Are Video Generation Models Geographically Fair? An Attraction-Centric Evaluation of Global Visual Knowledge

Anonymous ACL submission

Abstract

Recent advances in text-to-video generation have produced visually compelling results, yet it remains unclear whether these models encode geographically equitable visual knowledge. In this work, we investigate the geo-equity and geographically grounded visual knowledge of text-to-video models through an attraction-centric evaluation. We introduce **G**eo-**A**ttraction Landmark **P**robing (GAP), a systematic framework for assessing how faithfully models synthesize tourist attractions from diverse regions, and construct GEOATTRACTION-500, a benchmark of 500 globally distributed attractions spanning varied regions and popularity levels. GAP integrates complementary metrics that disentangle overall video quality from attraction-specific knowledge, including global structural alignment, fine-grained keypoint-based alignment, and vision–language model judgments, all validated against human evaluation. Applying GAP to the state-of-the-art text-to-video model Sora 2, we find that, contrary to common assumptions of strong geographic bias, the model exhibits a relatively uniform level of geographically grounded visual knowledge across regions, development levels, and cultural groupings, with only weak dependence on attraction popularity. These results suggest that current text-to-video models express global visual knowledge more evenly than expected, highlighting both their promise for globally deployed applications and the need for continued evaluation as such systems evolve.

1 Introduction

Recent advances in large-scale text-to-video models have enabled the synthesis of visually compelling and temporally coherent videos from natural language descriptions (Xie et al., 2025; Xing et al., 2024; Ma et al., 2026). Models such as Kling (Kling AI, 2026), Veo 3 (Google AI Studio, 2025), and Sora 2 (OpenAI, 2025b) demonstrate strong capabilities in rendering complex scenes, dynamic motion, and cinematic styles. As these systems are increasingly deployed in global contexts, an important question arises: *Do video generation models possess uniform geographic and cultural knowledge, or do they privilege regions that are more prevalent in their training data?*

Existing evaluations of AI-generated videos primarily focus on visual quality, temporal coherence, and instruction alignment (Liu et al., 2024a; He et al., 2024; Zheng et al., 2025; Liu et al., 2024b; Wu et al., 2023b; Xiang et al., 2025). While important, these evaluations do not directly assess whether generated videos accurately reflect region-specific visual knowledge. Meanwhile, studies of geo-bias in Large Language Models (LLMs) have shown that model knowledge often skews toward the Global North and West (Manvi et al., 2024b; Decoupes et al., 2024; Lalai et al., 2025a). However, analogous investigations for text-to-video models remain limited, despite video requiring richer spatial, environmental, and fine-grained visual understanding (Singh, 2023; Chen et al., 2025).

In this work, we study geo-equity and geographically grounded visual knowledge in text-to-video models. Unlike prior studies that evaluate cultural understanding through regional customs, festivals, or events (Chen et al., 2025), which can be ambiguous or stereotypical, we focus on tourist attraction landmarks. These landmarks are globally recognizable, tied to specific geographic identities, well documented in real-world imagery, and characterized by distinctive architectural and environmental cues, making them well suited for objective evaluation of regional visual knowledge (Merciu et al., 2020).

We introduce **G**eo-**A**ttraction Landmark **P**robing (GAP), the first framework for systematically evaluating how faithfully text-to-video models synthesize tourist attractions across diverse regions. We construct GEOATTRACTION-500, a benchmark of

500 tourist attractions derived from the Google Landmarks Dataset v2 (Ramzi et al., 2023) and enriched with attraction-level geographic and popularity metadata. This benchmark enables analysis across both globally prominent landmarks and regionally significant sites.

To evaluate generated videos, GAP combines complementary metrics that disentangle general video quality from attraction-specific visual knowledge. We use a reference-free quality metric (Liu and Zhang, 2025) alongside reference-based knowledge metrics (Radford et al., 2021b), including global structural alignment, our proposed keypoint-based local alignment metric for fine-grained geometric and textural fidelity, and VLM-as-a-Judge evaluation (Gu et al., 2025). Together, these metrics provide a fine-grained assessment of whether generated videos preserve the distinctive visual identity of each attraction.

Through experiments on Sora 2 (OpenAI, 2025b), we validate our metrics against human judgments and analyze the model's geographic visual knowledge. Contrary to the common assumption of strong bias toward the Global North (Odeh, 2010) and Global West (Turner and Khondker, 2010), our results show that Sora 2 maintains a relatively uniform level of visual knowledge across regions and popularity levels. We further find that detailed prompts improve fine-grained structural details, while the model remains robust even under short instructions.

Our contributions are summarized as follows:

- We propose GAP, the first systematic framework for evaluating geo-equity and geographically grounded visual knowledge in text-to-video generation.
- We construct GEOATTRACTION-500, a benchmark of 500 attraction landmarks spanning diverse regions, socio-cultural contexts, and popularity levels.
- We introduce complementary automatic metrics, including a keypoint-based local alignment metric, and show that they align with human judgments.
- Through experiments on Sora 2, we find that the model exhibits a largely uniform level of visual knowledge across regions and remains robust to prompt verbosity.

2 Related Works

2.1 AI-Generated Video Evaluation

AI-Generated Video Evaluation (AIGVE) extends traditional video quality assessment by considering both human perception and instruction alignment (Liu et al., 2024a). Perception-oriented methods evaluate visual fidelity, realism, and temporal coherence, ranging from statistical metrics such as FVD and Inception Score (Unterthiner et al., 2019; Barratt and Sharma, 2018) to learned evaluators such as DOVER (Wu et al., 2023c) and CLIP-based BVQI (Wu et al., 2023a). Instruction-aligned methods instead measure text–video consistency using metrics such as CLIPScore (Hessel et al., 2022) and TIFA (Hu et al., 2023), but often overlook perceptual degradation and physical implausibility.

Recent frameworks aim to unify these aspects through multi-dimensional evaluation. VBench (Huang et al., 2023) and EvalCrafter (Liu et al., 2023b) aggregate quality and alignment dimensions, while VideoScore (He et al., 2024), VideoScore2 (He et al., 2025), and AIGVE-MACS (Liu and Zhang, 2025) leverage vision–language models and human annotations for more interpretable, human-aligned assessment. However, existing AIGVE methods mainly evaluate overall quality and prompt adherence, and remain largely agnostic to whether generated videos faithfully capture structured real-world or region-specific visual knowledge.

2.2 Geo-Bias and Global Representation in Generative Models

Geographic bias and uneven global representation have been extensively documented in LLMs. Prior studies show that LLMs systematically favor entities, knowledge, and perspectives from the Global North and Global West, exhibiting higher accuracy and richer representations for Western regions while underperforming on or misrepresenting regions from the Global South (Manvi et al., 2024a; Lalai et al., 2025b; Tao et al., 2024). These disparities have been observed across tasks such as entity deduction, factual reasoning, and cultural judgment, indicating that geo-bias is a persistent and structural issue rooted in training data distributions (Manvi et al., 2024a).

In contrast, geo-bias in generative video models has received far less attention. Existing work such as T2VWorldBench (Chen et al., 2025) represents an early attempt to evaluate world knowledge

in text-to-video generation, but relies largely on culturally stereotypical prompts and coarse qualitative judgments, making systematic and objective evaluation difficult. As a result, there remains no systematic, large-scale evaluation of how fairly text-to-video models represent diverse regions of the world, motivating the need for geographically grounded and structurally measurable evaluation protocols.

3 Geo-Attraction Landmark Probing

To systematically investigate the geo-equity and geographically grounded visual knowledge of text-to-video models through their visual outputs, we propose a novel evaluation framework termed **G**eo-**A**ttraction Landmark **P**robing (GAP). This framework examines whether a video generation model can faithfully synthesize tourist attractions that capture the architectural, environmental, and socio-cultural characteristics of their corresponding locations.

We focus on tourist attractions because they provide globally recognizable and visually distinctive proxies for regional knowledge. Such landmarks are closely tied to specific geographic identities, extensively documented in real-world imagery, and characterized by rich visual signals spanning architecture, landscape, and socio-cultural context. Consequently, an attraction-centric evaluation offers a controlled yet expressive setting for assessing whether a model exhibits uneven geographic knowledge.

In GAP, we prompt a video generation model to synthesize short videos of tourist attractions from specified countries or regions. Generated videos are first evaluated using AIGVE-MACS (Liu and Zhang, 2025) to assess overall visual quality, and are then compared against real-world reference images using three complementary knowledge-oriented metrics: Patch-Level CLIP (Radford et al., 2021b) for global structural alignment, our novel Keypoint-Based Local Alignment for fine-grained structural fidelity, and VLM-as-A-Judge (Gu et al., 2025) for semantic alignment, as illustrated in Figure 1.

By applying a unified evaluation pipeline across geographically diverse attractions, GAP enables systematic analysis of whether text-to-video models exhibit region-dependent disparities or instead demonstrate a consistent level of globally grounded visual knowledge. In the remainder of this section, we describe the construction of GEOATTRACTION-500 benchmark dataset in Section 3.1 and detail the evaluation metrics used in GAP in Section 3.2.

3.1 GEOATTRACTION-500 Benchmark Dataset

We construct our benchmark dataset, GEOATTRACTION-500, based on the Google Landmarks Dataset v2 (GLDv2)[1] (Ramzi et al., 2023) and augment it with attraction-level metadata from the Google Landmarks Places dataset[2]. This combined dataset provides landmark images paired with attraction names and corresponding city, country, and regional geographical information.

To capture variation in global visibility, we incorporate Wikipedia page view statistics for each attraction landmark as a proxy for attraction popularity. Page view counts reflect how frequently an attraction landmark is accessed by users worldwide and enable analysis of model performance across attractions with varying levels of global prominence.

To ensure that each reference ground-truth image faithfully captures the defining characteristics of an attraction and its surrounding environment, we employ human annotators to manually select a representative image for each attraction from the GLDv2 dataset. These ground-truth images are chosen to reflect the canonical visual appearance of the landmark and serve as reference visuals for subsequent comparisons.

Based on each selected ground-truth image, we further generate textual instructions for video generation. Specifically, we prompt GPT-5.1 (OpenAI, 2025a) to produce a structured textual description of the scene conditioned on the attraction's name and location. The output includes (i) a detailed cinematic caption of 3–6 sentences describing composition, camera viewpoint, environmental and architectural details, lighting conditions, and stylistic cues, and (ii) a concise single-sentence caption summarizing the same scene. Both captions are designed to be visual, concrete, and directly suitable for scene recreation by video generation models. We present the instruction generation prompt in Appendix A.1.

In total, GEOATTRACTION-500 comprises 500 carefully curated tourist attraction landmarks distributed worldwide, covering diverse geographic

[1] https://github.com/cvdfoundation/google-landmark

[2] https://huggingface.co/datasets/visheratin/google_landmarks_places

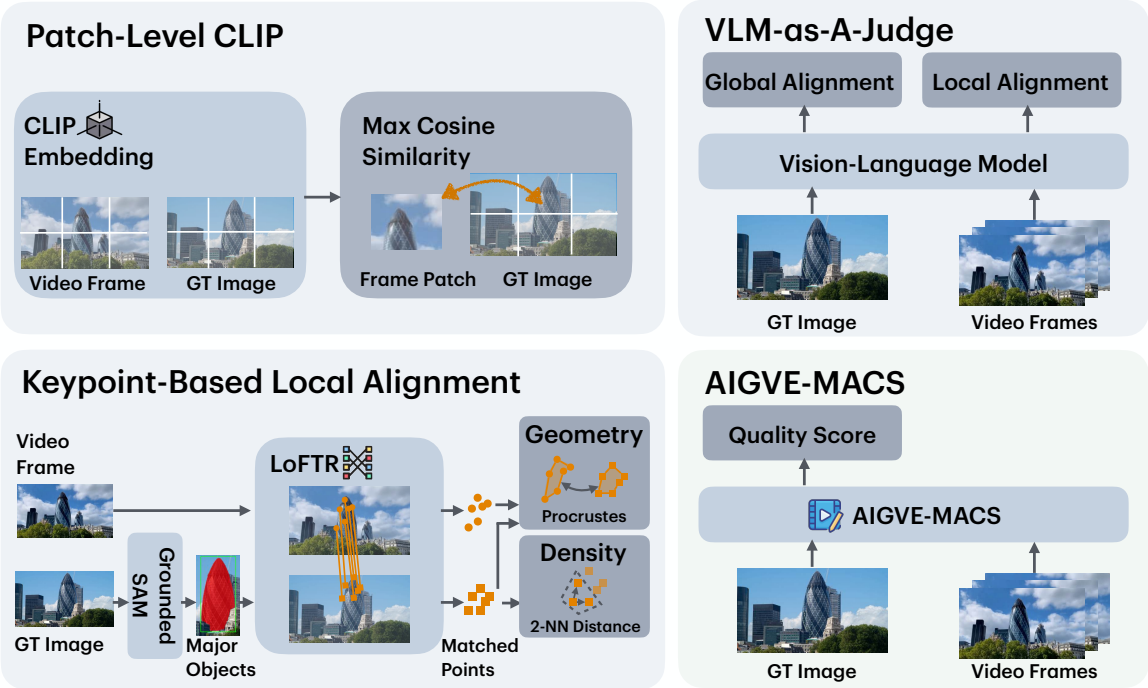

Figure 1: Overview of the evaluation metrics used in Geo-Attraction Landmark Probing (GAP). GAP integrates complementary knowledge-oriented and quality-oriented metrics to assess generated videos. Patch-Level CLIP measures global structural alignment via patch-wise similarity, Keypoint-Based Local Alignment evaluates fine-grained geometric and textural consistency, and VLM-as-A-Judge provides semantic assessments of global and local alignment. In contrast, AIGVE-MACS evaluates overall video quality independently of attraction-specific knowledge.

regions, socio-cultural contexts, and popularity levels. Detailed statistics and the geographic distribution of attractions are presented in Appendix B, Figure 3.

3.2 Evaluation Metrics

To comprehensively assess both the visual quality and the geographically grounded visual knowledge expressed by a video generation model, we employ a suite of complementary evaluation metrics. These metrics are designed to disentangle general perceptual quality from attraction-specific knowledge fidelity, enabling a fine-grained analysis of geo-equity across regions.

For each generated video, we uniformly sample $N{=}5$ frames $\mathcal{F} = \{f_1, \ldots, f_N\}$ and compute all metrics on this shared set. Each metric aggregates frame-level scores in a manner consistent with its evaluation objective, as detailed below. The overview of each metric is presented in Figure 1.

AIGVE-MACS (Liu and Zhang, 2025). We assess overall video generation quality using AIGVE-MACS, a multi-aspect evaluation model that scores generated videos across nine perceptual dimensions, including visual quality, coherence, and aesthetics, on an integer scale from 0 to 5. As a reference-free, quality-forwarded metric, AIGVE-MACS captures general visual fidelity independent of geographic or attraction-specific knowledge. The score is computed by inputting all sampled frames, and we report the overall score as the video quality measure.

Patch-level CLIP (Radford et al., 2021a). To evaluate attraction-specific knowledge at a global structural level, we compute Patch-level CLIP similarity between sampled video frames and the ground-truth reference image of the target attraction. This metric assesses whether the generated video captures the overall spatial layout, silhouette, and environmental context of the attraction, while remaining robust to variations in viewpoint, scale, and appearance.

Given a ground-truth image and a sampled frame f, we partition both images into fixed-size patches following the DINO (Oquab et al., 2024) preprocessing pipeline and extract patch-level visual tokens. Each image is thus represented as a set of patch embeddings corresponding to local spatial regions.

Let $\mathbf{T}^{\mathrm{gt}} = \{\mathbf{t}_i^{\mathrm{gt}}\}_{i=1}^{n}$ denote the patch embeddings extracted from the ground-truth image and $\mathbf{T}^{f} = \{\mathbf{t}_j^{f}\}_{j=1}^{m}$ denote those extracted from frame f, where n and m are the numbers of patches in the ground-truth image and the video frame, respectively. All embeddings are ℓ_2-normalized, and patch similarity is measured using cosine similarity. To account for asymmetric visual coverage between the two images, we compute a symmetric max-matching score:

$$\text{Patch-CLIP}(f) = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\max_{j}\text{cos-sim}\left(\mathbf{t}_i^{\mathrm{gt}}, \mathbf{t}_j^{f}\right) + \frac{1}{m}\sum_{j=1}^{m}\max_{i}\text{cos-sim}\left(\mathbf{t}_j^{f}, \mathbf{t}_i^{\mathrm{gt}}\right)\right). \tag{1}$$

The Patch-level CLIP score is averaged over the sampled frames to obtain the video-level score.

Keypoint-based Local Alignment. While Patch-level CLIP captures global structural correspondence, it does not explicitly assess fine-grained local geometric and textured fidelity. In particular, it may overlook whether localized surface details, such as architectural textures, facade patterns, and fine-scale structural elements, are faithfully reproduced. To evaluate whether a model preserves such detailed structural and textured characteristics of an attraction, we introduce a keypoint matching-based alignment metric that measures local detailed correspondence between generated frames and the ground-truth image.

By explicitly comparing dense keypoint correspondences within landmark-relevant regions and jointly accounting for match density and geometric consistency, this metric provides a fine-grained and interpretable measure of structural alignment, which is essential for detecting subtle errors in geographically grounded visual knowledge that may be overlooked by global similarity measures.

Specifically, given the ground-truth image, we first identify K landmark-relevant regions as binary masks $\{M_k\}_{k=1}^{K}$ using Grounded SAM (Ren et al., 2024). These regions correspond to visually salient components of the attraction, such as buildings, towers, or distinctive architectural elements. For each sampled video frame f, we extract dense keypoint correspondences using LoFTR (Sun et al., 2021), $\mathcal{P}^f = \{(p_i^{\mathrm{gt}}, p_i^{f})\}_{i=1}^{N_f}$, where p_i^{gt} and p_i^{f} denote matched keypoints in the ground-truth image and the generated frame, respectively. For each region M_k, we define the index set $\mathcal{I}_k^f$, containing matches whose ground-truth keypoints fall inside the region.

To account for differences in intrinsic visual complexity across regions, we introduce a region-specific detailness term to normalize the keypoint matching process. Human-made structures often exhibit dense, fine-grained details and yield many keypoint correspondences, whereas natural attractions typically have lower structural complexity. Without normalization, regions with greater intrinsic detail could disproportionately influence the alignment score.

We compute detailness from the Laplacian response of the ground-truth image. Let L denote the grayscale Laplacian, and define the detailness of region M_k as

$$d_k = 1 - \exp\left(-\mathrm{Var}(L \mid M_k)/\tau\right), \tag{2}$$

where τ is a fixed scaling constant. This normalization ensures that alignment scores reflect relative matching quality conditioned on regional visual complexity, rather than absolute keypoint density.

With the region-specific detailness defined above, we next estimate the density of matched keypoints within each region M_k for frame f. Match density serves as a measure of how consistently and densely local structures in the ground-truth image are recovered in the generated frame. Intuitively, a higher density of well-distributed matches suggests that local surface patterns and geometric details, such as architectural textures or repeated facade elements, are faithfully reproduced, rather than matched sparsely or incidentally.

Specifically, let $\{p_i^{\mathrm{gt}}\}_{i\in\mathcal{I}_k^f}$ be the ground-truth keypoints inside the region. We compute the inverse of the mean distance to the second nearest neighbor:

$$\rho_k^f = \left(\frac{1}{|\mathcal{I}_k^f|}\sum_{i\in\mathcal{I}_k^f}\mathrm{NN}_2(p_i^{\mathrm{gt}})\right)^{-1} \tag{3}$$

where $\mathrm{NN}_2(\cdot)$ denotes the Euclidean distance to the second nearest neighbor.

To normalize for region-specific scale and intrinsic visual complexity, we divide ρ_k^f by a self-matching baseline ρ_k^{ref}, which is computed from LoFTR correspondences between the ground-truth image and itself. This self-matching normalization mitigates biases introduced by the LoFTR matcher and isolates alignment differences attributable to the generated content. The resulting density is further normalized by the region detailness to remove the influence of intrinsic regional complexity:

$$r_k^f = \frac{\rho_k^f}{d_k \, \rho_k^{\text{ref}}}. \tag{4}$$

The resulting region-level match density score is obtained via a saturating transform:

$$D_k^f = 1 - \exp\Big(-r_k^f / \beta\Big), \tag{5}$$

where β is a fixed constant.

In addition to match density, we assess geometric consistency by applying Procrustes analysis (Andreella and Finos, 2020) to the matched keypoints and computing the Procrustes disparity δ_k^f, which is converted into a similarity score:

$$G_k^f = 1 - \delta_k^f. \tag{6}$$

We aggregate region-level scores using area-weighted averaging. Let $a_k = \sum M_k$ denote the area of region M_k, and define normalized weights $w_k = a_k / \sum_{\ell=1}^{K} a_\ell$. Frame-level density and geometry scores are then computed as

$$D(f) = \sum_{k=1}^{K} w_k D_k^f, \qquad G(f) = \sum_{k=1}^{K} w_k G_k^f. \tag{7}$$

The overall alignment score for frame f is defined as

$$\text{Keypoint-Match}(f) = D(f) + G(f). \tag{8}$$

Finally, the video-level keypoint matching-based alignment score is defined as the maximum score over the sampled frames. This max aggregation accounts for temporal variability in viewpoint and composition, reflecting the model's ability to generate at least one frame that faithfully captures the fine-grained structure of the target attraction.

VLM-as-A-Judge (Gu et al., 2025). To complement the vision-based metrics with a semantic perspective, we employ a VLM as an automated judge. Specifically, we use GPT-5.1 to compare the ground-truth image with the sampled frames $\mathcal{F}$, conditioned on the generation instruction. The model outputs two integer scores in $[0, 5]$ corresponding to global structural alignment and fine-grained structural and textural alignment.

We average global and fine-grained scores to obtain video-level scores, reflecting the overall alignment of the generated video with the target attraction.

Human Evaluation. We additionally conduct human evaluation following the same protocol as VLM-as-A-Judge. Human annotators are provided with the ground-truth reference image and the same uniformly sampled frames from the generated video. Annotators are asked to assess alignment along the same two dimensions as VLM, using an integer scale from 0 to 5. This human evaluation serves as a complementary benchmark for validating the reliability of the proposed automatic metrics and the VLM-based judgments.

As validated in Appendix D, these metrics capture distinct yet complementary aspects of generated videos while maintaining consistent alignment with human judgments.

Using the proposed evaluation metrics, we conduct comprehensive experiments on GEOATTRACTION-500. Detailed model configurations are provided in Appendix C.

4 Findings and Analysis

4.1 Geo-Knowledge Disparity Across Popularity

Our first analysis examines whether a model's geographically grounded visual knowledge is influenced by attraction popularity, which serves as a proxy for the amount of visual data available during training. Intuitively, more popular attractions are more frequently photographed and documented online, and thus are likely to be better represented in the training corpus. We measure attraction popularity using Wikipedia page view statistics. To quantify the relationship between popularity and visual knowledge alignment, we compute the SRCC, which captures monotonic trends, as well as the slope of a linear regression, which reflects the magnitude of alignment score changes as popularity increases.

As shown in Table 1, all knowledge-oriented metrics exhibit weak correlations with attraction popularity. Besides, their magnitudes are small, and the corresponding regression slopes are close to zero, indicating that increased popularity leads to only marginal gains in alignment. The qualitative examples in Figure 2 further support this finding: attractions with vastly different popularity levels exhibit comparable levels of visual fidelity, with errors typically localized rather than a global disparity. Together, these results suggest that the model maintains **a relatively uniform level of visual knowledge across attractions with vary-**

Metric	SRCC	p-value	Linear Slope
AIGVE-MACS	-1.94×10^{-2}	6.82×10^{-1}	-2.35×10^{-7}
Patch-level CLIP	2.47×10^{-1}	1.20×10^{-7} ***	4.73×10^{-7} ***
Keypoint Matching	9.63×10^{-2}	4.14×10^{-2} *	1.62×10^{-7}
VLM-as-A-Judge	1.93×10^{-1}	3.86×10^{-5} ***	4.42×10^{-6} ***
Human Evaluation	2.30×10^{-1}	8.90×10^{-7} ***	5.39×10^{-6} ***

Table 1: Trend analysis between attraction popularity and alignment scores. We report SRCC, two-sided p-values, and linear regression slopes. Significance stars indicate p-values for the corresponding statistic: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Figure 2: Qualitative case studies illustrating model performance across attractions with varying popularity. For each attraction, the top image shows the ground-truth reference, while the bottom image is generated by the video generation model. Orange dashed outlines highlight regions where the generated content deviates from the reference, indicating structural or geometric errors.

ing popularity, and that the visual knowledge expressed by Sora2 is not strongly driven by differences in training data volume associated with attraction popularity.

4.2 Geo-Knowledge Disparity Across Regions

Following Lalai et al. (2025a), we analyze whether the model's geographically grounded visual knowledge varies across regions. We group attractions by continent and further compare them along two axes: Global North vs. Global South (GN/GS), reflecting differences in development and data availability, and Global West vs. Global East (GW/GE), capturing broad cultural distinctions.

Table 2 shows that mean human alignment scores are broadly similar across continents, with comparable within-region variability. The higher variance in Africa and Oceania is mainly due to smaller sample sizes.

To assess practical significance, we perform pairwise bootstrap equivalence testing with a tolerance of $\delta = 1.0$, where a one-point difference on the 0–5 scale represents a perceptually meaningful change. As shown in Table 3, all continent-pair differences fall within this margin, indicating largely consistent alignment across continents.

Comparisons along the GN/GS and GW/GE axes also show small mean differences, with confidence intervals fully contained within the equivalence margin. Overall, these results suggest that the model's **geographically grounded visual knowledge is relatively uniform across continents, development levels, and cultural groupings**, with no evidence of large systematic regional disparities.

4.3 Impact of Instruction Specificity

We study the effect of prompt specificity using a subsample of 50 attractions, each paired with a short single-sentence prompt and a detailed multi-sentence prompt describing the same ground-truth image. All other generation and evaluation settings are held constant.

Table 4 shows that detailed prompts consistently improve all knowledge-oriented metrics, indicating

Region	N	Mean	Std
Africa	8	3.00	0.71
Americas	67	3.62	0.77
Asia	52	2.93	0.91
Europe	249	3.02	0.86
Oceania	5	3.60	0.42
Global East	147	3.03	0.83
Global West	302	3.13	0.87
Global South	178	3.00	0.81
Global North	271	3.16	0.89

Table 2: Human alignment scores by region.

Region Pair	Mean Diff.	95% CI	Eq.
Americas vs Asia	0.69	[0.38, 0.99]	✓
Americas vs Europe	0.60	[0.38, 0.81]	✓
Europe vs Asia	0.09	[−0.17, 0.36]	✓
Global West vs East	0.09	[−0.07, 0.26]	✓
Global North vs South	0.17	[0.01, 0.32]	✓

Table 3: Pairwise equivalence testing of regional differences in human alignment scores using bootstrap confidence intervals with a tolerance of $\delta = 1.0$. A checkmark indicates practical equivalence.

Metric	Mean_S	Mean_D	p	d
S-CLIP	0.66	0.68	0.00027***	0.57
Kpt-Almt	0.44	0.47	0.041*	0.26
VLM	3.49	3.78	0.017*	0.33

Table 4: Paired comparison between videos generated from short and detailed captions. Mean_S and Mean_D denote average scores under short and detailed caption settings, respectively. Statistical significance is assessed using the Wilcoxon signed-rank test (p), and effect size is measured by Cohen's d. S-CLIP denotes Structural CLIP, Kpt-Almt denotes keypoint-based local alignment, and VLM denotes the VLM-as-A-Judge score. $^{***}p < 0.001$, $^{*}p < 0.05$.

better global and fine-grained alignment. However, the effect sizes remain small to moderate, suggesting that prompt specificity improves the expression of visual knowledge without substantially changing the model's underlying knowledge. Overall, these results indicate that richer prompts help elicit more accurate visual details, while **geographically grounded visual knowledge remains relatively stable and not overly sensitive to prompt verbosity.**

4.4 Discussion and Future Directions

A counterintuitive finding is that, unlike LLMs (Manvi et al., 2024b; Decoupes et al., 2024), text-to-video models exhibit relatively stable geographically grounded visual knowledge across regions and attraction popularity levels. This suggests that their visual knowledge may be less sensitive to location-specific variation in training data than the textual knowledge patterns commonly observed in LLMs.

One possible explanation is that current text-to-video models have not yet reached expert-level visual knowledge. Most alignment scores fall in a medium range, approximately 3–4 on a 5-point scale, indicating relatively uniform underfitting across regions. As these models improve, geographic disparities may become more pronounced.

Another possibility lies in differences in generation mechanisms (Prabhudesai et al., 2025). Autoregressive LLMs generate content sequentially, allowing early errors to accumulate and amplify biases, whereas diffusion-based video models iteratively refine samples through denoising, which may mitigate error accumulation and reduce disparity.

These findings motivate future work comparing geographic disparities across diffusion-based and autoregressive models, as well as longitudinal studies of how such disparities evolve with model scaling. Overall, the uniformity observed in current text-to-video models suggests their potential for applications requiring broad global visual coverage.

5 Conclusion

In this work, we studied the geo-equity and geographically grounded visual knowledge of modern text-to-video models through their generated outputs. We introduced Geo-Attraction Landmark Probing(GAP), a structured evaluation framework that uses tourist attractions as geographically grounded proxies for regional visual knowledge and evaluates generated videos with complementary quality- and knowledge-oriented metrics. By grounding evaluation in real-world reference imagery and applying a consistent evaluation protocol across diverse regions, GAP enables systematic analysis of whether a model's visual knowledge is expressed uniformly across the globe.

Limitations

While GAP provides a systematic framework for evaluating geographically grounded visual knowledge in text-to-video generation, several limitations

remain for future investigation. First, our experiments focus on Sora 2, and the observed geographic uniformity may not generalize to other video generation models or future model generations.

Second, our benchmark uses tourist attractions as proxies for regional visual knowledge. Although attractions are visually distinctive and suitable for objective evaluation, they do not fully cover broader forms of geographic and cultural knowledge that are not embedded in non-landmark environments. Future work could extend GEOATTRACTION-500 to more diverse geographic scenarios, evaluate a wider range of video generation models, and develop adaptive video-level metrics that better capture temporal and fine-grained visual knowledge.

References

Angela Andreella and Livio Finos. 2020. Procrustes analysis for high-dimensional data. *Psychometrika*, 87:1422 – 1438.

Shane Barratt and Rishi Sharma. 2018. A note on the inception score. *arXiv preprint arXiv:1801.01973*.

Yubin Chen, Xuyang Guo, Zhenmei Shi, Zhao Song, and Jiahao Zhang. 2025. T2vworldbench: A benchmark for evaluating world knowledge in text-to-video generation. *arXiv preprint arXiv:2507.18107*.

R. Decoupes, R. Interdonato, Mathieu Roche, M. Teisseire, and S. Valentin. 2024. Evaluation of geographical distortions in language models. *Machine Learning*, 114.

Google AI Studio. 2025. Veo 3 — ai video generation model. https://aistudio.google.com/models/veo-3.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, and 5 others. 2025. Videoscore2: Think before you score in generative video evaluation. *Preprint*, arXiv:2509.22799.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. 2024. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *ArXiv*, abs/2406.15252.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023. VBench: Comprehensive Benchmark Suite for Video Generative Models. *arXiv preprint*. ArXiv:2311.17982 [cs].

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Kling AI. 2026. Kling ai: Next-gen ai video & ai image generator. https://app.klingai.com/global/.

Harsh Nishant Lalai, Raj Sanjay Shah, Jiaxin Pei, Sashank Varma, Yi-Chia Wang, and Ali Emami. 2025a. The world according to llms: How geographic origin influences llms' entity deduction capabilities. *Preprint*, arXiv:2508.05525.

Harsh Nishant Lalai and 1 others. 2025b. The world according to llms: How geographic origin influences llms' entity deduction capabilities. *arXiv preprint arXiv:2508.05525*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2023a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Weiqi Ye, and Jiawei Zhang. 2024a. A survey of ai-generated video evaluation. *Preprint*, arXiv:2410.19884.

Xiao Liu and Jiawei Zhang. 2025. Aigve-macs: Unified multi-aspect commenting and scoring model for ai-generated video evaluation. *Preprint*, arXiv:2507.01255.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2023b. EvalCrafter:

Benchmarking and Evaluating Large Video Generation Models. *arXiv preprint*. ArXiv:2310.11440 [cs].

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024b. Evalcrafter: Benchmarking and evaluating large video generation models. *Preprint*, arXiv:2310.11440.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, Zeyu Wang, Zhifeng Li, Xiu Li, Wei Liu, Dan Xu, Linfeng Zhang, and Qifeng Chen. 2026. Controllable video generation: A survey. *Preprint*, arXiv:2507.16869.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024a. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Rohin Manvi, Samar Khanna, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024b. Large language models are geographically biased. *ArXiv*, abs/2402.02680.

Florentina-Cristina Merciu, Oana-Ramona Ilovan, and Andreea-Loreta Cercleux. 2020. Visual imagery and construction of territorial identity through iconic buildings. case study: The romanian athenaeum. *Territorial Identity and Development*.

Lemuel Odeh. 2010. A comparative analysis of global north and global south economies. *J. Sustain. Dev.*, 12.

OpenAI. 2025a. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/.

OpenAI. 2025b. Sora 2 is here. https://openai.com/index/sora-2/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Mihir Prabhudesai, Mengning Wu, Amir Zadeh, Katerina Fragkiadaki, and Deepak Pathak. 2025. Diffusion beats autoregressive in data-constrained settings. *Preprint*, arXiv:2507.15857.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

E. Ramzi, N. Audebert, C. Rambour, A. Araujo, X. Bitot, and N. Thome. 2023. Optimization of Rank Losses for Image Retrieval. In *In submission to: IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *Preprint*, arXiv:2401.14159.

Aditi Singh. 2023. A survey of ai text-to-image and ai text-to-video generators. *2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, pages 32–36.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. Loftr: Detector-free local feature matching with transformers. *Preprint*, arXiv:2104.00680.

Yan Tao, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

B.S. Turner and Habibul Khondker. 2010. *Globalization East and West*.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2019. Towards accurate generative models of video: A new metric and challenges. *Preprint*, arXiv:1812.01717.

Haoning Wu, Liang Liao, Annan Wang, Chaofeng Chen, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023a. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment. *Preprint*, arXiv:2304.14672.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023b. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. *Preprint*, arXiv:2211.04894.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023c. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives. *arXiv preprint*. ArXiv:2211.04894 [cs, eess].

Xinhao Xiang, Xiao Liu, Zizhong Li, Zhuosheng Liu, and Jiawei Zhang. 2025. Aigve-tool: Ai-generated video evaluation toolkit with multifaceted benchmark. *arXiv preprint arXiv:2503.14064*.

Fan Xie, Dan Zeng, Qiaomu Shen, and Bo Tang. 2025. A comprehensive survey on text-to-video generation. *Chinese Journal of Electronics*, 34(4):1009–1036.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2024. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. 2025. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *Preprint*, arXiv:2503.21755.

A Prompts

A.1 Instruction Generation Prompt

Instruction Generation Prompt

You are a vision-language model tasked with analyzing an image and producing structured captions conditioned on a known landmark.

Task: Analyze the provided image and produce a JSON object with three fields:

1. "is_correct_place": A Boolean indicating whether the image matches the expected landmark.
2. "detailed_caption": A cinematic caption of 3–6 sentences describing the scene.
3. "short_caption": A single-sentence summary of the scene.

Context: The expected landmark is "<landmark_name>" in <city>, <country>.

Instructions:

1. First, determine whether the image depicts <landmark_name> in <city>, <country>.
   - If **YES**, set "is_correct_place" to true and generate both captions.
   - If **NO**, set "is_correct_place" to false and output:
     - "detailed_caption":
       "This image is not <landmark_name> in <city>, <country>. It appears to be <identify landmark if possible>."
     - "short_caption":
       "This is not <landmark_name> in <city>, <country>."

   Do not generate any additional descriptions.
2. If the image is <landmark_name> in <city>, <country>, generate:
   - **Detailed caption (3–6 sentences)** including:
     - Aspect ratio, e.g., 16:9, 9:16
     - Composition: camera angle, framing, foreground/background, depth
     - Location of the main landmark(s) within the frame
     - Weather, lighting, and overall tone
     - Environmental and architectural details
     - Cinematic or stylistic cues useful for video generation
   - **Short caption (1 sentence)** summarizing the same scene.
3. Writing constraints:
   - Do not mention verification, checking, or internal reasoning.
   - Do not refer to "this image" or "the viewer".
   - Captions must be visual, concrete, and suitable for scene recreation.
   - Output must be valid JSON.

Output Format: Return **only** a JSON object with the following structure:

```
{
  "is_correct_place": true/false,
  "detailed_caption": "<detailed caption>",
  "short_caption": "<1 sentence caption based on detailed caption>"
}
```

B GeoAttraction-500 Statistics

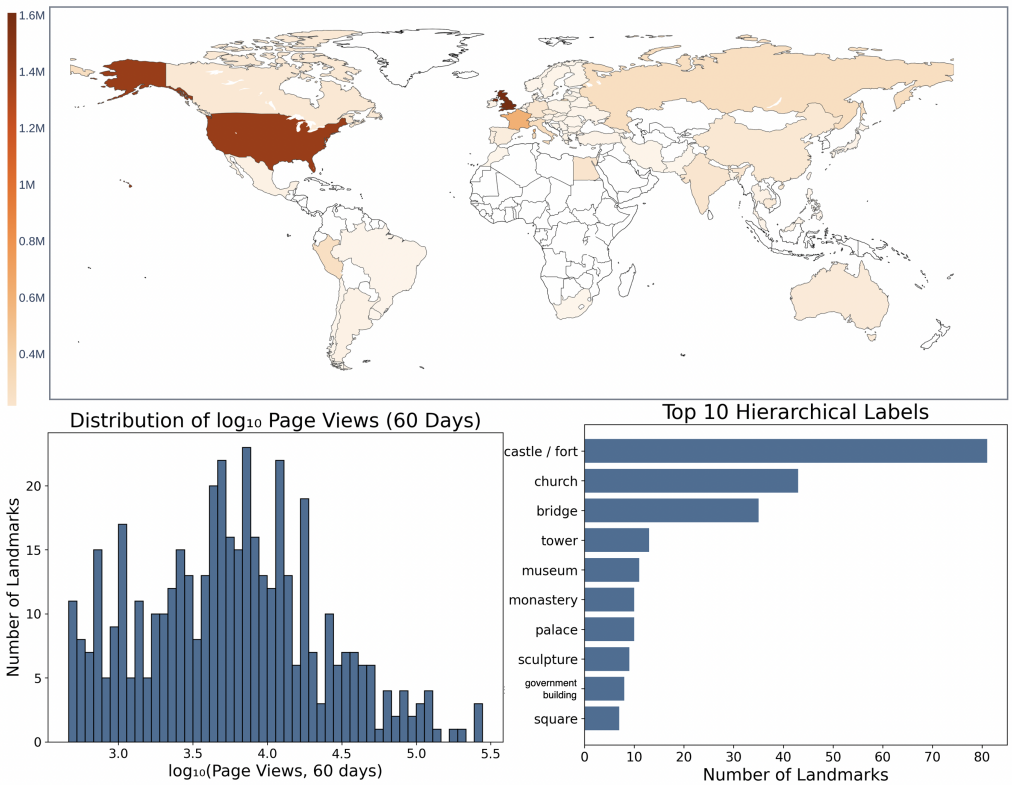

Figure 3: Dataset statistics. **Top:** Geographic distribution of attractions by country, showing broad global coverage. **Bottom left:** Distribution of pageviews of attraction landmarks in log scale. **Bottom right:** Top 10 hierarchical attraction landmark categories, dominated by architectural and cultural structures.

C Experiment Setup

For each attraction, we generate a single video using the OpenAI Sora 2 (OpenAI, 2025b) model conditioned on the corresponding detailed instructions. All videos are generated in portrait orientation with a fixed duration of 4 seconds to ensure consistent presentation across samples.

For evaluation, Patch-level CLIP is computed using the DINOv2-Large[3] (Oquab et al., 2023). For keypoint-based structural alignment, landmark-relevant regions are detected using GroundingDINO[4] (Liu et al., 2023a) and segmented with SAM[5] (Kirillov et al., 2023). Dense keypoint correspondences are extracted using LoFTR (Sun et al., 2021) in outdoor mode, which is well suited for large-scale scene and landmark matching. We set $\tau = 3000$ and $\beta = 1.5$ based on empirical validation.

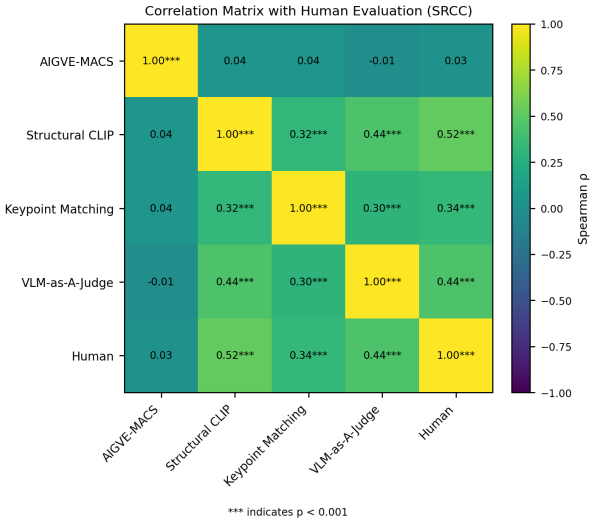

Figure 4: Spearman rank correlation coefficient between human evaluation scores and automatic metrics.

D Metric Validation

We first examine the alignment between our automatic metrics and human judgments. Figure 4 reports the Spearman rank correlation coefficient (SRCC) between human evaluation scores and the metrics used in our framework.

As shown in the figure, AIGVE-MACS exhibits little to no correlation with the other metrics or with human knowledge scores, indicating that it primarily captures video generation quality rather than attraction-specific visual knowledge. This result confirms that our evaluation successfully disentangles quality-related factors from knowledge-oriented assessment.

In contrast, all knowledge-oriented metrics, Patch-level CLIP, Keypoint-Based Local Alignment, and VLM-as-A-Judge, exhibit significant positive correlations with human evaluations. At the same time, the correlations among these automatic metrics themselves are moderate, indicating that they are not redundant. This pattern suggests that each metric captures a distinct yet complementary aspect of attraction-specific visual knowledge. Together, their consistent but differentiated alignment with human judgments demonstrates that the proposed metrics jointly provide a comprehensive and complementary evaluation of the global visual knowledge of video generation models.

[3] facebook/dinov2-large
[4] IDEA-Research/grounding-dino-base
[5] facebook/sam-vit-huge