Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on ArtificialIntelligence and Advanced Algorithms

Article

Evaluating BLIP-2, LLaVA, and GPT-4V for Accessible Artwork Description Generation in Virtual Museums: A Comparative Study Based on WCAG-Aligned Evaluation and User Satisfaction

Jiaying Li [1,*]

[1] Integrated Marketing Communications, Northwestern University, Chicago, IL, USA
* Correspondence: Jiaying Li, Integrated Marketing Communications, Northwestern University, Chicago, IL, USA

Abstract: Virtual museums have expanded cultural access beyond physical boundaries, yet people with visual impairments remain excluded from artwork experiences due to insufficient image descriptions. While vision-language approaches offer the potential to automate accessible content generation, their effectiveness in art-specific contexts has not been rigorously assessed. This study presents a comparative empirical evaluation of three representative vision-language approaches available as of early 2024 --- BLIP-2, LLaVA, and GPT-4V --- for generating accessible artwork descriptions in virtual museum environments. Using a curated evaluation set of 250 artworks spanning six genres from the SemArt dataset, we compare descriptions produced under baseline and art-optimized prompt conditions. Evaluation combines automated captioning metrics (BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr-D), a WCAG 2.1-aligned evaluation rubric scored by trained accessibility evaluators, and a user study with 18 blind and low vision participants. Results indicate that GPT-4V with art-optimized prompts achieves the highest CIDEr-D score (0.476) and WCAG sufficiency rating (3.87/5.00), while all three approaches exhibit notable performance degradation on abstract artworks. User preference data and qualitative feedback suggest that contextual richness, in addition to factual accuracy, may play an important role in shaping satisfaction among visually impaired users. These findings provide practical guidance for virtual museum developers seeking to deploy AI-generated accessible content at scale.

Keywords: accessible image description; vision-language evaluation; virtual museum accessibility; WCAG-aligned assessment

Received: 15 March 2026
Revised: 19 April 2026
Accepted: 01 May 2026
Published: 06 May 2026

1. Introduction

1.1. Accessibility Barriers in Virtual Museums

The digitization of museum collections has created unprecedented opportunities for global cultural participation. Institutions such as the Metropolitan Museum of Art and the Smithsonian now offer virtual access to hundreds of thousands of artworks through online platforms. Yet this expansion has not benefited all audiences equally. For the estimated 2.2 billion people worldwide who experience vision impairment, virtual museum experiences remain largely inaccessible because artwork images typically lack meaningful alternative text descriptions compatible with screen readers and assistive technologies.

The scale of this accessibility gap is substantial. A large-scale analysis of 1.09 million tweets with images found that only 0.1% included user-provided alternative text, even on

platforms that explicitly supported the feature [1]. Automated alt-text generation deployed at platform scale has demonstrated both the feasibility and the limitations of computer vision-based approaches, with descriptions tending toward generic object-level labels that inadequately convey artistic content [2]. Accessibility-focused datasets collected from blind photographers have revealed that real-world images taken by visually impaired users present unique challenges --- variable quality, atypical framing, and diverse content --- that stress-test visual question-answering and image-description algorithms [3]. A comprehensive survey of 836 accessibility research papers from 1994 to 2019 found that over 43% focused on blind and low vision users, yet the intersection of accessibility and cultural heritage technology remains underexplored [4].

1.2. Research Scope and Contributions

This study addresses a specific gap at the intersection of vision-language technology and museum accessibility: to our knowledge, no prior work has systematically compared state-of-the-art vision-language approaches for generating accessibility-oriented artwork descriptions evaluated against both technical standards and end-user needs.

1.2.1. Research Questions

Three research questions guide this investigation. RQ1 asks how BLIP-2, LLaVA, and GPT-4V compare in generating descriptive, accurate, and contextually rich artwork descriptions as measured by standard automated captioning metrics. RQ2 examines the extent to which generated descriptions align with a WCAG 2.1-informed evaluation framework for non-text content (Success Criterion 1.1.1). RQ3 investigates which approach best satisfies the information needs of users with visual impairments in a virtual museum context. We hypothesize that GPT-4V will produce the most contextually rich descriptions, while all approaches will exhibit reduced performance on abstract and non-representational artworks where object-centric visual features are less informative.

1.2.2. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work across vision-language captioning, image accessibility, and AI-driven art analysis. Section 3 details the experimental methodology, including artwork selection, prompt design, and evaluation protocols. Section 4 presents quantitative and qualitative results with cross-genre analysis. Section 5 discusses implications, limitations, and future research directions.

2. Related Work

2.1. Vision-Language Approaches for Image Description

2.1.1. Contrastive and Encoder-Decoder Architectures

The rapid advancement of vision-language pre-training has produced a family of approaches capable of generating detailed image descriptions. Contrastive Language-Image Pre-training (CLIP) demonstrated that training on 400 million image-text pairs enables zero-shot transfer to diverse vision tasks and has since served as the foundational visual encoder for many subsequent architectures [5]. Building on this foundation, BLIP-2 introduced a lightweight Querying Transformer (Q-Former) that bridges frozen image encoders and frozen large language models, achieving state-of-the-art captioning performance with 54 times fewer trainable parameters than comparable approaches [6]. These encoder-decoder architectures excel at structured caption generation while maintaining computational efficiency.

2.1.2. Large Multimodal Approaches

An alternative paradigm connects visual encoders directly to large language models through visual instruction tuning. Early work, such as Flamingo, demonstrated that few-shot visual language models can rapidly adapt to new tasks by interleaving visual and textual tokens. LLaVA pioneered a further step by training an end-to-end multimodal assistant using GPT-4-generated instruction-following data, demonstrating strong open-ended conversational capabilities about images [7]. GPT-4V, building on the architecture

described in OpenAI's technical report, accepts both image and text inputs and has demonstrated strong performance across professional benchmarks [8]. A comprehensive qualitative evaluation of GPT-4V across diverse domains revealed unprecedented capability in processing interleaved multimodal inputs, alongside systematic limitations in spatial reasoning and fine-grained visual detail [9]. The architectural differences between encoder-decoder and large multimodal paradigms suggest they may exhibit distinct strengths for art-specific description tasks, motivating direct comparison.

2.2. Image Accessibility for Visually Impaired Users

Research on image descriptions for blind and low vision (BLV) users has established that description preferences are highly context-dependent. A qualitative study of 28 BLV participants across seven digital contexts found that users expect different description content depending on whether they encounter images in news articles, social media, or eCommerce --- a finding that challenges one-size-fits-all approaches to accessible alt-text [10]. The VizWiz-Captions dataset, comprising 39,181 images taken by blind photographers paired with 195,905 human-written captions, demonstrated that images from BLV users present unique captioning challenges, including poor lighting, atypical framing, and text-heavy scenes [11]. A four-level semantic content model for accessible descriptions of visualizations --- ranging from construction properties to domain-specific insights --- provides a principled hierarchy for evaluating description depth that can be adapted from data visualization to artwork contexts [12]. These studies collectively establish that effective, accessible descriptions must be contextually appropriate, sufficiently detailed, and structured around user information needs rather than algorithmic convenience.

2.3. AI-Driven Art Analysis and Description

Computational approaches to art understanding have progressed from style classification to natural language description generation. The ArtEmis dataset paired 455,000 emotion attributions and grounded verbal explanations with 80,000 artworks, revealing that art-related language is substantially more abstract, metaphorical, and affective than standard captioning vocabularies [13]. KALE, a recent art-specific captioning approach, enhanced vision-language architectures by integrating artwork metadata via a heterogeneous knowledge graph, thereby improving CIDEr performance on the SemArt benchmark [14]. These developments confirm that art description requires capabilities beyond object detection --- including stylistic interpretation, historical contextualization, and affective resonance --- yet to our knowledge, no prior work has evaluated general-purpose vision-language approaches specifically against accessibility requirements in museum contexts.

3. Methodology

3.1. Artwork Selection and Evaluation Set Construction

The evaluation set was constructed from the SemArt dataset, which contains 21,384 fine-art images paired with expert-written artistic comments and structured metadata including author, title, date, technique, type, and artistic school. From this corpus, 250 artworks were selected through stratified sampling across six genre categories to ensure balanced representation, as detailed in Table 1. The multi-topic annotation scheme from prior work on painting description generation provided reference descriptions decomposed into form, content, and context dimensions [15]. Each selected artwork was verified to include a reference artistic comment of at least 30 words and high-resolution imagery suitable for input to all three vision-language approaches. Supplementary metadata from the Metropolitan Museum of Art Open Access collection (CC0 license) enriched 87 of the 250 items with additional curatorial fields, including culture, period, and medium. The selection aimed to capture a range of visual complexity: portraits and still-life paintings present clearly delineated subjects amenable to object-level description, while abstract works and complex religious compositions test the limits of visual

grounding, where compositional and symbolic elements dominate over discrete, identifiable objects.

Table 1. Distribution of Artworks in the Evaluation Set by Genre

Genre	Count	Proportion (%)	Time Period Range	Unique Artists
Portrait	52	20.8	1434–1896	41
Landscape	48	19.2	1510–1903	38
Religious	43	17.2	1280–1785	36
Still Life	38	15.2	1596–1888	29
Abstract	34	13.6	1907–1965	27
Genre Scene	35	14.0	1509–1892	31
Total	250	100.0	1280–1965	189

Source: Artworks sampled from SemArt. Metadata verified against original dataset documentation.

3.2. Description Generation Protocol

3.2.1. Prompt Design and Configuration

Each of the three approaches was evaluated under two prompting conditions: a baseline prompt and an art-optimized prompt. Throughout this paper, "BLIP-2" refers to the InstructBLIP variant (instruction-tuned on top of the BLIP-2 architecture) to ensure a fair comparison under instruction-following conditions. The baseline prompt used a general accessibility instruction: "Describe this image in detail for a person who cannot see it." The art-optimized prompt incorporated domain-specific guidance structured around the multi-aspect description framework identified in prior accessibility research [16]: role-based framing ("As an art museum audio guide for visually impaired visitors"), explicit requests for visual elements, artistic technique, historical context, and emotional impression, and output length guidance (75--150 words). This multi-aspect decomposition is loosely inspired by structured reasoning approaches in multimodal settings, where task decomposition has been associated with improved output quality [17]. Table 2 reports the specific configuration parameters for each approach.

Table 2. Configuration Parameters for Each Vision-Language Approach

Parameter	BLIP-2 (ViT-G + FlanT5-XXL)	LLaVA-1.5 (13B)	GPT-4V (gpt-4-vision-preview)
Visual Encoder	ViT-G/14 (EVA-CLIP)	CLIP ViT-L/14-336	Proprietary
Language Component	FlanT5-XXL (11B)	Vicuna-13B-v1.5	GPT-4
Max Output Tokens	256	512	512
Temperature	1.0 (nucleus sampling)	0.7	0.7
Image Resolution	224 × 224	336 × 336	Variable (auto)
Instruction Tuned	Via InstructBLIP variant	Yes (visual instruction tuning)	Yes (details proprietary)
Inference Platform	Local (A100 80GB)	Local (A100 80GB)	API (2024-01 version)

Note: BLIP-2 was evaluated using the InstructBLIP variant for fair comparison under instruction-following conditions.

3.2.2. Art-Optimized Prompt Strategies

The art-optimized condition was implemented through two alternative prompt strategies, each compared against the baseline. The role-based strategy positioned the approach as a museum audio-guide specialist and requested descriptions that addressed both visual content and interpretive context. The multi-aspect strategy explicitly decomposed the description task into four sequential components: (1) visual content and composition, (2) artistic technique and medium, (3) historical and cultural context, and (4) emotional tone and aesthetic qualities. This decomposition was informed by the form-content-context taxonomy developed for knowledgeable art description and by the affective language patterns documented in the ArtEmis corpus. The optimized results reported throughout this paper reflect the more suitable of the two prompt strategies for each approach in this exploratory study: the multi-aspect strategy for InstructBLIP and the role-based strategy for both LLaVA and GPT-4V.

3.3. Evaluation Design

3.3.1. Automated Metrics and WCAG-Aligned Assessment

Automated evaluation employed five standard captioning metrics computed against SemArt's expert-written artistic comments as reference descriptions: BLEU-1, BLEU-4, METEOR, ROUGE-L, and CIDEr-D. While these metrics were originally developed for natural image captioning and may not fully capture the nuances of art-specific language, they provide a standardized and reproducible basis for cross-approach comparisons, enabling positioning relative to prior art captioning benchmarks. Alignment with accessibility needs relevant to WCAG 2.1 was assessed using a four-dimension rubric informed by Success Criterion 1.1.1: sufficiency (does the description convey artwork purpose and content), accuracy (are visual elements correctly identified), contextual richness (is relevant art-historical context included), and clarity (is the language accessible to diverse audiences, including those with cognitive disabilities). Three trained accessibility evaluators --- each with at least two years of experience in digital accessibility auditing --- independently scored each dimension on a 1--5 Likert scale for a random subset of 20 descriptions per approach, yielding 60 ratings per dimension (20 artworks × 3 evaluators). Evaluators received a calibration session with 10 practice descriptions before scoring the experimental set; during this session, a score of 4.0 was adopted as a practical internal benchmark indicating that a description substantially satisfied the intended criterion in a given dimension. The semantic content model, which distinguishes four levels from construction properties to domain-specific insights, was applied to classify the depth of each generated description. Inter-rater reliability was computed using Fleiss' kappa.

3.3.2. User Study Protocol

A user study with 18 BLV participants (11 blind, 7 low vision; ages 24--67, M = 41.3) was conducted following IRB approval. Participants were recruited through the National Federation of the Blind and local accessibility organizations. Each participant evaluated 15 artwork descriptions (5 per approach, blinded to source) presented through their preferred screen reader (JAWS, NVDA, or VoiceOver). Measures included information satisfaction on a 7-point Likert scale, forced-choice preference ranking across approaches, and semi-structured qualitative feedback. The study design followed established protocols from accessibility-focused captioning research. Table 3 summarizes participant demographics.

Table 3. User Study Participant Demographics (N = 18)

Characteristic	Category	Count
Vision Status	Totally blind	11

	Low vision	7
Screen Reader	JAWS	8
	NVDA	6
	VoiceOver	4
Museum Visit Frequency	Monthly or more	5
	Several times per year	7
	Rarely or never	6
Art Knowledge (self-rated)	High	4
	Moderate	9
	Low	5

Source: Self-reported participant data collected during pre-study intake.

4. Results and Analysis

4.1. Automated Captioning Metrics

4.1.1. Cross-Approach Performance Comparison

Table 4 presents automated metric scores for all three approaches under both baseline and art-optimized prompt conditions. GPT-4V consistently achieved the highest scores across all metrics, with its art-optimized configuration reaching a CIDEr-D of 0.476, representing a 19.6% relative improvement over BLIP-2's optimized score (0.398) and a 23.0% improvement over LLaVA's (0.387). The gain from prompt optimization varied across approaches: LLaVA showed the largest relative CIDEr-D improvement (21.7%), followed by BLIP-2 (16.7%) and GPT-4V (15.5%). These patterns, illustrated in Figure 1, suggest that structured, art-specific prompting yields differential benefits depending on the underlying architecture's language-generation capacity. Broad-coverage multimodal benchmarks have similarly shown that model performance varies substantially across different ability dimensions [18, 19].

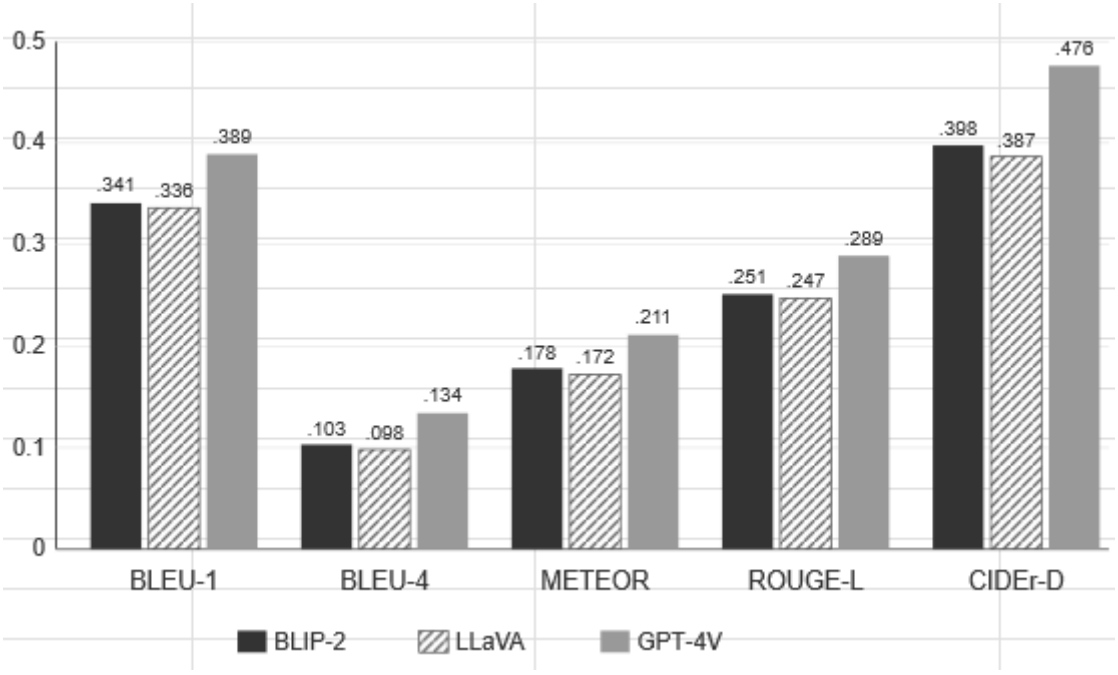

Figure 1. Automated Metric Scores Under Art-Optimized Prompts

Comparison of five automated captioning metrics for BLIP-2, LLaVA, and GPT-4V under art-optimized prompt conditions. GPT-4V achieves the highest scores on all metrics, with its largest advantage on CIDEr-D (0.476 vs. 0.398 for BLIP-2 and 0.387 for LLaVA).

Table 4. Automated Captioning Metrics: Baseline vs. Art-Optimized Prompts

Metric	BLIP-2 Base	BLIP-2 Opt.	LLaVA Base	LLaVA Opt.	GPT-4V Base	GPT-4V Opt.
BLEU-1	0.312	0.341	0.298	0.336	0.354	0.389
BLEU-4	0.087	0.103	0.079	0.098	0.112	0.134
METEOR	0.156	0.178	0.149	0.172	0.183	0.211
ROUGE-L	0.224	0.251	0.213	0.247	0.258	0.289
CIDEr-D	0.341	0.398	0.318	0.387	0.412	0.476
Avg. Length (words)	38.4	67.2	42.1	81.6	53.7	103.4

Source: Computed on the 250-artwork evaluation set using SemArt expert-written artistic comments as references. All metrics calculated using standard implementations from the pycocoevalcap library.

4.1.2. WCAG-Aligned Evaluation Scores

Table 5 reports mean WCAG-aligned evaluation scores across the four assessment dimensions. Inter-rater reliability among the three evaluators was substantial (Fleiss' κ = 0.71). GPT-4V achieved the highest scores on all four dimensions, with a particularly pronounced advantage in contextual richness (3.91/5.00) compared to LLaVA (2.89) and BLIP-2 (2.54). This 53.9% gap between GPT-4V and BLIP-2 on contextual richness may reflect GPT-4V's capacity to incorporate art-historical knowledge --- period attribution, stylistic movement identification, and iconographic interpretation --- that the encoder-decoder architecture does not consistently produce. The clarity dimension showed the smallest inter-approach variance (range: 3.58--3.74), as shown in Figure 2, indicating that all three approaches generate linguistically accessible text at comparable levels. Prior work on CNN-based art style classification has demonstrated that computational approaches can learn meaningful art-historical patterns from visual features alone [20, 21], yet translating such visual representations into detailed natural-language descriptions may pose additional challenges for approaches with smaller language components.

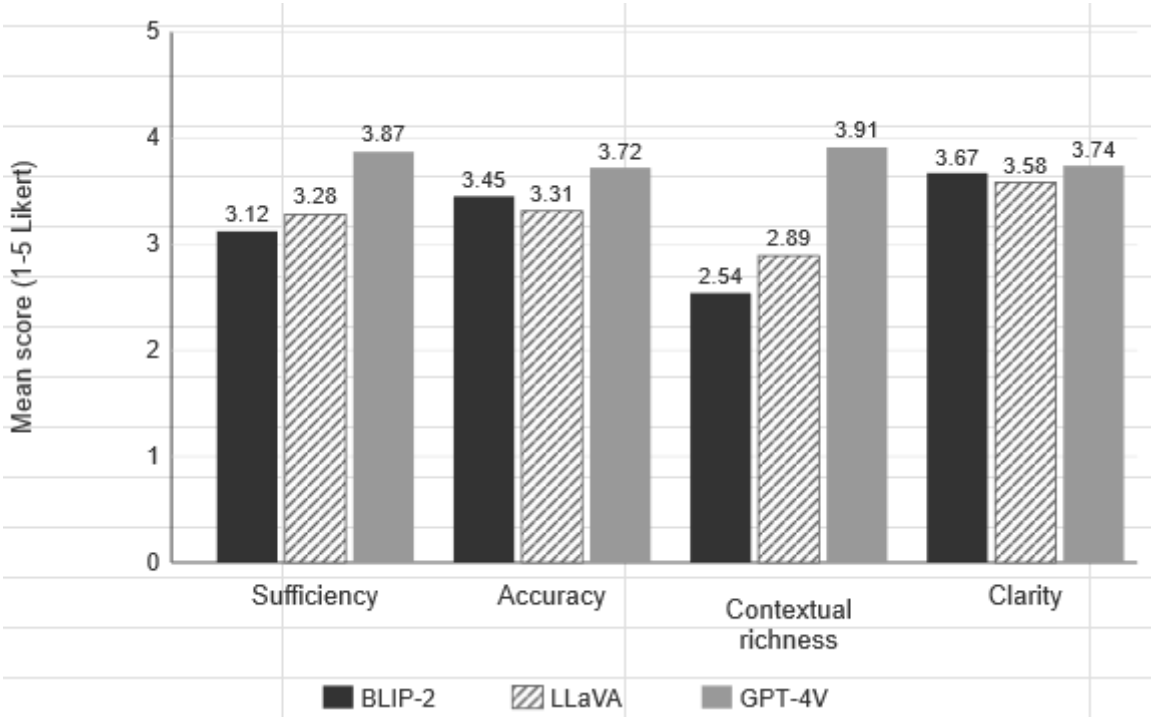

Figure 2. WCAG-Aligned Evaluation Dimension Scores Across Three Approaches

Mean WCAG-aligned evaluation ratings for BLIP-2, LLaVA, and GPT-4V across four assessment dimensions. GPT-4V achieves the highest score in every dimension, with its most pronounced advantage in contextual richness (3.91 vs. 2.54 for BLIP-2). The clarity dimension exhibits the least variation, with all three scoring between 3.58 and 3.74.

Table 5. Mean WCAG 2.1-Aligned Evaluation Scores by Dimension (1--5 Scale, Art-Optimized Prompts)

Dimension	BLIP-2	LLaVA	GPT-4V	Fleiss' κ
Sufficiency	3.12	3.28	3.87	0.68
Accuracy	3.45	3.31	3.72	0.74
Contextual Richness	2.54	2.89	3.91	0.69
Clarity	3.67	3.58	3.74	0.73
Mean (all dimensions)	3.20	3.27	3.81	0.71

Source: Independent ratings by 3 trained accessibility evaluators on 20 descriptions per approach (60 ratings per dimension per approach). Scale: 1 = does not meet criterion, 5 = fully meets criterion.

4.2. User Study Findings

Participant responses revealed consistent preferences aligned with the automated evaluation patterns. Mean information satisfaction scores (7-point scale) were 4.13 (SD = 1.21) for BLIP-2, 4.38 (SD = 1.14) for LLaVA, and 5.24 (SD = 0.93) for GPT-4V. A Friedman test indicated a statistically significant difference across approaches ($\chi^2(2) = 14.67$, $p < 0.001$, Kendall's W = 0.41). Post-hoc Wilcoxon signed-rank tests with Bonferroni correction confirmed that GPT-4V ratings were significantly higher than both BLIP-2 ($p = 0.003$, $r = 0.49$) and LLaVA ($p = 0.011$, $r = 0.42$), while the BLIP-2 versus LLaVA difference did not reach significance ($p = 0.184$). Figure 3 summarizes the preference distribution and satisfaction scores.

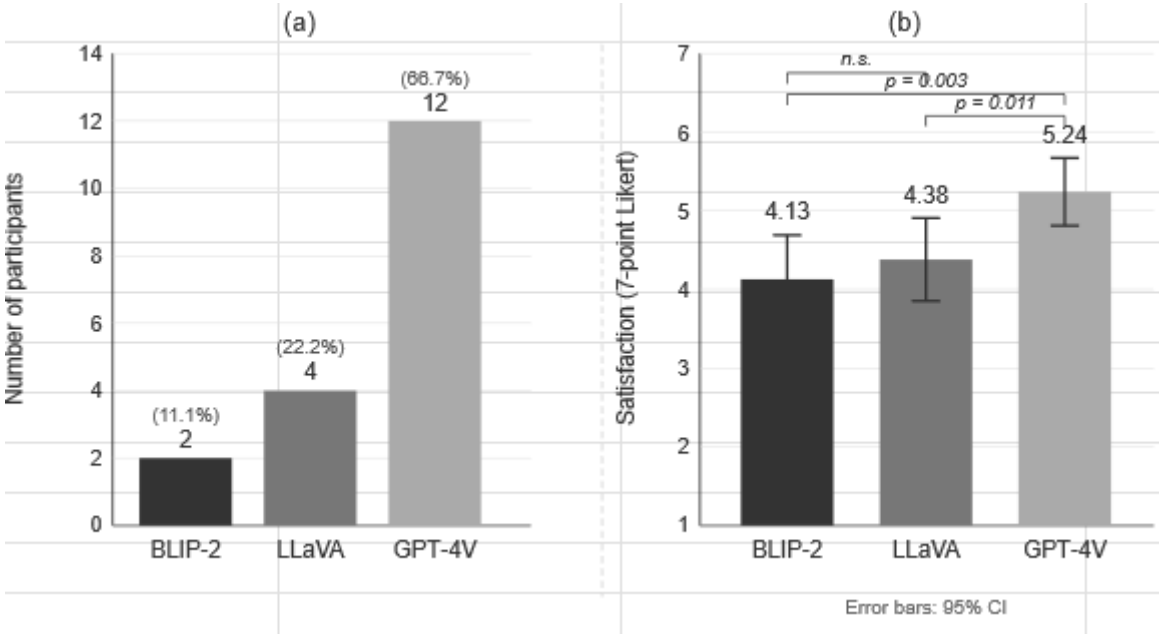

Figure 3. User Preference Distribution and Satisfaction Scores

Distribution of first-choice preferences among 18 BLV participants: GPT-4V was preferred by 12 participants (66.7%), LLaVA by 4 (22.2%), and BLIP-2 by 2 (11.1%). (b) Mean information satisfaction

scores (7-point Likert scale) with 95% confidence intervals: GPT-4V (M = 5.24, SD = 0.93) significantly outperformed BLIP-2 (M = 4.13, SD = 1.21; p = 0.003) and LLaVA (M = 4.38, SD = 1.14; p = 0.011). The gap between BLIP-2 and LLaVA was not statistically significant (p = 0.184).

In forced-choice preference rankings, 12 of 18 participants (66.7%) ranked GPT-4V as their first choice, 4 (22.2%) preferred LLaVA, and 2 (11.1%) preferred BLIP-2. Qualitative feedback suggested that many participants valued descriptions that situated artworks within an interpretive context --- for example, by identifying artistic movements, conveying emotional atmosphere, or providing brief historical framing. Several participants with moderate or high self-rated art knowledge expressed that contextual information was essential for meaningful engagement, whereas factual accuracy alone produced descriptions they characterized as "flat."

4.3. Genre-Specific Performance Analysis

4.3.1. Abstract Versus Figurative Art

Table 6 presents CIDEr-D scores stratified by genre under art-optimized prompts. All three approaches demonstrated substantially lower performance on abstract artworks than on figurative artworks. GPT-4V's CIDEr-D on abstract art (0.312) was 42.0% lower than its portrait score (0.538), while BLIP-2 showed a 53.3% decline (from 0.467 to 0.218) and LLaVA a 45.2% decline (from 0.451 to 0.247). This degradation reflects the challenge posed by abstract artworks, which lack identifiable objects, and by vision-language architectures pre-trained on natural image-text pairs, which rely heavily on object-centric features. The bootstrapping pre-training approach underlying encoder-decoder captioning, where a captioner generates synthetic captions filtered for quality [22-24], inherently favors training distributions dominated by concrete visual referents. BLIP-2 exhibited the steepest abstract-art penalty, possibly because its Q-Former extracts fixed-length query representations that compress visual information. GPT-4V's more moderate decline may indicate that its larger language component can partially compensate by generating descriptions grounded in color relationships and compositional dynamics when object-level features are unavailable.

Table 6. CIDEr-D Scores by Genre (Art-Optimized Prompts)

Genre	N	BLIP-2	LLaVA	GPT-4V	GPT-4V vs. BLIP-2 (Δ%)
Portrait	52	0.467	0.451	0.538	+15.2
Landscape	48	0.432	0.419	0.512	+18.5
Religious	43	0.389	0.378	0.487	+25.2
Still Life	38	0.441	0.428	0.521	+18.1
Abstract	34	0.218	0.247	0.312	+43.1
Genre Scene	35	0.401	0.392	0.489	+21.9

Source: Computed on genre-stratified subsets of the 250-artwork evaluation set. Reference descriptions from SemArt expert-written artistic comments.

4.3.2. Cultural Context and Terminology Handling

Qualitative analysis of generated descriptions revealed systematic patterns in how each approach handled culturally specific iconography and specialized art terminology. Two researchers independently coded each description for correct identification of iconographic elements and artistic movements; inter-coder agreement exceeded 90%, and disagreements were resolved through discussion. BLIP-2 produced the most conservative descriptions, rarely venturing beyond visual element enumeration and occasionally misidentifying religious iconography --- describing a halo as a "circular shape behind the head" or a vanitas arrangement as "a table with objects." LLaVA demonstrated moderate cultural awareness, correctly identifying common Christian iconographic elements in

approximately 62% of the 43 religious artworks, yet struggling with less common symbolic references. GPT-4V exhibited the strongest cultural contextualization, correctly identifying artistic movements in 78% of cases across all 250 artworks and producing art-historical terminology (chiaroscuro, tenebrism, impasto) that matched expert reference descriptions. These findings may inform future multilingual audio guide deployment, where culturally informed descriptions must navigate divergent art-historical traditions across linguistic communities.

5. Discussion

5.1. Practical Implications for Virtual Museum Accessibility

The comparative results yield differentiated recommendations depending on institutional deployment constraints. GPT-4V with art-optimized prompts produced descriptions that scored highest on the depth and contextual richness dimensions valued by BLV users, achieving the highest scores across all automated metrics, all four WCAG-aligned evaluation dimensions, and user satisfaction ratings. Its mean score of 3.81/5.00 on the WCAG-informed rubric, while the strongest among the three approaches evaluated, remains below the internal 4.0 benchmark adopted during evaluator calibration to indicate substantial criterion satisfaction --- indicating that even the best-performing approach requires human review before deployment at scale. For institutions with API budget constraints, BLIP-2 and LLaVA offer locally deployable alternatives with moderate performance: their WCAG mean scores (3.20 and 3.27, respectively) suggest utility as draft generators for human editors, reducing the labor required to produce accessible descriptions by shifting the workflow from writing descriptions from scratch to reviewing and supplementing machine-generated drafts.

The qualitative finding that participants appeared to value contextual richness over factual accuracy alone carries important design implications. Virtual museum accessibility efforts should prioritize prompt strategies that elicit interpretive and contextual content rather than focusing exclusively on visual fidelity. The multi-aspect prompting strategy --- decomposing description requests into visual, technical, historical, and affective components --- proved effective for BLIP-2, while the role-based strategy served LLaVA and GPT-4V better, suggesting that prompt optimization should be approach-specific rather than universal.

5.2. Limitations

Several limitations constrain the generalizability of these findings. First, the art-optimized results reported for each approach reflect the more suitable of the two prompt strategies identified within this exploratory study. Because a separate validation set or pre-registered prompt-selection rule was not used, the optimized scores should be interpreted as indicative rather than strictly confirmatory. Second, automated metrics were computed against SemArt expert-written artistic comments, which are art-historical in nature and do not constitute accessibility-oriented reference descriptions; higher n-gram overlap with such references does not necessarily indicate greater usefulness for BLV users, and future work should develop purpose-built accessibility reference corpora. Third, the user study presented each participant with five descriptions per approach drawn from different artworks rather than fully matched sets of the same artwork across all three approaches. Accordingly, the user-study findings are better interpreted as reflecting overall user preference patterns under realistic exposure conditions rather than as a strictly controlled within-artwork comparison, although randomized assignment partially mitigated artwork-level variation. Fourth, the evaluation set of 250 artworks, while genre-stratified, represents a small fraction of major museum collections and skews toward Western European painting traditions (1280--1965). The user study sample of 18 BLV participants, though adequate for identifying preference patterns, limits the statistical power to detect smaller effect sizes. All evaluations were conducted in English, leaving multilingual performance unassessed.

Future research should extend this evaluation along several dimensions. Multilingual description generation would address the needs of non-Anglophone Museum audiences. Integrating text-to-speech quality evaluation would assess the full pipeline from image to the audio guide experience. The development of a purpose-built benchmark dataset pairing museum artworks with accessibility-graded reference descriptions would establish a standardized evaluation infrastructure. Longitudinal deployment studies in operational virtual museum settings would provide ecological validity, measuring whether AI-generated descriptions translate into sustained engagement gains for visitors with disabilities.

References

1. C. Gleason, P. Carrington, C. Cassidy, M. R. Morris, K. M. Kitani, and J. P. Bigham, "It's almost like they're trying to hide it: How user-provided image descriptions have failed to make Twitter accessible," in *Proceedings of The World Wide Web Conference (WWW '19)*, pp. 549–559, ACM, 2019. Available: https://doi.org/10.1145/3308558.3313605
2. S. Wu, J. Wieland, O. Farivar, and J. Schiller, "Automatic alt-text: Computer-generated image descriptions for blind users on a social network service," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, pp. 1180–1192, ACM, 2017. Available: https://doi.org/10.1145/2998181.2998364
3. D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "VizWiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '18)*, pp. 3608–3617, IEEE, 2018.
4. K. Mack, E. McDonnell, D. Jain, L. L. Wang, J. E. Froehlich, and L. Findlater, "What do we mean by 'accessibility research'? A literature survey of accessibility papers in CHI and ASSETS from 1994 to 2019," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, Article 371, ACM, 2021. Available: https://doi.org/10.1145/3411764.3445412
5. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML '21)*, PMLR 139, pp. 8748–8763, 2021.
6. J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning (ICML '23)*, PMLR 202, pp. 19730–19742, 2023.
7. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems 36 (NeurIPS '23)*, 2023.
8. OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
9. Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of LMMs: Preliminary explorations with GPT-4V(ision)," *arXiv preprint arXiv:2309.17421*, 2023.
10. A. Stangl, M. R. Morris, and D. Gurari, "'Person, shoes, tree. Is the person naked?' What people with vision impairments want in image descriptions," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, Article 315, ACM, 2020. Available: https://doi.org/10.1145/3313831.3376404
11. D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in *Proceedings of the European Conference on Computer Vision (ECCV '20)*, pp. 417–434, Springer, 2020.
12. A. Lundgard and A. Satyanarayan, "Accessible visualization via natural language descriptions: A four-level model of semantic content," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 902–912, 2022. Available: https://doi.org/10.1109/TVCG.2021.3114770
13. P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas, "ArtEmis: Affective language for visual art," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '21)*, pp. 11564–11574, 2021.
14. Y. Jiang, K. A. Ehinger, and J. H. Lau, "KALE: An artwork image captioning system augmented with heterogeneous graph," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI '24)*, pp. 7663–7671, 2024.
15. Z. Bai, Y. Nakashima, and N. Garcia, "Explain me the painting: Multi-topic knowledgeable art description generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '21)*, pp. 5422–5432, 2021.
16. A. Stangl, N. Verma, K. R. Fleischmann, M. R. Morris, and D. Gurari, "Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision," in *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*, Article 12, ACM, 2021. Available: https://doi.org/10.1145/3441852.3471233
17. Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. J. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.
18. W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems 36 (NeurIPS '23)*, 2023.

19. Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, "MMBench: Is your multi-modal model an all-around player?" in *Proceedings of the European Conference on Computer Vision (ECCV '24)*, pp. 216–233, 2024.
20. A. Elgammal, B. Liu, D. Kim, M. Elhoseiny, and M. Mazzone, "The shape of art history in the eyes of the machine," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI '18)*, pp. 2183–2191, 2018. Available: https://doi.org/10.1609/aaai.v32i1.11894
21. J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, ... K. Simonyan, "Flamingo: A visual language model for few-shot learning," in *Advances in Neural Information Processing Systems 35 (NeurIPS '22)*, 2022.
22. J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning (ICML '22)*, PMLR 162, pp. 12888–12900, 2022.
23. N. Garcia and G. Vogiatzis, "How to read paintings: Semantic art understanding with multi-modal retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV '18) Workshops*, pp. 676–691, Springer, 2018.
24. W3C, *Web Content Accessibility Guidelines (WCAG) 2.1*, W3C Recommendation, 2018. Available: https://www.w3.org/TR/WCAG21/