CUBE-MT: A Cultural Benchmark for Multimodal Knowledge Graph Construction with Generative Models

Albert Meroño-Peñuela[1], Xin Fan Guo[2], Nitisha Jain[1], Filip Birčanin[1], Timothy Neate[1], Thomas van Erven[3], Sándor Daranyi[3], and Nasrine Olson[3]

[1] King's College London, 30 Aldwych, London, UK
{albert.merono}@kcl.ac.uk
[2] Imperial College London, Exhibition Rd, London, UK
[3] University of Borås, Allégatan 1, Borås, Sweden

Abstract. Cultural heritage institutions (GLAM) have a societal role of guaranteeing access to their collections independently of location or background, as a pledge for knowledge equity. However, around one billion people (15% of the world's population) experience some form of disability that hinders such access, especially when considering the multimedia gap: differences in the use of different types of media to convey content to different audiences. Knowledge Graphs are increasingly becoming more multimodal by supporting images and text. However, the extent to which they address the multimedia gap for people with disabilities is not well understood, due to the lack of appropriate cultural evaluation frameworks. To address this, here we propose CUBE-MT, a benchmark and dataset that leverages generative models to build Multimodal Knowledge Graphs (MMKGs) with surrogate, multimedia representations that adapt to the sensory capacities of cultural heritage collection users. We extend the CUBE (CUltural BEnchmark for Text-to-Image models) and Muse-IT datasets[4] for paintings to encompass 6 modalities (text, images, Braille, speech, music, and 3D models); a collection of prompts to account for their cultural awareness and diversity; and a dataset with the resulting MMKG mapped to Wikidata. We show usage and evaluate the effectiveness of our approach in: (1) a quantitative assessment of cultural diversity; (2) an expert survey; and (3) a user study with people with aphasia focusing on perceptual and comprehension differences between model-generated and original MMKGS objects.

Keywords: Multimodal Knowledge Graphs · Cultural Heritage · Knowledge Access

1 Introduction

Cultural heritage is commonly preserved and promoted by GLAM (Galleries, Libraries, Archives, and Museums) institutions with a mission to provide access to

[4] MuseIT is an Horizon Europe project on innovative technologies for broadening access to cultural heritage for people with disabilities: https://www.muse-it.eu/

knowledge [18,46]. Large digital platforms like Europeana [41] and the Cultural Heritage Cloud[5] provide this access globally on the Web, including in machine-readable form [32]. The number of cultural heritage and GLAM collection knowledge graphs (KGs) has dramatically exploded in recent years, providing a wealth of semantically enriched cultural heritage data [5,16,21,12,11,37,51].

Despite such advances, contemporary KGs face a critical **multimedia gap**. In large cross-domain KGs such as Wikidata [54], this gap manifests as under-representation of media beyond text, for instance around 95% of Wikidata items lack even basic image representations (property *P18*)[6], and coverage of audio, video, tactile, and other modalities is nearly non-existent. This creates sample biases favouring one group of readers while magnifying discrimination across gender, ethnicity, age, geography, language, and disability [2,43]. The Wikimedia Foundation has acknowledged "*the potential of different forms of media to convey content to different audiences*" and recommends "*building the necessary technology to make free knowledge content accessible in various formats and support more diverse modes of consumption and contribution*". The multimedia gap is especially problematic for people with disabilities. Around 1 billion people (15% of the world's population)[7] experience some form of disability, of which 110-190M people experience significant disabilities. Facilitating equal access to knowledge via modality alternatives (images for those who cannot hear, sound for those who cannot see, tactile representations for those with combined impairments) is a right granted by the UN's Universal Declaration of Human Rights[8]. Yet existing KG approaches, including recent multimodal efforts, fall short to serve these populations. For instance, Multimodal Knowledge Graphs (MMKGs) in [35], cover merely few images.

To address these gaps, we propose **CUBE-MT (CUltural BEnchmark with Multimodal Transformations)**, a systematic framework that: (1) leverages multimodal generative models to automatically synthesize six modalities (text, images, Braille, speech, music, and 3D models) for cultural heritage concepts, thereby producing multimodal knowledge graphs that close multimedia gaps and support users with different sensory capabilities; and (2) systematically evaluates both the cultural awareness and cultural diversity of the resulting multimodal representations. Unlike prior work limited to images, CUBE-MT generates representations across modalities relevant for users with diverse disabilities while explicitly addressing cultural competence throughout the generation process. CUBE-MT builds upon and extends two complementary resources. First, we extend CUBE [33], a text-to-image (T2I) benchmark for evaluating cultural diversity and awareness, by adding 5 additional modalities (text descriptions,

[5] See e.g. Commission Recommendation of 10.11.2021 on a common European data space for cultural heritagehttps://ec.europa.eu/newsroom/dae/redirection/document/80911

[6] https://meta.wikimedia.org/wiki/Research:Recommending_Images_to_Wikipedia_Articles

[7] https://www.worldbank.org/en/topic/disability

[8] https://www.un.org/en/about-us/universal-declaration-of-human-rights

Braille, speech, music, and 3D models) for cultural artifacts from 8 countries (Brazil, France, India, Italy, Japan, Nigeria, Turkey, and the USA) across three domains (cuisine, landmarks, and art-cultural attire). Second, we extend the Muse-IT dataset [52], which originally made cultural heritage paintings accessible by converting them into Multimodal Digital Objects (MMDOs) through sound (sonification with harmonic mapping) and touch (haptification with vibration mapping). The Muse-IT artifacts were collected from 16 countries (Austria, Denmark, Germany, Ireland, Mexico, Netherlands, Poland, Russia, Spain, Switzerland, Brazil, France, Italy, Japan, UK, USA). This combined extension allows our benchmark to support the generation and evaluation of rich multimodal representations that capture diverse cultural aspects across countries and sensory channels. In summary, our contributions are:

- A systematic benchmark for assessing cultural awareness and diversity in multimodal knowledge graph generation and completion using generative models, including an extended evaluation framework for cultural diversity that goes beyond existing approaches
- An extension to the *modalities* supported by CUBE and Muse-IT, supporting 6 modalities: images, text, Braille, speech, music, and 3D—relevant for the provision of knowledge to other human senses (hearing, touch)
- A set of multimodal *prompts* to account for the cultural awareness and cultural diversity in generating data for those modalities
- A sample *dataset* resulting from one run of the benchmark with 10,942 multimodal representations linked to 802 unique Wikidata items for cultural awareness, and 9,600 representations for cultural diversity using publicly available generative models

2 Related Work

Knowledge equity concerns ensuring "*inclusive and equitable quality education and promote lifelong learning opportunities for all*" [9], including people with disabilities. Cultural heritage institutions are key enablers for such learning, but biases in their collections and databases are well known and documented [59]. In the Semantic Web, Wikidata [54] follows the Wikimedia Foundation's vision of "*a world in which every single human being can freely share in the sum of all knowledge*" [27] and can be considered a large cultural heritage knowledge base. However, Wikidata may contain knowledge gaps, i.e. differences in reader, contributor or content coverage, which can magnify biases and favour discrimination across various dimensions (e.g. gender, recency, geographic, socio-economic [2]). [43] proposes a taxonomy of knowledge gaps that includes multimedia gaps as potential blockers for conveying content to different audiences. Multimodality has been studied to address these gaps and comprehension in people with disabilities, by providing multimodal versions of podcasts (text and images) with generative models to aid understanding [10].

[9] UN's Sustainable Development Goal (SDG) 4: `https://sdgs.un.org/goals/goal4`

Generative models, thanks to the transformer architecture [53] and large-scale diffusion models [45], can generate realistic multimodal content as text, image, video, and audio [6,57,31,25], with state-of-the-art models for tasks like text-to-image, text-to-speech, text-to-audio and music, etc Multimodal generative models can be used for tasks such as acquisition, fusion and inference on Multimodal Knowledge Graphs [17]. However, these models are also under scrutiny due to potential harms [42]. Some of these harms can be seen as manifestations of cultural biases: the CUBE benchmark [33] aims at evaluating cultural competence of close and open-source text-to-image models in cultural awareness and cultural diversity, and is evaluated combining different measures.

Cultural heritage knowledge graphs [11] are a primary example of the challenges around cultural biases and how more polyvocal approaches are needed [24]. [5] gives an account of existing GLAM projects that support the publication of cultural heritage semantic data as Linked Data or KGs, typically led by large digital infrastructures like Europeana [41] and models like CIDOC-CRM [15] and EDM [22]. These models are typically extended to model specific cultural heritage collections, e.g. Italian [16] and Dutch [12] cultural heritage. Knowledge graphs of music [37] and its metadata [20,8] are representatives of a modality rarely considered within KGs. More commonly, MMKGs focus on the modalities of text and literals [29] and are often used to benchmark KG embedding models [30]). Beyond literals, MMPedia [56] is an image-rich MMKG that grounds entities in KGs on images found in DBpedia [7] and the Web. In [3], the closest work with ours, the authors propose a KG-based RAG architecture to derive prompts from Wikidata that are then used to generate missing images in item pages of fictional characters. Despite their relevance, none of these approaches address the issue of benchmarking cultural competence in MMKGs with multiple modalities, beyond literals and text, to address multimedia gaps.

3 The CUBE-MT Benchmark & Dataset

In this Section we describe CUBE-MT's construction process and properties.

3.1 Benchmark Construction

CUBE-MT extends CUBE [33], an existing benchmark for evaluating cultural competence of T2I models in cultural awareness and cultural diversity, together with Muse-IT dataset. The following section outlines the main components of CUBE-MT's construction.

Cultural Awareness Cultural awareness assesses a model's capacity to reliably and accurately portray objects and artifacts associated with specific cultures [33]. Figure 1 illustrates our pipeline for collecting entities and evaluating cultural awareness across diverse modalities, extending the methodology introduced in the CUBE benchmark [33]. The pipeline integrates two primary data sources. First, we leverage CUBE-1K, a curated dataset of high-quality prompts

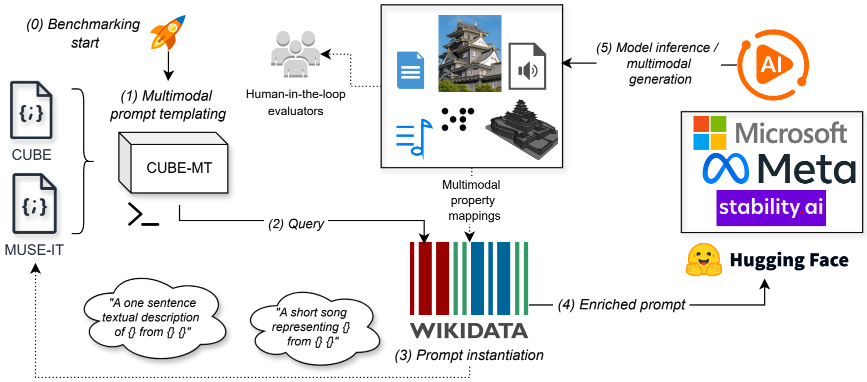

(0) Benchmarking start
(1) Multimodal prompt templating
Human-in-the-loop evaluators
CUBE
CUBE-MT
MUSE-IT
(2) Query
Multimodal property mappings
(5) Model inference / multimodal generation
Microsoft
Meta
stability ai
Hugging Face
(4) Enriched prompt
WIKIDATA
(3) Prompt instantiation
"A one sentence textual description of {} from {} {}"
"A short song representing {} from {} {}"

Fig. 1: The pipeline for cultural awareness, from collection to evaluation.

representing widely recognized cultural artifacts linked to grounded knowledge bases (Step 1 in Figure 1). Each prompt targets a specific entity (e.g., "A high resolution image of *carne de panela* from Brazilian cuisine, realistic") and is associated with its unique Wikidata item identifier (Q-id) and relevant properties (an example entry is given in subsection A.2). Second, we map artwork entities from the Muse-IT dataset to their corresponding Wikidata Q-ids. This alignment is achieved by traversing the Wikidata graph structure to identify corresponding entity linkages (Step 2 in Figure 1). The resulting unified CUBE-MT concept space encompasses 1,300 unique cultural artifacts spanning three domains: cuisine, art, and landmarks. The dataset covers 18 distinct geo-cultural regions, expanding upon the original 8 countries in CUBE (Brazil, France, India, Italy, Japan, Nigeria, Turkey, USA) [33] to include Austria, Denmark, Germany, Ireland, Mexico, the Netherlands, Poland, Russia, Spain, Switzerland, and the UK. Subsequently, we adopt RAG approach and initialize each prompt with its relevant properties and submit the enriched prompt for multimodal generation (Steps 3 and 4). The extended prompt templates used to generate entities for the multi-modal evaluation are detailed in Table 8 within subsection A.2, and the queried models are listed in Table 1. Finally, the generated multi-modal entities are evaluated for faithfulness and realism, with the corresponding proof of use detailed in Section 5.1.

Cultural Diversity Cultural diversity is defined as a model's ability to avoid oversimplified or homogenized representations of a culture. We benchmark cultural diversity to assess a model's ability to avoid stereotypical depictions when responding to "under-specified" prompts [33]. The pipeline is illustrated in Figure 2. First, we generate a set of under-specified prompts (Step 1 in Figure 2), as detailed in Table 9 within subsection A.2. These prompts are propagated to

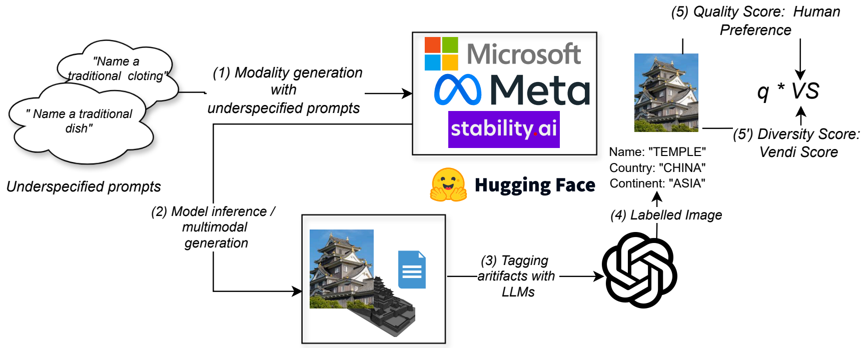

"Name a traditional cloting"
" Name a traditional dish"
Underspecified prompts
(1) Modality generation with underspecified prompts
Microsoft
Meta
stability ai
Hugging Face
(2) Model inference / multimodal generation
(3) Tagging aritifacts with LLMs
(4) Labelled Image
Name: "TEMPLE"
Country: "CHINA"
Continent: "ASIA"
(5) Quality Score: Human Preference
q * VS
(5') Diversity Score: Vendi Score

Fig. 2: The pipeline for cultural diversity, from collection to evaluation.

state-of-the-art open-source models hosted on Hugging Face to perform multimodal generation (Step 2). For this study, we extend the CUBE framework [33] to include text modalities but exclude audio and music, as robust automatic metrics for these modalities based on human preference scores are currently unavailable. Consequently, we benchmark only text and image outputs. Next, the generated artifacts undergo an automated tagging process (Step 3). We parse the artifacts using GPT-4, enforcing the extraction of structured metadata: *Name*, *Country*, and *Continent* (Step 4). Finally, to assess diversity both qualitatively and quantitatively, we compute the Quality-Weighted Vendi Score ($q \times VS$) (Step 5). This metric integrates a diversity score (Vendi Score) with a quality score derived from human preference proxies, following the methodology described in CUBE [33]. The evaluation process is described in Section 4.

3.2 Output Dataset and Properties

With its modalities, prompt templates, and RAG approach defined, CUBE-MT can be instantiated with any generative model and produce example MMKG outputs, in this study, we publish a dataset generated using the framework described below.

Models Table 1 shows the generative models we use for each modality. For this convenience, here we select models provided by the Hugging Face Serverless API[10]. Importantly, this list of models can (and should be) changed for each instantiation of the benchmark: the purpose of CUBE-MT is to offer a framework with replaceable models that facilitates their testing and evaluation. We encourage prospective users of CUBE-MT to experiment with other models and compare their results with ours (see Section 5).

[10] `https://huggingface.co/models`

Modality	**Generative model**	**URL**	**#Application**
Images	Qwen Image [55]	`https://huggingface.co/Qwen/Qwen-Image`	B
	Stability AI Stable Diffusion 3 Medium [26]	`https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers`	B
	Flux.1 Schnell [34]	`https://huggingface.co/black-forest-labs/FLUX.1-schnell`	B
Text	Google Gemma 2 [47]	`https://huggingface.co/google/gemma-2b`	D
	Meta LLama 3.1 Instruct [23]	`https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`	D
	Qwen 2.5 Instruct [48]	`https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`	D
	Microsoft Phi3 Mini4K Instruct [1]	`https://huggingface.co/microsoft/Phi-3-mini-4k-instruct`	A
	Qwen3-Next-80B [49]	`https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct`	A
	Microsoft Phi3 Mini4K Instruct [1]	`https://huggingface.co/microsoft/Phi-3-mini-4k-instruct`	A
Speech	Meta FastSpeech 2 [44]	`https://huggingface.co/facebook/fastspeech2-en-ljspeech`	A
	Kokoro-82M	`https://huggingface.co/hexgrad/Kokoro-82M`	A
	Chatterbox	`https://huggingface.co/ResembleAI/chatterbox`	A
Music	Meta MusicGen Small 300M [19]	`https://huggingface.co/facebook/musicgen-small`	A
	MusicGen Melody Large	`https://huggingface.co/ylacombe/musicgen-melody-large`	A
3D	Tencent Hunyuan3D 2.0 [58]	`https://huggingface.co/tencent/Hunyuan3D-2`	A
Braille	Algorithmic (using `PyBraille`)	`https://pypi.org/project/pybraille/`	A

Table 1: Models employed for modality generation and cultural competence evaluation. The final column, #Application, specifies the application scenario, where (A) denotes Cultural Awareness, (D) denotes Cultural Diversity, and (B) denotes both.

Execution dataset Executing the CUBE-MT benchmark yields 10,942 multimodal representations (images, text, Braille, speech, music, and 3D models) for 802 unique Wikidata/CUBE-1k/Muse-IT items across 15 generative models, as well as 9,600 multimodal representations for cultural awareness and cultural diversity. Figure 3 shows six example Wikidata items and their corresponding multimodal representations produced while running the benchmark. We make this dataset available in Dataverse and Hugging Face, with unique identifiers as shown in Table 2. The Muse-IT Dataverse repository automatically generates unique DOIs for each item in the resource, e.g. Kojinyama Fortress in Figure 3 is at `https://doi.org/10.5072/FK2/L9BLYI`. The dataset links back to all Wikidata items for which it generates modalities through the *id* field (e.g. `https://www.wikidata.org/wiki/Q71053154` for Kojinyama Fortress). An example of the extended metadata for one item is shown in subsection A.2.

Meroño-Peñuela, A. et al. (2025). *CUBE-MT: A Cultural Benchmark for Multimodal Knowledge Graph Construction with Generative Models.* doi:10.5281/zenodo.15398577

Resource	Link
CUBE-MT benchmark	https://github.com/albertmeronyo/CUBE-MT
CUBE-MT output dataset	https://dataverse.museit.eu/dataverse/cube-mt
Portal page	https://museit.eu/landing-page
Dataset generation code	https://github.com/albertmeronyo/CUBE-MT/blob/main/mt.ipynb
Documentation and tutorials	https://github.com/albertmeronyo/CUBE-MT/wiki
Full dump download	https://github.com/albertmeronyo/CUBE-MT/archive/refs/heads/main.zip
Subset of evaluation items	https://drive.google.com/drive/folders/1a78A7DtmHog8fACzxOEonRlojpKPn7vy?usp=sharing
Example de-referenceable resource	https://doi.org/10.5072/FK2/L9BLYI
Hugging Face & dataset card	https://huggingface.co/datasets/albertmeronyo/CUBE-MT
Zenodo	https://doi.org/10.5281/zenodo.15398577

Table 2: Links to key resources of the CUBE-MT benchmark.

4 Benchmarking Cultural Diversity

In this section, we present an automated framework to evaluate the cultural diversity of outputs from generative models, extending the methodology established in CUBE [33] to include text modalities.

Quality-Weighted Vendi Score We employ the Quality-Weighted Vendi Score (qVS) to simultaneously assess the quality and diversity of generated artifacts [40,28]. While the quality measure is defined by human preference scores specific to the modality, the diversity calculation relies on a composite similarity kernel. We utilize the kernel $k(x_i, x_j)$ defined in CUBE, which is weighted by terms $w_1, w_2,$ and w_3, corresponding to continent, country, and artifact similarity, respectively. We adopt the following weight configurations to enforce preferences for distinct cultural aspect from CUBE [33]:

- **Continent-level diversity** : $(w_1 = 1, w_2 = 0, w_3 = 0)$ Evaluates the distribution of generated artifacts across continents.
- **Country-level diversity** : $(w_1 = 0, w_2 = 1, w_3 = 0)$. Evaluates the distribution of generated artifacts across countries .
- **Artifact-level diversity** : $(w_1 = 0, w_2 = 0, w_3 = 1)$. Considers only the distinctness of the artifacts, disregarding geographical origin.
- **Hierarchical geographical diversity** : $(w_1 = 1/2, w_2 = 1/2, w_3 = 0)$. Captures a hierarchical notion of diversity where both continent and country similarities are penalized equally.
- **Uniformly weighted diversity** : $(w_1 = 1/3, w_2 = 1/3, w_3 = 1/3)$. Provides equal weight to all three similarity dimensions.

Text-to-Image Following CUBE [33], we assess cultural diversity across three concepts (Art, Cuisine, Landmarks) using under-specified prompts. For each concept, we generate 400 images by combining 5 prompt template variations (Table 1) with 10 seed batches (seeds 0–79, 8 images per batch). GPT-4o is used

Table 3: Breakdown of the mean quality component (q) and mean diversity component ($\overline{\text{VS}}$) averaged over repetitions for Image modality. SDXL refers to the Stable Diffusion XL model. Higher scores indicate better performance (greater diversity and quality).

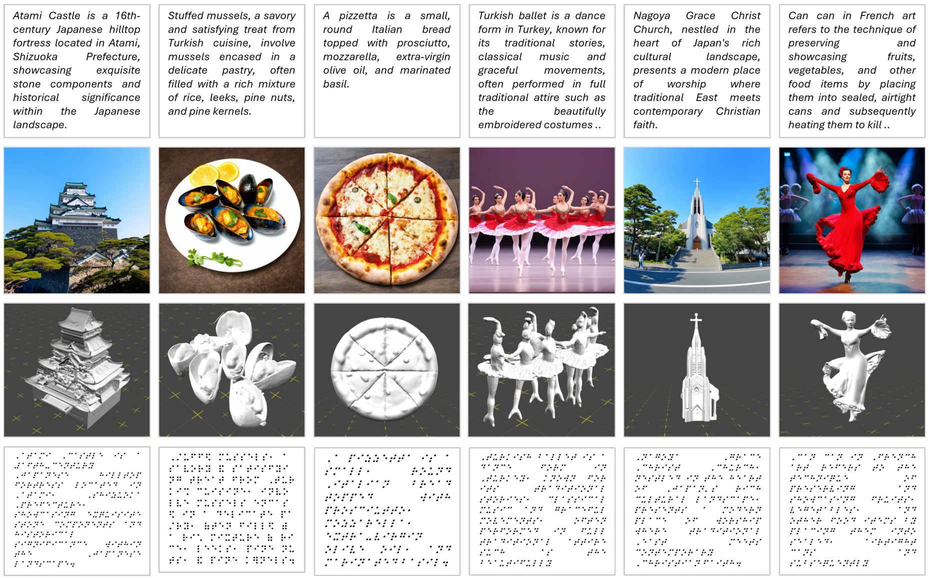
	Cuisine			**Landmarks**			**Art (CUBE)**			**Art (MUSE-IT)**		
	FLUX	**QWEN**	**SDXL**	**FLUX**	**QWEN**	**SDXL**	**FLUX**	**QWEN**	**SDXL**	**FLUX**	**QWEN**	**SDXL**
q (→)	0.2602	0.2826	0.2075	0.2409	0.2876	0.2605	0.1937	0.2908	0.2623	0.2152	0.2780	0.2623
$\overline{\text{VS}}(w_1, w_2, w_3)$												
$\overline{\text{VS}}$ (1, 0, 0)	0.2513	0.2616	0.3852	0.2523	0.2201	0.2047	0.1480	0.1913	0.1886	0.3108	0.2550	0.1965
$\overline{\text{VS}}$ (0, 1, 0)	0.6872	0.5048	0.7079	0.6718	0.4024	0.5738	0.5857	0.4666	0.4679	0.7635	0.5216	0.5737
$\overline{\text{VS}}$ (0, 0, 1)	0.9561	0.7385	0.9514	0.9157	0.7110	0.8979	0.8414	0.7465	0.7575	0.9773	0.7308	0.8224
$\overline{\text{VS}}$ ($\frac{1}{2}$, $\frac{1}{2}$, 0)	0.5503	0.4344	0.6235	0.5309	0.3515	0.4495	0.4126	0.3939	0.3725	0.6240	0.4427	0.4454
$\overline{\text{VS}}$ ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)	0.7597	0.5903	0.7918	0.7243	0.5186	0.6669	0.6125	0.5715	0.5570	0.8084	0.6002	0.6414

Table 4: Breakdown of the mean quality component (q) and mean diversity component ($\overline{\text{VS}}$) averaged over repetitions for Text modality. Model abbreviations are as follows: GEM (Google Gemma), LLAM (Meta Llama), and Qwen (Qwen Instruct). Higher scores indicate better performance (greater diversity and quality).

	Cuisine			**Landmarks**			**Art (CUBE)**			**Art (MUSE-IT)**		
	GEM	**LLAM**	**QWEN**	**GEM**	**LLAM**	**QWEN**	**GEM**	**LLAM**	**QWEN**	**GEM**	**LLAM**	**QWEN**
q (→)	0.0143	0.0127	0.0786	0.0155	0.0238	0.1448	0.0128	0.0190	0.0727	0.0120	0.0108	0.0621
$\overline{\text{VS}}(w_1, w_2, w_3)$												
$\overline{\text{VS}}$ (1, 0, 0)	0.3184	0.5364	0.1905	0.2345	0.3369	0.1752	0.2730	0.2752	0.1262	0.1626	0.1822	0.1326
$\overline{\text{VS}}$ (0, 1, 0)	0.4119	0.7785	0.2350	0.3529	0.3369	0.1845	0.4091	0.4308	0.1378	0.3477	0.4387	0.1821
$\overline{\text{VS}}$ (0, 0, 1)	0.5023	0.7785	0.2949	0.3874	0.4172	0.2009	0.4130	0.5120	0.1434	0.3546	0.4387	0.2164
$\overline{\text{VS}}$ ($\frac{1}{2}$, $\frac{1}{2}$, 0)	0.4122	0.7255	0.2241	0.3249	0.3369	0.1823	0.3940	0.3917	0.1348	0.2938	0.3651	0.1688
$\overline{\text{VS}}$ ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)	0.4672	0.7551	0.2621	0.3845	0.3851	0.1929	0.4371	0.4698	0.1404	0.3412	0.4430	0.1958

to identify the artifact name and corresponding country for each generated image, after which mSigLIP-based retrieval is applied to obtain the top-5 most similar reference images from CUBE-CSpace. Cultural diversity is measured using the quality-weighted Vendi Score (qVS), which integrates artifact diversity (computed via kernel-based similarity) with HPS-v2 image quality scores. We present the mean breakdown of quality and VS component values, averaged over repetitions, in Table 3; weighted results are provided in subsection A.1.

Text-to-Text For text generation, we follow the established workflow but exclude image generation, directly prompting models for cultural concepts (Name, Country, Origin) via the template in Table 1. We employ PairRM to calculate the quality score $s(x)$ for the qVS metric. Serving as a proxy for textual quality, PairRM aligns with the visual metrics of the CUBE framework by leveraging training on human preference data. Results are shown in Table 4 (for weighted results refer to subsection A.1).



Atami Castle is a 16th-century Japanese hilltop fortress located in Atami, Shizuoka Prefecture, showcasing exquisite stone components and historical significance within the Japanese landscape.

Stuffed mussels, a savory and satisfying treat from Turkish cuisine, involve mussels encased in a delicate pastry, often filled with a rich mixture of rice, leeks, pine nuts, and pine kernels.

A pizzetta is a small, round Italian bread topped with prosciutto, mozzarella, extra-virgin olive oil, and marinated basil.

Turkish ballet is a dance form in Turkey, known for its traditional stories, classical music and graceful movements, often performed in full traditional attire such as the beautifully embroidered costumes ..

Nagoya Grace Christ Church, nestled in the heart of Japan's rich cultural landscape, presents a modern place of worship where traditional East meets contemporary Christian faith.

Can can in French art refers to the technique of preserving and showcasing fruits, vegetables, and other food items by placing them into sealed, airtight cans and subsequently heating them to kill ..

Fig. 3: Multimodal transformations for the artefacts chosen for evaluation (shown here - Text, Image, 3D render and Braille; Speech and Music files available at `https://shorturl.at/wcH5V`)

Results The results presented in both tables indicate that contemporary T2I and Text-to-Text (T2T) models exhibit limited geo-continental diversity. The highest diversity scores observed were ≈0.39 for T2I models (Stable Diffusion) and ≈0.32 for T2T models (Google Gemma). However, T2T models demonstrated a relative advantage in producing diverse artifacts when conditioned on under-specified prompts. For example, Flux achieved ≈0.98 for artwork artifacts in the MUSt-IT benchmark, whereas the highest score among text models was only ≈0.50. This discrepancy suggests that, under our evaluation settings, current image generation models may encode comparatively less bias than text generation models, particularly when prompts provide limited contextual specificity for cultural artifacts.

5 Use and Impact

We evaluate and show evidence of usefulness and adoption of the CUBE-MT resources in two ways: (1) by running a survey through the network of partners in an Horizon Europe project which have adopted the benchmark, asking questions about cultural diversity and accuracy of generative models; and (2) by running a workshop with people with disabilities (a diverse range of aphasia-related challenges) on strengths and weaknesses for CUBE-MT.

5.1 Survey for Evaluation of Cultural Awareness

We evaluate the accuracy and diversity of the example dataset generated by the benchmark instantiated with the models of Section 3.2 with human experts. We collected a total of $N = 51$ responses from partners with a Semantic Web and/or Cultural Heritage background in the Horizon Europe Muse-IT project[11], covering a diverse set of artefacts and modalities (image, audio, 3D render). Responses were evenly distributed across the sample of 10 artefacts and their 3 modalities, ensuring a fair representation of perspectives. Participants were asked to rate their level of agreement with a series of statements using a Likert scale from 1 (very low agreement) to 5 (very high agreement), with an additional option of 0 for 'Not relevant.' This allowed us to capture both nuanced opinions and the relevance of each statement. The a sample of results are illustrated in Fig.4 (the complete set of results is provided in subsection A.3.). The multimodal representations were generally well received and considered useful. However, faithfulness ratings showed limited confidence in how accurately they reflected the original artefact texts, with few participants strongly agreeing. Emotional engagement was modest, with most responses indicating only low to moderate connection. Participants moderately agreed that key features were conveyed, though some important aspects may have been missed. Most also recognized the content as AI-generated, highlighting a clear distinction from human-created material. These findings suggest that while the representations are perceived as generally useful and informative, there is room for improvement in enhancing their faithfulness,

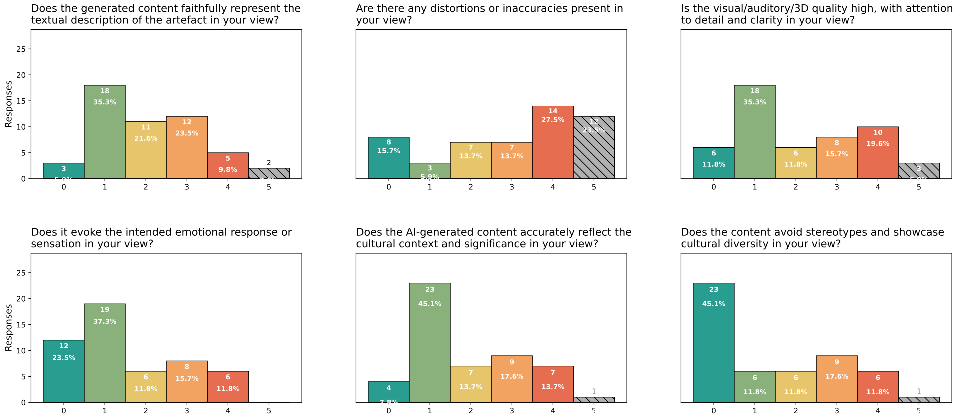


Fig. 4: Selection of questions from the online survey, where responses are expressed on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The total sample size is 51 (for the full set of results, refer to subsection A.3).

[11] `https://www.muse-it.eu/`

emotional impact, and subtlety to make them more aligned with human expectations and experiences.

5.2 Workshop with Target Beneficiaries

Here we describe the user study held with a focus group at the Aphasia Re-Connect charity's premises in London, which offers peer befriending and training to support people living with aphasia. Aphasia is a language disorder frequently resulting from stroke, impacting approximately one in three stroke survivors and causing varied impairments in speaking, understanding, reading, and writing [9]. We focus on people with aphasia, for whom difficulties with recall and expression are common. Multimodal enhancements (e.g., images paired with text) can help mitigate these challenges by supporting comprehension of text-heavy documents and websites [10].

We therefore choose CUBE-MT for its multimodality and follow a similar approach to [10] to explore (in)accuracies in its generated datasets. Three researchers (paper authors) facilitated the workshop, including two experienced Human-Computer Interaction (HCI) researchers with 8 years combined experience of supporting users with aphasia in design and testing technologies. They were responsible for guiding activities and probing participant reflections.

Participants Five individuals with aphasia (mean age = 60.4, SD = 3.6) participated in the focus group (see Fig. 5). Their mean Aphasia Severity Rating (ASR) score was 3.2 (SD = 0.8), indicating mild to moderate aphasia. They showed relatively preserved naming abilities but had difficulties with complex sentence construction, verbal expression, and auditory comprehension. Participants relied on visual cues and context, experienced word-finding challenges, and required extra time to process and respond, but all were able to actively engage in the discussion. We deliberately used a small group size to better support the communication needs of people with aphasia. Larger groups often create additional barriers, as participants with aphasia usually need more time and focused support to express their views [36].

Study Method The overall approach to this study draws from user-centred design, employing reflexive thematic analysis (TA) [13] to develop candidate topics/themes. These themes were then evaluated against three key criteria with the CUBE-MT dataset: (1) Are multimodal AI-generated formats helpful for people with aphasia? (2) Do they facilitate understanding of cultural objects? and (3) Does the presence of inaccuracies impact usefulness? This resulted in a final set of four themes.

We used a multimodal semantic guessing game, adapted from [50], to explore communication strategies in people with aphasia. Participants identified cultural artefacts (10 CUBE-MT items; Fig. 3) as modalities were gradually revealed (e.g., spoken description, images, 3D-printed object). By varying how

Fig. 5: Participants of the evaluation workshop (left); interactions with text and 3D printed models (right).

guessable each artefact was, we observed and discussed communication breakdowns, compensatory strategies, and how different representations supported comprehension and expression in aphasia. After each round, participants reflected on which modalities helped or hindered them, highlighting how layering modalities can scaffold understanding of complex cultural content.

Analysis The video-recorded sessions were transcribed using automated speech recognition (OpenWhisper), followed by manual review and correction. We then applied Thematic Analysis (TA) [14], an iterative method for identifying patterns and themes in qualitative data. Results are presented in Table 5. Following the six-step TA framework [14], we began by manually reading transcripts to identify recurring ideas. Initial observations were noted, generating insights relevant to the three requirements: (1), (2), and (3). Through systematic analysis, 24 codes were identified—each representing a label or phrase capturing explicit or implicit meanings tied to helpfulness, cultural understanding, or tolerance for inaccuracies. These codes were then grouped into broader patterns or themes. Four candidate themes were developed based on similarity, recurrence, and relevance. Themes were reviewed and refined for internal consistency, coherence, and clear differentiation. Final themes were explicitly defined, named, and summarized alongside corresponding codes to ensure analytical clarity and transparency.

Findings Thematic analysis reveals clear strengths and limitations of CUBE-MT modalities for users with aphasia. Visual formats (3D objects, pictures) significantly enhance accessibility and understanding, supporting engagement with cultural objects. In contrast, auditory formats remain challenging and may require improvement or clearer integration with visual/textual context. Accuracy, while important, did not critically affect usefulness as long as errors were minor. However, transparently acknowledging potential inaccuracies may be essential to maintain user trust.

Criteria 1: Are these AI-generated formats helpful for people with aphasia? Visual clarity and immediacy emerged as the most clearly helpful theme, indicat-

Theme	**Codes Matched to Theme**
Visual clarity and immediacy	Preference for images over text Immediate comprehension through visuals Practical understanding through 3D objects Easier cognitive load with images Initial attention to visual elements Images give clear cultural context Scale and physical attributes clear from visuals
Challenges of auditory representations	Ambiguity of auditory information Difficulty identifying culture from sound Music unclear for cultural identification Incorrect guesses from audio examples Auditory details often misleading Audio representations less informative than visual Cultural ambiguity due to unclear audio
Tolerance for inaccuracies	Visual inaccuracies better than absence of information Misleading visual details still somewhat useful Errors not affecting overall cultural understanding Tolerance higher for visual than textual inaccuracies Inaccuracies do not significantly hinder comprehension
Skepticism and trust issues	Skepticism about Wikipedia-style editing Concern about misinformation risk Trust issues limiting acceptance of AI outputs Questioning accuracy of textual information Suspicion toward AI reliability

Table 5: Initial Codes (24) Matched to Themes (4)

ing that images and 3D models significantly facilitate immediate understanding. Participants noted these formats reduced the cognitive load of reading textual descriptions. However, auditory formats were considerably less helpful, as they led to confusion and difficulty clearly interpreting cultural contexts.

Criteria 2: Do they make it easier to understand cultural objects? Participants clearly indicated improved understanding of cultural objects through visual representations (images, 3D) due to their clarity, tangible attributes, and immediate accessibility. Visual cues provided substantial support, especially where textual information was limited or cognitively demanding. Conversely, auditory representations did not simplify understanding; rather, they often increased ambiguity and uncertainty regarding the cultural objects discussed.

Criteria 3: Do inaccuracies in generations affect how useful these modalities are? Participants generally demonstrated tolerance for small inaccuracies, especially in visual formats, emphasizing that minor errors were not critically detrimental to overall usefulness. They preferred imperfect visuals over no information at all, suggesting pragmatic acceptance. However, they were less forgiving of inaccuracies in textual information and auditory contexts, highlighting skepticism and potential issues of misinformation and distrust. Clear accuracy expectations existed for text-based formats and cultural accuracy in sounds. Overall, minor inaccuracies in visuals were acceptable, whereas inaccuracies in textual and auditory forms significantly affected trust and perceived reliability.

6 Availability, Sustainability, and FAIRness

All CUBE-MT resources are available online[12], including the benchmark, dataset, and links to the original Wikidata items. All resources are created and made usable within the Wikidata namespace at `https://www.wikidata.org/wiki/`. Permanent URIs (DOIs) are created automatically when depositing to Dataverse for each individual item, and globally for the deposits in Zenodo, Datahub and Hugging Face as shown in Section 3.2. CUBE-MT is under version control on public GitHub repositories[13] (see Table 2). All resources are published under the CC-BY 4.0 licence and in accordance with usage of outputs of all generative models of Table 1. The storage of all resources on GitHub, Hugging Face and Dataverse guarantees their persistence beyond the project.

7 Conclusion

In this paper we present CUBE-MT, a benchmark and resulting dataset for evaluating cultural awareness and diversity of generative models in multiple modalities for multimodal knowledge graph construction. We extend an existing benchmark with 6 modalities (images, text, speech, music, Braille, and 3D models), corresponding prompts, and a Wikidata-based RAG architecture for MMKG construction and completion. To evaluate CUBE-MT, we run an expert survey among internal project users, finding that the benchmark can be used to identify cultural inaccuracies and diversity issues in current generative models; and a workshop with people with aphasia that illustrates that CUBE-MT outputs hold considerable promise for enhancing learning outcomes and improving accessibility to cultural heritage for users with different types of disabilities and to promote inclusivity. Image and tactile 3D prints of generated 3D models were found more useful than audio representations, and users were more tolerant to inaccuracies in them.

Many opportunities and challenges remain open for the future. First, we will address the inherent limitation of human evaluations in MMKGs generated with CUBE-MT, investigating metrics and evaluation frameworks that have been previously raised [3]. Second, we will devise multimodal integrations of CUBE-MT with tools such as Langchain[14] and the Model Context Protocol[15] (MCP). Finally, we will devise ways to enhance trust, ethical scrutiny, and governance in CUBE-MT's generations by using, and extending, existing open standards and licenses for AI-readiness [38,39], such as MLCommons Croissant [4].

Acknowledgements This work was partially supported by the HE project Muse-IT, co-funded by the European Union under Grant Agreement No. 101061441. The views and opinions expressed in this work are those of the authors and do not necessarily reflect the views of the European Union or the European Research Executive Agency. This project has received funding from the EPSRC through the CA11y project (EP/X012395/1).

[12] `https://dataverse.museit.eu/dataverse/cube-mt`

[13] `https://github.com/albertmeronyo/CUBE-MT`

[14] `https://python.langchain.com/docs/introduction/`

[15] `https://modelcontextprotocol.io/introduction`

References

1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.C., Chen, Y.L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A.D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R.J., Hu, W., Huynh, J., Iter, D., Jacobs, S.A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y.J., Kurilenko, L., Lee, J.R., Lee, Y.T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C.C.T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L.L., Zhang, Y., Zhang, Y., Zhang, Y., Zhou, X.: Phi-3 technical report: A highly capable language model locally on your phone (2024), `https://arxiv.org/abs/2404.14219`
2. Abián, D., Meroño-Peñuela, A., Simperl, E.: An analysis of content gaps versus user needs in the wikidata knowledge graph. In: International Semantic Web Conference. pp. 354–374. Springer (2022)
3. Ahmad, R.A., Critelli, M., Efeoglu, S., Mancini, E., Ringwald, C., Zhang, X., Meroño Peñuela, A.: Draw me like my triples: Leveraging generative ai for wikidata image completion. In: The 4th Wikidata Workshop (2023)
4. Akhtar, M., Benjelloun, O., Conforti, C., Foschini, L., Giner-Miguelez, J., Gijsbers, P., Goswami, S., Jain, N., Karamousadakis, M., Kuchnik, M., et al.: Croissant: A metadata format for ml-ready datasets. Advances in Neural Information Processing Systems **37**, 82133–82148 (2024)
5. Alexiev, V.: Museum linked open data: Ontologies, datasets, projects. Digital Presentation and Preservation of Cultural and Scientific Heritage (VIII), 19–50 (2018)
6. Anil, R., Borgeaud, S., Wu, Y., Alayrac, J., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., Silver, D., Petrov, S., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T.P., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P.R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., von Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., et al.: Gemini: A family of highly capable multimodal models. CoRR **abs/2312.11805** (2023). `https://doi.org/10.48550/ARXIV.2312.11805`, `https://doi.org/10.48550/arXiv.2312.11805`
7. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: international semantic web conference. pp. 722–735. Springer (2007)
8. de Berardinis, J., Carriero, V.A., Jain, N., Lazzari, N., Meroño-Peñuela, A., Poltronieri, A., Presutti, V.: The polifonia ontology network: Building a semantic

backbone for musical heritage. In: International Semantic Web Conference. pp. 302–322. Springer (2023)

9. Berthier, M.L.: Poststroke aphasia: epidemiology, pathophysiology and treatment. Drugs & aging **22**, 163–182 (2005)
10. Bircanin, F., Nevsky, A., Perera, H., Agarwal, V., Song, E., Cruice, M., Neate, T.: Sounds accessible: Envisioning accessible audio-media futures with people with aphasia. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3706598.3714000, https://doi.org/10.1145/3706598.3714000
11. de Boer, V.: Knowledge graphs for cultural heritage and digital humanities. In: SUMAC@ ACM Multimedia. p. 3 (2023)
12. de Boer, V., Wielemaker, J., van Gent, J., Oosterbroek, M., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Amsterdam museum linked open data. Semantic Web **4**(3), 237–243 (2013)
13. Bowman, R., Nadal, C., Morrissey, K., Thieme, A., Doherty, G.: Using thematic analysis in healthcare hci at chi: A scoping review. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–18 (2023)
14. Braun, V., Clarke, V.: Reflecting on reflexive thematic analysis. Qualitative research in sport, exercise and health **11**(4), 589–597 (2019)
15. Bruseker, G., Carboni, N., Guillem, A.: Cultural heritage data management: the role of formal ontology and cidoc crm. Heritage and archaeology in the digital age: acquisition, curation, and dissemination of spatial cultural heritage data pp. 93–131 (2017)
16. Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: Arco: The italian cultural heritage knowledge graph. In: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18. pp. 36–52. Springer (2019)
17. Chen, Z., Zhang, Y., Fang, Y., Geng, Y., Guo, L., Chen, X., Li, Q., Zhang, W., Chen, J., Zhu, Y., Li, J., Liu, X., Pan, J.Z., Zhang, N., Chen, H.: Knowledge graphs meet multi-modal learning: A comprehensive survey. CoRR **abs/2402.05391** (2024). https://doi.org/10.48550/ARXIV.2402.05391, https://doi.org/10.48550/arXiv.2402.05391
18. Cole, J.B., Lott, L.L.: Diversity, equity, accessibility, and inclusion in museums. Rowman & Littlefield (2019)
19. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., Défossez, A.: Simple and controllable music generation (2024), https://arxiv.org/abs/2306.05284
20. Daquino, M., Daga, E., d'Aquin, M., Gangemi, A., Holland, S., Laney, R., Penuela, A.M., Mulholland, P.: Characterizing the landscape of musical data on the web: State of the art and challenges (2017)
21. Daquino, M., Mambelli, F., Peroni, S., Tomasi, F., Vitali, F.: Enhancing semantic expressivity in the cultural heritage domain: exposing the zeri photo archive as linked open data. Journal on Computing and Cultural Heritage (JOCCH) **10**(4), 1–21 (2017)
22. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Van de Sompel, H.: The europeana data model (edm): object representations, context and semantics. Gothenburg; 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY (2010)

23. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv–2407 (2024)
24. van Erp, M., de Boer, V.: A polyvocal and contextualised semantic web. In: The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18. pp. 506–512. Springer (2021)
25. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net (2024), https://openreview.net/forum?id=FPnUhsQJ5B
26. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), https://arxiv.org/abs/2403.03206
27. Foundation, W.: Wikimedia vision. https://wikimediafoundation.org/about/vision/ (2018), retrieved May 9, 2025
28. Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning. arXiv preprint arXiv:2210.02410 (2022)
29. Gesese, G.A., Alam, M., Sack, H.: Literallywikidata-a benchmark for knowledge graph completion using literals. In: The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20. pp. 511–527. Springer (2021)
30. Gesese, G.A., Biswas, R., Alam, M., Sack, H.: A survey on knowledge graph embeddings with literals: Which model links better literal-ly? Semantic Web **12**(4), 617–647 (2021)
31. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind one embedding space to bind them all. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 15180–15190. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01457, https://doi.org/10.1109/CVPR52729.2023.01457
32. Haslhofer, B., Isaac, A.: data. europeana. eu: The europeana linked open data pilot. In: Proceedings of the international conference on Dublin Core and metadata applications. Dublin Core Metadata Initiative (2011)
33. Kannen, N., Ahmad, A., Andreetto, M., Prabhakaran, V., Prabhu, U., Dieng, A.B., Bhattacharyya, P., Dave, S.: Beyond aesthetics: Cultural competence in text-to-image models (2024), https://arxiv.org/abs/2407.06863
34. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
35. Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D., Rosenblum, D.S.: Mmkg: Multi-modal knowledge graphs (2019), https://arxiv.org/abs/1903.05485
36. Luck, A.M., Rose, M.L.: Interviewing people with aphasia: Insights into method adjustments from a pilot study. Aphasiology **21**(2), 208–224 (2007)
37. Meroño-Peñuela, A., Hoekstra, R., Gangemi, A., Bloem, P., de Valk, R., Stringer, B., Janssen, B., de Boer, V., Allik, A., Schlobach, S., Page, K.: The midi linked

data cloud. In: The Semantic Web – ISWC 2017. Springer International Publishing, Cham (2017)

38. Meroño-Peñuela, A., Massey, J., Newman, A., Simperl, E.: How an ai-ready national data library would help uk science. arXiv preprint arXiv:2501.17013 (2025)
39. Meroño-Peñuela, A., Simperl, E., Kurteva, A., Reklos, I.: Kg. gov: Knowledge graphs as the backbone of data governance in ai. Journal of Web Semantics **85**, 100847 (2025)
40. Nguyen, Q., Dieng, A.B.: Quality-weighted vendi scores and their application to diverse experimental design. arXiv preprint arXiv:2405.02449 (2024)
41. Purday, J.: Think culture: Europeana. eu from concept to construction (2009)
42. Quaye, J., Parrish, A., Inel, O., Rastogi, C., Kirk, H.R., Kahng, M., Van Liemt, E., Bartolo, M., Tsang, J., White, J., Clement, N., Mosquera, R., Ciro, J., Janapa Reddi, V., Aroyo, L.: Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. p. 388–406. FAccT '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3630106.3658913, https://doi.org/10.1145/3630106.3658913
43. Redi, M., Gerlach, M., Johnson, I., Morgan, J., Zia, L.: A taxonomy of knowledge gaps for wikimedia projects (second draft) (2021), https://arxiv.org/abs/2008.12314
44. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y.: Fastspeech: Fast, robust and controllable text to speech. In: NeurIPS 2019 (November 2019), https://www.microsoft.com/en-us/research/publication/fastspeech-fast-robust-and-controllable-text-to-speech/
45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 10674–10685. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01042, https://doi.org/10.1109/CVPR52688.2022.01042
46. Sandell, R., Nightingale, E.: Museums, equality and social justice. Taylor & Francis (2012)
47. Team, G.: Gemma (2024). https://doi.org/10.34740/KAGGLE/M/3301, https://www.kaggle.com/m/3301
48. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
49. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
50. Turnbull, D., Liu, R., Barrington, L., Lanckriet, G.R.: A game-based approach for collecting semantic annotations of music. In: ISMIR. vol. 7, pp. 535–538. Citeseer (2007)
51. Van Den Akker, C., Legêne, S., Van Erp, M., Aroyo, L., Segers, R., Van der Meij, L., Van Ossenbruggen, J., Schreiber, G., Wielinga, B., Oomen, J., et al.: Digital hermeneutics: Agora and the online understanding of cultural heritage. In: Proceedings of the 3rd International Web Science Conference. pp. 1–7 (2011)
52. Van Erven, T., Darányi, S.: Turning paintings into multimodal digital objects. In: International Conference on Human-Computer Interaction. pp. 185–201. Springer (2025)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

54. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
55. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), `https://arxiv.org/abs/2508.02324`
56. Wu, Y., Wu, X., Li, J., Zhang, Y., Wang, H., Du, W., He, Z., Liu, J., Ruan, T.: Mmpedia: A large-scale multi-modal knowledge graph. In: International semantic web conference. pp. 18–37. Springer (2023)
57. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v(ision). CoRR **abs/2309.17421** (2023). `https://doi.org/10.48550/ARXIV.2309.17421`, `https://doi.org/10.48550/arXiv.2309.17421`
58. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., Shi, H., Liu, S., Wu, J., Lian, Y., Yang, F., Tang, R., He, Z., Wang, X., Liu, J., Zuo, X., Chen, Z., Lei, B., Weng, H., Xu, J., Zhu, Y., Liu, X., Xu, L., Hu, C., Yang, S., Zhang, S., Liu, Y., Huang, T., Wang, L., Zhang, J., Chen, M., Dong, L., Jia, Y., Cai, Y., Yu, J., Tang, Y., Zhang, H., Ye, Z., He, P., Wu, R., Zhang, C., Tan, Y., Xiao, J., Tao, Y., Zhu, J., Xue, J., Liu, K., Zhao, C., Wu, X., Hu, Z., Qin, L., Peng, J., Li, Z., Chen, M., Zhang, X., Niu, L., Wang, P., Wang, Y., Kuang, H., Fan, Z., Zheng, X., Zhuang, W., He, Y., Liu, T., Yang, Y., Wang, D., Liu, Y., Jiang, J., Huang, J., Guo, C.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation (2025), `https://arxiv.org/abs/2501.12202`
59. Zhitomirsky-Geffet, M., Kizhner, I., Minster, S.: What do they make us see: a comparative study of cultural bias in online databases of two large museums. Journal of Documentation **79**(2), 320–340 (2023)

A Additional Tables and Results

A.1 Cultural Diversity

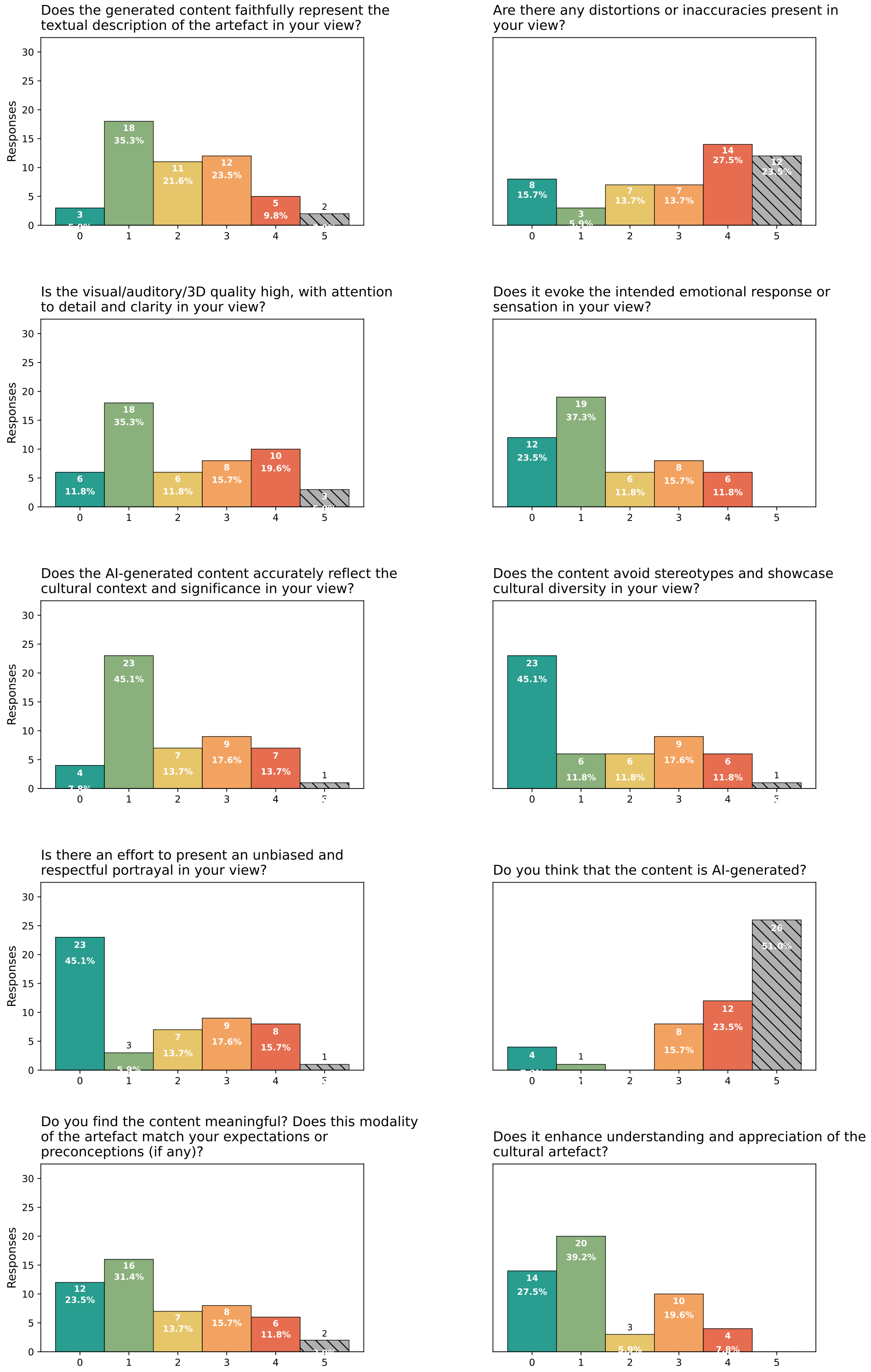
Additional tables reporting the evaluation of cultural diversity, with the diversity component weighted by quality, are provided in Table 6 for image modalities and in Table 7 for text.

A.2 CUBE-MT space Examples

An illustrative example of a CUBE-MT item related to cultural awareness is shown in Listing 1.1. Prompt templates addressing cultural awareness are presented in Table 8, while those pertaining to cultural diversity are provided in Table 9.

A.3 User study results

The complete results of the user study are presented in Fig. 6.


Does the generated content faithfully represent the textual description of the artefact in your view?

Are there any distortions or inaccuracies present in your view?

Is the visual/auditory/3D quality high, with attention to detail and clarity in your view?

Does it evoke the intended emotional response or sensation in your view?

Does the AI-generated content accurately reflect the cultural context and significance in your view?

Does the content avoid stereotypes and showcase cultural diversity in your view?

Is there an effort to present an unbiased and respectful portrayal in your view?

Do you think that the content is AI-generated?

Do you find the content meaningful? Does this modality of the artefact match your expectations or preconceptions (if any)?

Does it enhance understanding and appreciation of the cultural artefact?

Fig. 6: Selection of questions from the online survey, where responses are expressed on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The total sample size is 51.

Table 6: Breakdown of the diversity score over multiple repetitions for images, weighted by the quality score ($q\overline{\mathrm{VS}}$) for Images. SDXL refers to the Stable Diffusion XL model. Higher scores indicate better performance (greater diversity and quality).

	Cuisine			**Landmarks**			**Art (CUBE)**			**Art (MUSE-IT)**		
	FLUX	**QWEN**	**SDXL**	**FLUX**	**QWEN**	**SDXL**	**FLUX**	**QWEN**	**SDXL**	**FLUX**	**QWEN**	**SDXL**
q (→)	0.2602	0.2826	0.2075	0.2409	0.2876	0.2605	0.1937	0.2908	0.2623	0.2152	0.2780	0.2623
$\overline{\mathrm{qVS}}(w_1, w_2, w_3)$												
$\overline{\mathrm{qVS}}$ (1, 0, 0)	0.0656	0.0736	0.0787	0.0611	0.0632	0.0538	0.0292	0.0554	0.0496	0.0665	0.0697	0.0508
$\overline{\mathrm{qVS}}$ (0, 1, 0)	0.1790	0.1409	0.1460	0.1617	0.1161	0.1504	0.1146	0.1361	0.1259	0.1643	0.1416	0.1490
$\overline{\mathrm{qVS}}$ (0, 0, 1)	0.2490	0.2063	0.1968	0.2202	0.2046	0.2341	0.1615	0.2171	0.1990	0.2105	0.1999	0.2168
$\overline{\mathrm{qVS}}$ ($\frac{1}{2}$, $\frac{1}{2}$, 0)	0.1435	0.1215	0.1284	0.1280	0.1013	0.1179	0.0812	0.1146	0.0998	0.1342	0.1204	0.1155
$\overline{\mathrm{qVS}}$ ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)	0.1979	0.1650	0.1635	0.1744	0.1492	0.1743	0.1190	0.1662	0.1474	0.1740	0.1639	0.1678

Table 7: Breakdown of the diversity score over multiple repetitions for text, weighted by the quality score. Model abbreviations are as follows: GEM (Google Gemma), LLAM (Meta Llama), and Qwen (Qwen Instruct). Higher scores indicate better performance (greater diversity and quality).

	Cuisine			**Landmarks**			**Art (CUBE)**			**Art (MUSE-IT)**		
	GEM	**LLAM**	**QWEN**	**GEM**	**LLAM**	**QWEN**	**GEM**	**LLAM**	**QWEN**	**GEM**	**LLAM**	**QWEN**
q (→)	0.0143	0.0127	0.0786	0.0155	0.0238	0.1448	0.0128	0.0190	0.0727	0.0120	0.0108	0.0621
$\overline{\mathrm{qVS}}(w_1, w_2, w_3)$												
$\overline{\mathrm{qVS}}$ (1, 0, 0)	0.0046	0.0070	0.0115	0.0032	0.0083	0.0215	0.0033	0.0052	0.0091	0.0021	0.0020	0.0081
$\overline{\mathrm{qVS}}$ (0, 1, 0)	0.0059	0.0101	0.0131	0.0045	0.0083	0.0228	0.0050	0.0081	0.0096	0.0045	0.0047	0.0103
$\overline{\mathrm{qVS}}$ (0, 0, 1)	0.0074	0.0101	0.0146	0.0053	0.0100	0.0241	0.0049	0.0098	0.0097	0.0047	0.0047	0.0114
$\overline{\mathrm{qVS}}$ ($\frac{1}{2}$, $\frac{1}{2}$, 0)	0.0059	0.0095	0.0127	0.0042	0.0083	0.0225	0.0048	0.0074	0.0095	0.0039	0.0039	0.0097
$\overline{\mathrm{qVS}}$ ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)	0.0068	0.0098	0.0137	0.0051	0.0093	0.0234	0.0052	0.0089	0.0096	0.0045	0.0048	0.0107

Modality	**Prompt pattern**	**Variables**
Images	"A high resolution image of {} from {} cuisine, realistic"	English label, P17 (country), P31/P279 ∈ (dish, food, type of food or dish)) **[cuisine]**
	"An image from {} performance in {}, realistic"	English label, P17 (country), P31/P279 ∈ (type of dance, folk art, performing arts genre) **[art]**
	"An image of {} from {} clothing, realistic"	English label, P17 (country), P31/P279 ∈ (clothing, costume, traditional costume) **[art]**
	"A panoramic view of {} in {}, realistic"	English label, P17 (country), P31/P279 ∈ (see [33] for landmark concept list) **[landmarks]**
	"A painting of {} by {} from {} period in {} style, realistic"	English label, P17 (country), P31/P279 ∈ (art genre, for details see [52]) **[art]**
Text	"A one sentence textual description of {} from {} {}"	English label, P17 (country), domain (art, cuisine, landscapes)
Music	"A short song representing {} from {} {}"	English label, P17 (country), domain (art, cuisine, landscapes)

Table 8: Prompt templates used for cultural awareness, querying across modalities: Art, Cuisine, and Landmarks (adopted from CUBE [33] and extended with Muse-IT [52].

```
{
 "P31": "['Q57831']",
 "P279": "[]",
 "P495": "[]",
 "P17": "['Q17']",
 "P2012": " ",
 "P361": "[]",
 "id": "Q71053154",
 "name": "Kojinyama Fortress",
 "country": "Japan",
 "domain": "landscapes",
 "prompt": "A panoramic view of Kojinyama Fortress in Japan, realistic",
 "prompt_text": "A one sentence textual description of Kojinyama Fortress from
     ↪ Japanese landscapes",
 "gen_text": "txt/Q71053154.txt",
 "gen_braille": "braille/Q71053154.txt",
 "prompt_speech": "Kojinyama Castle, located in Nara Prefecture, dates back to
     ↪ the Nara period and is renowned for its elegant hilltop design and
     ↪ reflections of the surrounding beauty in its moats and gardens.",
 "gen_speech": "speech/Q71053154.wav",
 "prompt_image": "A panoramic view of Kojinyama Fortress in Japan, realistic",
 "gen_image": "img/Q71053154.png",
 "prompt_music": "A short song representing Kojinyama Fortress from Japanese
     ↪ landscapes",
 "gen_music": "music/Q71053154.wav",
 "gen_3d": "3d/Q71053154.glb"
}
```

Listing 1.1: Example of a CUBE-MT item based on Wikidata Q71053154, "Kojinyama Fortress". 'prompt_' and 'gen_' attributes contain instantiated modality prompts and paths to generated objects, respectively.

Art	**Cuisine**	**Landmarks**	**Muse-IT (Art)**
Name a traditional clothing	Name one traditional dish	Name an architectural heritage site	Name a painting
State the name of a traditional outfit	State one popular meal	State a famous landmark	State one cultural painting
Give me the name of a traditional attire	Give me the name of a festive food	Give me a historic site	Give me the name of a local painting
Name a traditional garment	Name a meal	Name a tourist attraction	Name an artwork
Name a cultural wear	Name a local food	Name a monument	Name a cultural artwork

Table 9: Under-specified prompt templates used for cultural diversity evaluation (adapted from CUBE-MT [33] and extended with Muse-IT-MT concepts [52]).