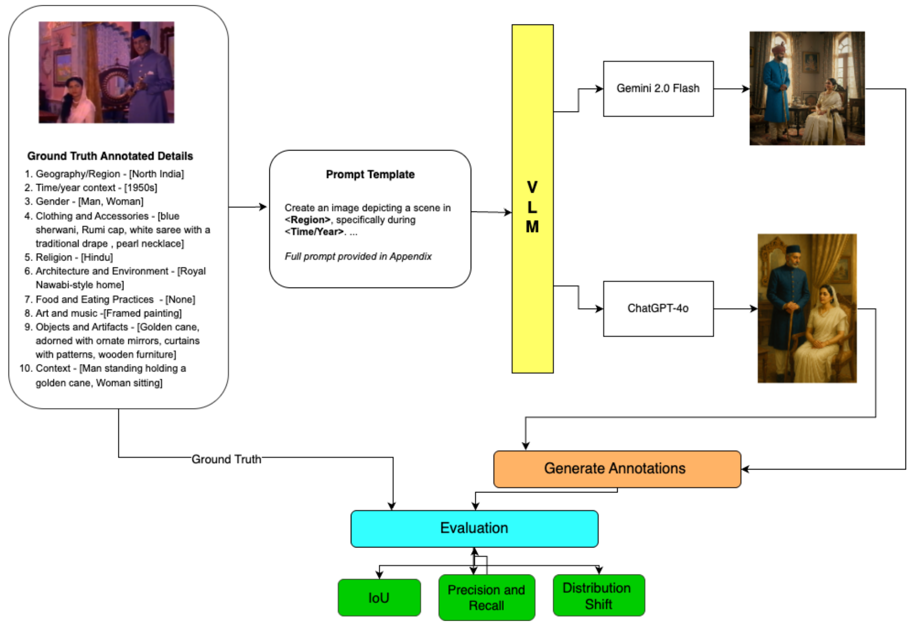

Figure 1. Workflow with the evaluation of temporal cultural awareness in VLMs from ground truth annotation and prompt generation to image creation, followed by annotation and assessment

into two primary classes: Demographic Proxies (DP), attributes linked to population identity and Semantic Proxies (SP), as outlined in [1], which capture cultural aesthetics, practices, and belief systems. The taxonomy of nine high-level categories guiding the dataset's structure are:

1. Geography/Region (DP)
2. Gender
3. Clothing and Accessories (incorporating both DP/SP aspects)
4. Architecture and Environment (DP)
5. Food and Eating Practices (SP)
6. Religion (DP)
7. Art and Music (including Dance forms, Musical instruments, Festivals) (SP)
8. Time/Year context
9. Objects and Artifacts (SP)

This categorical framework allows for a systematic approach to studying cultural representation across different facets of visual scenes.

3.2. Annotation

The CineCulture dataset employs a rigorous annotation process to capture fine-grained cultural nuances. Each image is labeled using a structured one-hot encoding scheme, where cultural attributes are grouped into predefined classes (e.g., *Headwear*, *Footwear*, *Building Style*), and each vector dimension corresponds to a discrete, exhaustively defined attribute (e.g., 'Mojaris', 'Geta sandals'). Trained human annotators follow standardized guidelines to ensure consistency and cultural fidelity across annotations, which are nested within the broader taxonomy defined during dataset construction.

1. Clothing and Accessories: Specific garment types (Sherwani, saree, jeans, Kimono), symbolic colors/patterns, jewelry (Nose rings, Mangalsutra), headwear (Turbans, Hijabs, Sombreros), and footwear (Mojaris, Geta sandals).
2. Architecture and Environment: Housing styles (Traditional, modern), setting (Urban vs. Rural), landscaping features (Gardens, marketplaces), construction materials (Wood, stone), design patterns (Islamic geometric, Colonial arches), and transportation modes (Rickshaws, Camels, Bullet trains).
3. Food and Eating Practices: Specific food types (Sushi, Thali meals, Tacos), dining styles (Floor seating, chopsticks), and related household items (Brass utensils, Tatami mats).
4. Religion: Identifiable religious items (Statues, prayer beads) and structures (Temples, Mosques, Churches).
5. Art and Music: Recognizable dance forms, specific musical instruments, and indicators of festivals.
6. Objects and Artifacts: Culturally specific tools/utensils, depicted technology levels, logos/emblems (Flags, symbols), and visible written languages/scripts.

This human-in-the-loop process ensures high-fidelity, multi-attribute cultural labeling suitable for comprehensive visual cultural analysis.

3.3. Evaluation

3.3.1. IoU, Precision and Recall on the One-Hot Vectors

To rigorously assess the performance of our cultural attribute detection system on movie screenshots, we employ three standard evaluation metrics: Intersection over Union (IoU), Precision and Recall. Our ground-truth dataset, annotated with one-hot vectors indicating the presence of various cultural artifacts, serves as the reference for these assessments.

1. **Intersection over Union (IoU)**
2. **Precision**
3. **Recall**

3.3.2. Measuring Temporal Distribution Shifts

To evaluate how well Vision-Language Models (VLMs) capture the temporal evolution of cultural elements in cinema, we introduce a framework that compares the distribution of cultural attributes over time between real movie shots and VLM-generated images.

The pipeline consists of the following steps:

1. **Context Extraction:** From a curated dataset of movie shots s_{gt} spanning various eras and regions, we extract contextual information c—including activity description, country, and time period.
2. **Image Generation:** Using c as input, the VLM generates an image s_{gen} corresponding to each s_{gt}.
3. **Cultural Annotation:** Both s_{gt} and s_{gen} are annotated with a binary vector $v \in \{0,1\}^N$ indicating the presence of N predefined cultural attributes (e.g., fashion, food, architecture).
4. **Temporal Distribution Estimation:** For each attribute i, we compute its empirical distribution across time bins (e.g., decades) for both ground truth ($D_{gt,i}$) and generated images ($D_{gen,i}$), reflecting its frequency over time.
5. **Significance Testing:** A χ^2 goodness-of-fit test compares $D_{gt,i}$ and $D_{gen,i}$, yielding p-values p_i to assess whether the temporal distributions differ significantly.
6. **Divergence Scoring:** For attributes with $p_i < \alpha$ (e.g., $\alpha = 0.05$), we compute the Jensen-Shannon Divergence (JSD) between $D_{gt,i}$ and $D_{gen,i}$. We then calculate an overall score:

$$\mathcal{S} = \sum_{i=1}^{N} \mathbb{I}(p_i < \alpha) \cdot (1 - p_i) \cdot JSD(D_{gt,i} \,\|\, D_{gen,i})$$

where $\mathbb{I}(\cdot)$ is the indicator function. A lower $\mathcal{S}$ indicates closer alignment between VLM-generated outputs and historical ground truth data.

This method enables fine-grained, temporal analysis of cultural fidelity in VLMs, identifying both broad trends and specific eras or attributes where the model may exhibit biases or inaccuracies.

4. Discussion

While our research remains in progress, the methodological approach offers several promising areas for understanding how VLMs conceptualize and represent cultural elements across temporal and geographical contexts.

The IoU metric in our evaluation will offer pointers into how VLMs perceive and replicate distinct cultural components. This quantitative approach is useful for identifying which cultural features are accurately captured and which tend to be consistently neglected. Initial findings indicate that general characteristics, such as gender presentation and basic spatial arrangement, often yield higher IoU scores. In contrast, nuanced cultural elements like the traditional draping of a saree (Fig 3) or the authentic design of a Rumi cap tend to be less accurately recognized.

The most compelling insights may arise from elements with particularly low IoU scores, as these often highlight cultural blind spots in current VLM systems. These overlooked features frequently encompass culturally rich details, such as symbolic objects, specific traditional attire, or architectural motifs, that hold deep cultural significance.

We also analyze various kinds of shifts like:

1. **Temporal modernization**: VLMs may introduce contemporary elements into historical settings, such as modern architectural features in representations of 1950s homes.
2. **Cultural homogenization**: Models might blend distinctive cultural elements, replacing historically accurate elements with more generalized representations.
3. **Western-centric normalisation**: The distribution analysis may reveal tendencies to subtly westernise non-Western cultural contexts, particularly in spatial arrangements, posture, or stylistic elements.

5. Conclusion

We propose a dataset and evaluation method aimed at addressing the critical yet underexplored capability of Vision-Language Models to comprehend the evolution of culture over time through visual media. We intend to create CineCulture, a novel dataset curated from chronologically diverse movie screenshots, providing a unique benchmark. Future work should focus on expanding the dataset's scope, by including more demographies, incorporating multimodal information (like dialogue or sound) and benchmarking on multiple SoTA models. Ultimately bridging this gap is essential for creating AI systems that possess a deeper, more historically and culturally informed understanding of the human experience as represented visually.

References

[1] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in llms: A survey, 2024. 3

[2] Shirshendu Ganguli Anshuman Mohanty, Aditi Mudgal. Mapping movie genre evolution (1994 - 2019) using the role of cultural and temporal shifts: a thematic analysis, 2023. 1

[3] Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration, 2024. 2

[4] Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models, 2023. 1

[5] Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[6] Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. How culturally aware are vision-language models?, 2025. 2

[7] Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. Eticor: Corpus for analyzing llms for etiquettes, 2023. 1

[8] Ridouane Ghermi, Xi Wang, Vicky Kalogeiton, and Ivan Laptev. Long story short: Story-level video understanding from 20k short films, 2025. 2

[9] Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach, 2024. 1

[10] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural nlp, 2022. 1

[11] Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models, 2025. 2

[12] Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models, 2024. 2

[13] Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries, 2025. 2

[14] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022. 1

[15] Anjishnu Mukherjee, Ziwei Zhu, and Antonios Anastasopoulos. Crossroads of continents: Automated artifact extraction for cultural adaptation with large multimodal models, 2024. 2

[16] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding, 2024. 2

[17] Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond, 2024. 1

[18] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence, 2022. 1

[19] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024. 2

[20] Adhithya Prakash Saravanan, Rafal Kocielnik, Roy Jiang, Pengrui Han, and Anima Anandkumar. Exploring social bias in downstream applications of text-to-image foundation models, 2023. 1

[21] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Asma Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, San-

jay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Amirudin, Muhammad Ridzuan, Daniya Kareem, Ketan More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Fabian Farestam, Muztoba Rabbani, Sanoojan Baliah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Xavier, Amit Bhatkal, Hawau Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman Khan, and Fahad Khan. All languages matter: Evaluating lmms on culturally diverse 100 languages, 2025. 2

[22] J. Virdi. *The Cinematic ImagiNation [sic]: Indian Popular Films as Social History*. Rutgers University Press, 2003. 1

[23] Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation, 2025. 2

[24] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding, 2023. 2

A. Prompt Template

The Prompt Template for feeding the data from the annotated ground truth image to the VLM is as follows:

Prompt Template

Create an image depicting a scene in <Region>, specifically during <Time/Year>. The genders of the people involved in the image are listed as <Gender>. The pertinent context for the setting of this image is <Context>.

B. Pictoral Depiction of Cultural Shifts Over Time

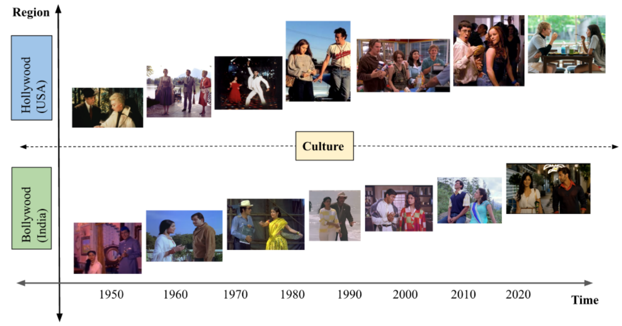

Figure 2. Pictoral Depiction of Cultural Shifts Over Time

C. Workflow with the evaluation for an Example of an Image from a Hollywood Movie

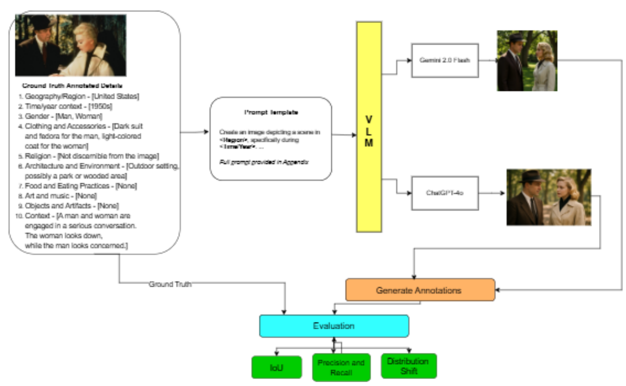

Figure 3. Workflow for an example image taken from a Hollywood Movie