Navigating Cultural Chasms: Exploring and Unlocking the Cultural POV of Text-To-Image Models

Mor Ventura[1] **Eyal Ben-David**[1] **Anna Korhonen**[2] **Roi Reichart**[1]
[1]Faculty of Data and Decision Sciences, Technion, IIT, Israel
[2]Language Technology Lab, University of Cambridge, UK
mor.ventura@campus.technion.ac.il

Abstract

Text-To-Image (TTI) models, such as DALL-E and StableDiffusion, have demonstrated remarkable prompt-based image generation capabilities. Multilingual encoders may have a substantial impact on the cultural agency of these models, as language is a conduit of culture. In this study, we explore the cultural perception embedded in TTI models by characterizing culture across three tiers: cultural dimensions, cultural domains, and cultural concepts. Based on this ontology, we derive prompt templates to unlock the cultural knowledge in TTI models, and propose a comprehensive suite of evaluation techniques, including intrinsic evaluations using the CLIP space, extrinsic evaluations with a Visual-Question-Answer models and human assessments, to evaluate the cultural content of TTI-generated images. To bolster our research, we introduce the CulText2I dataset, based on six diverse TTI models and spanning ten languages. Our experiments provide insights regarding *Do*, *What*, *Which*, and *How* research questions about the nature of cultural encoding in TTI models, paving the way for cross-cultural applications of these models.[1]

1 Introduction

''We seldom realize, for example that our most private thoughts and emotions are not actually our own. For we think in terms of languages and images which we did not invent, but which were given to us by our society.'' (Watts, 1989)

Generative Text-To-Image models (TTI, e.g., DALL-E [Ramesh et al., 2021, 2022] and Stable-Diffusion [Rombach et al., 2021]) have recently witnessed a surge in popularity, due to their remarkable zero-shot capabilities. They are guided by textual prompts to generate images, offering a visual representation of their textual interpretation.

TTI models exhibit multilingual proficiency, acquired *explicitly*, through the model architecture and objective, or *implicitly*, through exposure to multiple languages only (§2). These models find widespread use in domains such as art, education, and communication, exerting substantial societal influence (Ko et al., 2023; Vartiainen and Tedre, 2023; Maharana et al., 2022). Their cultural significance stems from their multilingual competence and extensive adoption, as language is a vessel for cultural identity and heritage. Indeed, Yiu et al. (2023) have demonstrated AI models' pivotal role in enhancing cultural transmission.

This study aims to gain insight into the cultural perception inherent in TTI models. We embark on a novel characterization, dissecting the complex ties between language, culture, and TTI models. Our approach is inspired by well-established cultural research (Hofstede, 1983; Rokeach, 1967; Haerpfer et al., 2022), allowing us to systematically deconstruct the wide notion of culture across three tiers, progressing from the broader to the finer levels of abstractness: *cultural dimensions, cultural domains*, and *cultural concepts*.

This ontology allows us to derive prompt templates, with which we aim to unlock the cultural knowledge encoded within TTI models, and a suite of evaluation measures which reflects different aspects of the cultural information in a generated image. These methods consist of intrinsic evaluations using the CLIP (Radford et al., 2021; Cherti et al., 2022) space, extrinsic evaluations with Visual Question Answering (VQA) models (Li et al., 2023; Achiam et al., 2023), and human assessments.

To address the lack of an appropriate dataset, we introduce CulText2I, comprising images generated by six distinct TTI models, varying in multilingual capability and architecture (§2). These models include StableDiffusion 2.1v and 1.4v, AltDiffusion, DeepFloyd, DALL-E, and a Llama

[1] Our code and data are available at https://github.com/venturamor/CulText-2-I.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 142–166, 2025. https://doi.org/10.1162/tacl_a_00732
Action Editor: Luke Zettlemoyer. Submission batch: 1/2024; Revision batch: 8/2024; Published 2/2025.

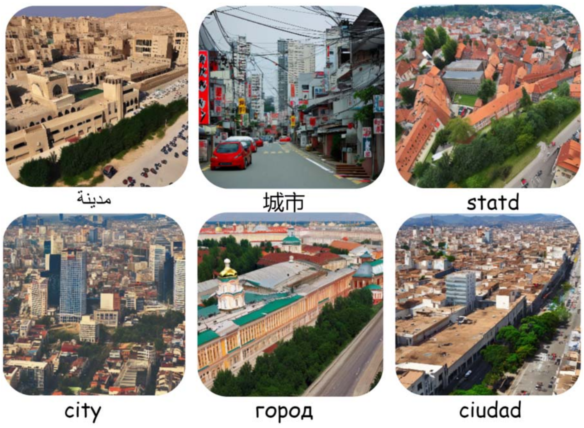


Figure 1: StableDiffusion 2.1v images of *‘‘A photo of <city>’’*, while *city* is translated to (left to right) Arabic, Chinese, German (top) English, Russian, Spanish (bottom).

2 + SD 1.4 UNet model based on Llavi-Bridge (Rombach et al., 2021; Ye et al., 2023; DeepFloyd, 2023; Ramesh et al., 2022; Zhao et al., 2024). We generate images using prompts representing identical cultural concepts across ten languages (see Figure 1).

In §3 we present our research questions, aiming to address the way cultural knowledge is encoded by and can be unlocked in TTI models, the effective ways of unlocking this knowledge and the resulting conclusions about the world cultures. §4 presents our cultural ontology, and then §5 and §6 present the derived prompt templates and evaluation measures, respectively. §7, §8, and §9 present our experiments, results, and potential factors behind our key findings, which demonstrate the cultural capacity of TTI models, the important role of the multilingual textual encoder, and the impact of the different unlocking decisions as manifested by the prompt design. By interrogating the cultural nuances within TTI models, we hope to challenge existing NLP paradigms and inspire innovative applications that harness the models’ potential for cross-cultural understanding.

2 Related Work

Generative Text-To-Image Models TTI models typically incorporate two core components: an *LM-based text encoder*, which interprets and processes linguistic inputs; and an *image generator* (typically based on diffusion) that synthesizes corresponding images.

Multilingual text encoders (Devlin, 2018; Xue et al., 2020; Goyal et al., 2021; Scao et al., 2022) opened the door to a wide range of cultural influences on TTI models, which are at the heart of this study. The multilingual capabilities, varying in their semantic interpretation and how well they align images to the requested concept in the prompt (i.e., *conceptual coverage*) (Saxon and Wang, 2023) are acquired either by *explicit* training objectives (e.g., DeepFloyd IF [DeepFloyd, 2023] and AltDiffusion [Ye et al., 2023]), such as in the training of XLM-R and T5, or *implicitly*—only through exposure to different languages in their training corpus, as in the case of StableDiffusion v2.1 or v1.4 (Rombach et al., 2021) which employ CLIP text encoder.[2]

Foundational Frameworks in Cultural Studies Rokeach (1967) introduced the Rokeach Value Survey, illuminating how classified values shape behaviors at both individual and societal levels. Building on this, Hofstede (1983) laid grounding work in studying cultural variances by introducing key cultural dimensions like femininity versus masculinity, fostering a systematic approach to exploring cultural differences. Bond (1988) identified 36 ‘‘universal values’’, such as love and freedom, underscoring the shared human values across diverse cultures and regions. Schwartz (1994) further refined our understanding by proposing a universal framework of ten fundamental human values, revealing how they are prioritized and interpreted diversely across cultures, impacting behaviors and attitudes. Later advancements by Triandis and Gelfand (1998) and McCrae and Allik (2002) further enriched the domain of cross-cultural psychology. Triandis emphasized the nuances of individualism and collectivism, while McCrae delved into the variances in the manifestations of the Big Five personality traits across different cultural settings. Complementing these, the World Values Survey (Haerpfer et al., 2022) introduced an innovative cultural map, portraying the global shift towards more secular and self-expression values as societies advance and prosper.

This work synthesizes the cultural literature presented above to establish a comprehensive repository of cultural concepts and dimensions. In §4, we introduce our culture characterization for

[2] For some models (e.g., DALL-E [Ramesh et al., 2022]) the nature of the multilingual encoder is unknown.

detailed analyses of the representation and perception of cultural perspectives within TTI models.

Culture in LMs and TTI Models With the burgeoning interest in Pre-trained Language Models (PLMs) and TTI diffusion models, there is increasing scrutiny of the cultural gaps and biases inherent within these models. These gaps manifest as discrepancies in the representation of norms, values, beliefs, and practices across diverse cultures (Rao et al., 2024; Prabhakaran et al., 2022; Struppek et al., 2022; Abid et al., 2021; Ahn and Oh, 2021; Touileb et al., 2022; Smith et al., 2022). Arora et al. (2022) and Ramezani and Xu (2023) explored the cross-cultural values and moral norms in PLMs and assessed their alignment with theoretical frameworks, revealing a conspicuous inclination towards western norms. Similarly, other studies demonstrated the challenges with English probes and monolingual LMs, which diminish the representation of non-Western (e.g., Arab) norms in model responses (Cao et al., 2023; Naous et al., 2023; Masoud et al., 2023; Atari et al., 2023; Putri et al., 2024). While these works focused on the cultural implications of PLMs, detecting or mitigating cultural biases, we develop a methodology that inspects the TTI models' cultural values, and seek to understand their internal representations of cultures.

While cultural exploration in TTI model research is relatively limited, there have been advancements to enhance cultural diversity within these models. Multilingual benchmarks, focusing on Chinese and Western European/American cultures, have been introduced (Liu et al., 2023, 2021). Efforts to uncover cultural biases include evaluations of nationality-based stereotypes (Jha et al., 2024), skin tone biases (Cho et al., 2022), and associated risks (Bird et al., 2023). Analyses also address social biases in English-only TTI models (Naik and Nushi, 2023), covering gender, race, age, and geography. Kannen et al. (2024) and Basu et al. (2023) have evaluated the cultural competence of these models. Our work extends beyond Western tendencies, focusing on multilingual TTI models and exploring fine-grained cultural concepts and dimensions.

3 Research Questions and Overview

Our primary research inquiry revolves around: **How does multilingual TTI models capture cultural differences?** To delve into this overarching question, we craft four research questions:

- *RQ1:* ***Do*** *TTI models encode cultural knowledge?*
- *RQ2:* ***What*** *are the cultural dimensions encoded in TTI models?*
- *RQ3:* ***Which*** *cultures are more similar according to the model?*
- *RQ4:* ***How*** *to unlock the cultural knowledge?*

RQs 1-3 form a hierarchy, with each question building upon the previous one. RQ4 stands as an independent, high-level question. Our methodology consists of three pillars: (1) crafting a cultural ontology; (2) experimenting with TTI models featuring diverse multilingual text encoders, and (3) employing a triad of evaluation methodologies: *intrinsic evaluation* using OpenClip, *extrinsic evaluation* involving VQA models, and *human assessment*.

4 Cultural Ontology

We aim to design an ontology that will allow us to (1) consolidate diverse perspectives on culture; and (2) quantitatively assess cultural aspects within the context of TTI models. Research on cultural definitions is typically based on breaking down the big idea of culture into different aspects like individualism or science, which often involve intricate and abstract details or queries unsuitable for visual examination. To utilize culture studies to our needs, we hence develop two key pillars: (a) ***cultural domains***, comprising ***cultural concepts***; and (b) ***cultural dimensions***.

Cultural Domains and Concepts. Drawing inspiration from established categorizations in works like Hofstede (1983), Rokeach (1967), and Haerpfer et al. (2022), we combine ten common and broad aspects to form the cultural domains. Each domain reflects a collection of values, tendencies, and beliefs, which we represent through concise concepts. For instance, the cultural concept of *Heaven* in the *Religion* domain is derived from the question in the religion section of the World Values Survey, which asks: ‘‘Do you believe in heaven?’’. We define twelve domains and 200 cultural concepts (see Table 9 in the Appendix). These domains include *Moral Discipline*

Prompt Description	Prompt Template (T - translated, EN - English)	Example
English Reference	EN: ‘‘a photo of <concept>’’	‘‘a photo of food’’
Fully Translated Prompt	T: ‘‘a photo of <concept>’’	‘‘фото еда’’
Translated Concept	EN: ‘‘a photo of’’ + T: <concept>	‘‘a photo of еда’’
English with Nation	EN: ‘‘a photo of <nationality> <concept>’’	‘‘a photo of Russian food’’
English with Gibberish	EN: ‘‘a photo of <concept>’’ + T: ‘‘<gibberish>’’	‘‘a photo of food йкуаскымдо’’

Table 1: Prompt Templates: Language and Gibberish prompts.

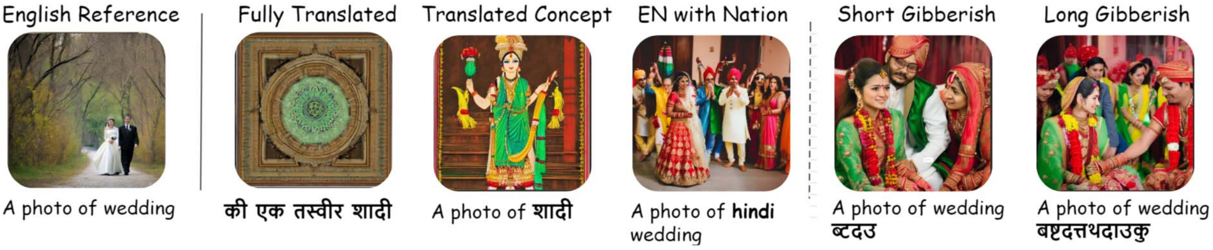


Figure 2: StableDiffusion images generated from all the prompt templates for the cultural concept (CC) of *Wedding* and the Hindi language.

and Social Values (example concepts: Housewife, Divorce), *Education* (Teacher, Engineer), *Economy* (Market, Job), *Religion* (God, Wedding), *Health* (Doctor, Medicine), *Security* (War, Weapon), *Aesthetics* (Art, Fashion), *Material Culture* (Car, Camera), *Personality Characteristics and Emotions* (Lazy person, Proud person), and *Social Capital and Organizational Membership* (City, Police).

Cultural Dimensions. Certain cultural aspects function more like axes (e.g., from Individualism to Collectivism) than as comprehensive domains (e.g., Science). We grouped these aspects under the category of cultural dimensions. The dimensions we use are defined as follows: (1) *Traditional versus Rational* values;[3] (2) *Survival versus Self-expression* values (Haerpfer et al., 2022); (3) *Critical versus Kindness*;[4] (4) *Extraversion versus Introversion* (McCrae and Allik, 2002); (5) *Modern versus Ancient* values; (6) *Masculine versus Feminine* attributes; (7) *Individualism versus Collectivism* and (8) *Nature versus Human* (Hofstede, 1983; Schwartz, 1994).[5] The cultural dimensions don’t cover all suggested aspects from the original research. We focused on aspects that are more visually representable and quantifiable.

5 Unlocking Culture in TTI Models

We now introduce our prompt templates which feed the cultural concepts as input into the TTI models, aiming to unlock the effect of different cultures on their outputs.

Cultural Concepts and Dimensions. In §4 we defined cultural concepts, denoted below with $\{CC\}_{i=1}^{200}$, and cultural dimensions, denoted as $\{CDM\}_{i=1}^{8}$. Cultural concepts are dynamic parts of the TTI model templated input (see Table 1). Every *CC* is expressed by one or two English words (e.g., *Food*), acting as a concise representation of a more expansive domain (e.g., *Aesthetics*). In contrast, the cultural dimensions are used in our outcome measures but not in the prompts.

Prompt Templates. We construct five prompt templates (PTs; see Table 1 and resulting images in Figure 2). These templates aim to discern the cultural implications carried solely by the linguistic characters of the language in question. In our setup, a *PT* is defined as a function ϕ of the target language, *L*, and a cultural concept, *CC*: $PT = \phi(L, CC)$. The first two PTs (*Translated PTs*) with the third (all together - *Language PTs*) enable us to investigate whether the language can convey cultural information, while the last PT (*Gibberish PT*) aims to explore if linguistic characters alone can convey such information,

[3] The original term is *Secular-Rational*.

[4] McCrae and Allik (2002) originally named this dimension by *Neuroticism versus Adjustment*.

[5] Schwartz (1994) originally included this in the values *Universalism* and *Harmony*.

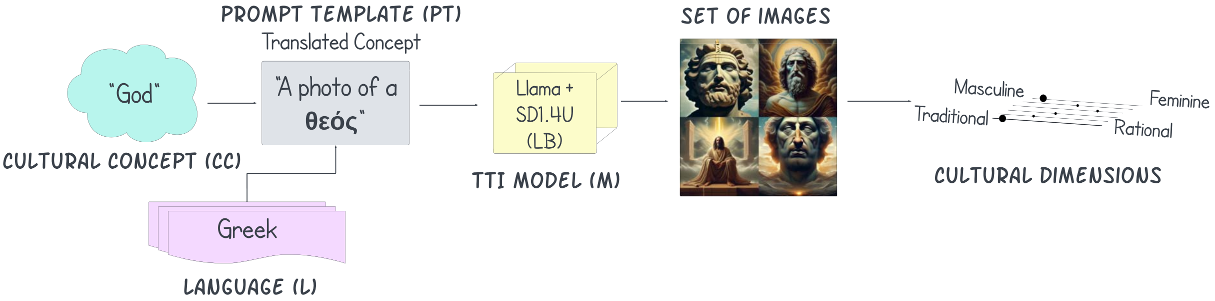


Figure 3: TTI model workflow scheme. The visual representations of each Cultural Concept (CC) are image sets generated with different languages (L) and prompt templates (PTs) by different TTI models (M). Then, the images' cultural content is evaluated. Here, for example, CC is *God*, PT is *Translated Concept*, M is Llama2 + SD1.4 UNet (LB) and the evaluation uses the cultural dimensions metrics (§6.1).

Metric	X**: Template**	X**: Instance**	D
	Identifying Cultural Origin [RQ1, RQ4]		
National Association (NA)	*a photo with <national> style*	*a photo with spanish style*	softmax $\left(\vec{\text{D}}\right)$
Extrinsic NA (XNA)	*What is the country of origin for the depicted photo?*	–	VQA, Majority
	Depicting Cultural Dimensions [RQ2]		
Cultural Dimensions Projection (DP)	*a photo with <cultural dimension> aspects*	*a photo with modernity aspects*	$\left(\text{I} \cdot X^t\right)$
Extrinsic DP (XDP (d_0, d_1))	*Are there more <d_0> features in the photo or more <d_1>?*	*Are there more modern features in the photo or more ancient?*	VQA, Majority
	Finding Cultural Similarities [RQ3]		
Cultural Distance (CD)	EN*:a photo of <cultural concept>*	*a photo of city*	$1 - \left(\text{I} \cdot X^t\right)$
Cross-Cultural Similarity (CCS ($l1$, $l2$))	*a photo of* T_{l2}*:<cultural concept>*	*a photo of ciudad*	$I_{l1} \cdot (X^v)_{l2}$

Table 2: Automatic Metrics - Grouped by their aims and research questions. I corresponds to the OpenClip visual representation of an image in the set generated by a TTI model. X^t and X^v stand for the textual and visual representations of the baseline prompt (X:Instance), respectively. $l1$ encodes the language of the inspected I while $l2$ encodes the other language of the evaluation prompt X. $\vec{\text{D}} = \left[I \cdot X_1^t, I \cdot X_2^t, \ldots, I \cdot X_k^t\right]$, for the k nationalities. Finally, T stands for 'Translated', and d_0 and d_1 stand for the dimension extremes, respectively.

especially when the language lacks extensive translated data in the TTI models' training data. Interestingly, in our experiments we found the more non-English information (words, characters) the prompt contains, the lower the conceptual coverage (see Figure 12 and details in Appendices B.1.2 and B.2.4).

TTI Model Workflow. A TTI model, *M*, operates by receiving a textual prompt as input, denoted as *In*, and in turn, generating a corresponding set of images as output. We form *In* through prompt templates, *PTs* (Table 1), to study how changing parts of *In* affects the image generated by the model (*M*). By employing a *PT*, we are able to keep all elements of the input constant except the one under examination. As depicted in Figure 3, the *PT* shapes the input (*In*) to *M*, culminating in a generated image that reflects the interplay between cultural concepts within the model's parameters (see examples of generated images in Figures 20, 21 in the Appendix).

6 Evaluating Cultural Aspects in Images

In this section, we discuss how we evaluate the images' cultural content. We employ two **automatic** measures for cultural characteristics (**intrinsic** and **extrinsic**), as well as **human assessment**.

6.1 Automatic Metrics

We introduce six metrics, corresponding to our research questions (Table 2 and §3), which fall into two categories: *intrinsic*, utilizing internal representations, and *extrinsic*, relying on an external VQA model. Consistent with previous culture-in-TTI research (Wang et al., 2023; Naik and Nushi, 2023; Liu et al., 2023), we construct intrinsic measures using both textual and visual representations from image-text encoders (OpenClip in our case). The intrinsic metrics follow the equation: $\mathbb{D}\left(I \cdot X\right)$.

Here, I is a visual representation of an image generated by a TTI model in response to a prompt of interest. X is either a textual representation (X^t) or a visual representation (X^v) of the metric's baseline prompt (the X**: Instance** column of the table), and the choice of X being textual or visual representation is metric specific. D is the metric operator, that is applied to the cosine similarity scores between I and X. The eventual metric is an average of the D values over the n images in the set generated by the TTI model in response to a given prompt (see §5). This section will exemplify our metrics with a running example of I curated from the prompt: ''a photo of <stadt>'' (*stadt* is the German word for *city*).

Identifying Cultural Origin [RQ1, RQ4] The first metric, ***National Association (NA)***, aims to identify the origin culture of the image. Intrinsically (top table row), we calculate the cosine similarity scores between the generated image representation (I) and the textual representations of the nationality templates (X; e.g., a national prompt for the ''Spanish'' style). We do so with all the nationalities[6] in our data and apply the softmax operator on the scores vector. If the image does correspond to the German culture, we would expect the German coordinate in the vector to be high. Extrinsically (***XNA***, second row), we direct a query to the VQA model, inquiring about the origin country of the image I. Subsequently, we compute the majority vote over the set of generated images. If there is no clear majority, the XNA answer is ''can't tell''. The score is the fraction of times a question on the image is answered correctly by the majority vote.

Depicting Cultural Dimensions [RQ2]: By the ***Cultural Dimensions Projection metrics (DP, XDP)***, we evaluate the extent to which cultural dimensions are manifest within the images. As can be seen in rows 3 and 4 of the table, the computation is very similar to the above nationality measures (NA and XNA).

Finding Cultural Similarities [RQ3]: Inspired by Hofstede's (2010) definition of culture,[7] we introduce the ***Cultural Distance (CD)*** and ***Cross-Cultural Similarity (CCS)*** metrics to probe cultural distinctions. In CD, we assess how closely various cultures align with the English culture.[8] In CCS we measure image similarities to explore how different languages influence the visual representation of the same cultural concept (cc).

6.2 Human Evaluation

We create a questionnaire (see questionnaire example and guidelines in Figures 18, 19 in Appendix B.3) for human evaluation of cultural dimensions in images generated by TTI models. The questionnaire considers 4 languages (RU, ZH, ES, DE), 12 concepts, 3 prompt templates (English with Nation, Translated Prompt, and English with Gibberish) and 3 models (2 with implicit multilingual encoding - SD and DL; and 1 with explicit multilingual encoding - AD), involving 15 annotators (3 per item) on the LabelStudio (Tkachenko et al., 2020–2022) platform. For each (TTI model, prompt template, concept) triplet we generated 4 images per language, for a total of 1728 images in the entire evaluation set. Each triplet is represented by 1 page in the questionnaire, consisting of four 4-image grids,[9] 1 grid per language. For each 4-image grid, annotators were asked to make 3 binary decisions: one for each of 3 arbitrary dimensions (*Modern versus Ancient*, *Traditional versus Rational*, *Critical versus Kindness* in §4), and specify the culture of origin from a given set of options (the 4 languages as well as USA). We calculate the inter-annotator agreement with the Fleiss kappa (*Modern versus Ancient*: 0.54, *Traditional versus Rational*: 0.39, *Critical versus Kindness*: 0.41, national association: 0.4; on a $[-1, 1]$ scale) and the agreement of human annotation (after taking the majority vote) with the ground-truth culture (74.4%). Below we report the agreement of the automatic evaluation metrics with the majority vote between the annotators for each example.

7 Experimental Setup

Languages. We experiment with ten languages, serving as proxies of geographically diverse cultures: English (EN), Spanish (ES), German (DE), Russian (RU), French (FR), Greek (EL), Hebrew (IW), Arabic (AR), Chinese (ZH), and Hindi (HI).

[6] see Table 8 in the Appendix.

[7] ''Collective mental programming distinguishing one group from another''.

[8] Using the EN reference in the CD measure aligns with prior studies which employed EN as a reference for western cultures (Atari et al., 2023), and acknowledges the English predominance in the training data of our TTI models.

[9] Figure 3 presents an example of a 4-image grid.

	EN	ES	DE	FR	RU	EL	AR	IW	ZH	HI
StableDiffusion 2.1v	v	v	v	v	v	v	v	v	v	v
StableDiffusion 1.4v	v	v	v	v	v	v	v	v	v	v
Llama2 + SD1.4 Unet	v	v	v	v	v	v	v	v	v	v
AltDiffusion m9	v	v	v	v	v	v	v	v	v	v
DeepFloyd v1.0	v	v	v	v	v					
DALL-E v2	v	v	v	v	v				v	

	Text Encoder	Objective
StableDiffusion 2.1v	OpenCLIP (ViT-H/14)	I
StableDiffusion 1.4v	CLIP (ViT-L/14)	I
Llama2 + SD1.4 UNET	Llama2 7b	I
AltDiffusion m9	XLM-R (in AltClip)	E
DeepFloyd v1.0	T5-XXL	E
DALL-E v2	Unknown	Unknown

Table 3: Top: Model coverage of different languages. Bottom: Model's text encoder. The coverage is based on the existence of multilingual characters (letters) in the embedding layers of the text encoder of each model (except for DALL-E, where it is based on empirical tests). Multilingual capabilities are acquired through an *explicit* (E) or *implicit* (I) training objective.

We consider two inclusion criteria: (1) The lingual coverage of TTI models (Saxon and Wang, 2023), and (2) Etymological Diversity. Balancing between both, we cover mainly the Indo-European language family.

Models. We experiment with six SOTA TTI models, namely, StableDiffusion 2.1v (SD), StableDiffusion 1.4v (SD1.4), AltDiffusion (AD), DeepFloyd (DF), DALL-E (DL), and Llama 2 + SD 1.4 UNet based on Llavi-Bridge (LB), differing in their multilingual textual encoders and the languages they cover (Table 3; Appendix Table 7). The multilingual capabilities of a model are affected by the languages it is trained on, and the training objective it follows. For encoders like XLM-R and T5-XXL, the multilingual aspect is *explicitly* represented in the training objective, by bringing similar words in different languages closer in the learned embedding space. Also, multilingual aspects can be *implicitly* represented, with different alphabets encoded differently while the objective does not impose any explicit cross-lingual constraint (e.g., as in SD). For the evaluation (§6.1) which requires a VQA model, we employ BLIP2 (Li et al., 2023), which applies the Flan-T5-XL encoder.

Experimental Dataset. We employ the 6 models to generate an image set, where each image is characterized by 4 properties: (1) the generating TTI model (M); (2) the cultural concept (CC) of interest; (3) the applied prompt template (PT); and (4) the target culture (L), encoded through the prompt, either by its language or through the culture name it mentions. We generate a K-image set, for the value of $K = 4$, for each configuration of these properties, maintaining a constant initiation seed (42) for the first image in each set. This methodology yields for each TTI model T unique cultural tuples of the form (CC, PT, L, 4 images), where the number of tuples depends on the number of languages covered by the model, see Table 3 ($T_{SD} = T_{AD} = 10,500$, $T_{SD1.4} = T_{LB} = T_{DF} = 6,300$ and $T_{DL} = 2,310$).[10,11]

8 Experiments and Results

1. TTI Models Encode Cultural Identity Information (RQ1). Figure 4 illustrates the extrinsic national association (XNA) scores measured on images generated by the experimental models. It is important to note that this metric uses free text answers from the VQA, which can vary widely based on national origins, making it difficult to achieve high scores. Despite that, 2 languages (HI and RU) score consistently above 0.4, with the highest mean scores across models (0.69, 0.73); 3 languages (FR, DE, AR) score consistently above 0.3 in 5 out of 6 models; and only 2 languages with a mean score lower than 0.3 (ES and EN). See detailed results in Figure 4 in the Appendix. We hypothesize that low English Association scores are due to the overrepresentation of English in the training data, resulting in a lack of cultural specificity. In contrast, the more limited training data for other languages is likely more culturally specific, as it is likely to be carefully selected. Additionally, it can be associated with the global influence of American culture in the data, which may obscure distinct cultural traits, further impacting model performance in English. Finally, the results ascertain that all the examined models can distinguish image origins.[12]

[10] Due to API usage constraints, the DALL-E subset was limited to half of the cultural concepts (105) and three prompt templates (''English with nation'', ''Translated concept,'' and ''English with gibberish'', see Table 1). SD 1.4v and LB are limited to these 3 PTs as well.

[11] The images in the human evaluation set of §6.2 are selected from this dataset.

[12] The automatic XNA metric agrees with the human answers to the cultural origin question in 69.6% of the cases in the human evaluation set.

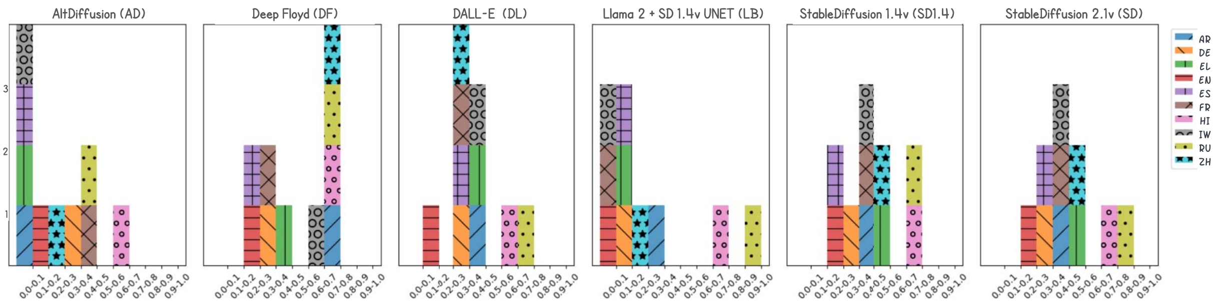


Figure 4: National Association Scores by BLIP2 (XNA) presented as histograms. The x-axis represents bins of mean XNA scores ranging from 0 to 1 across three representative Prompt Templates (PTs): ‘Translated Concept’, ‘EN with Nation’, and ‘English with Gibberish’ (refer to Table 1 for details). Higher scores indicate better performance. Colors encode languages.

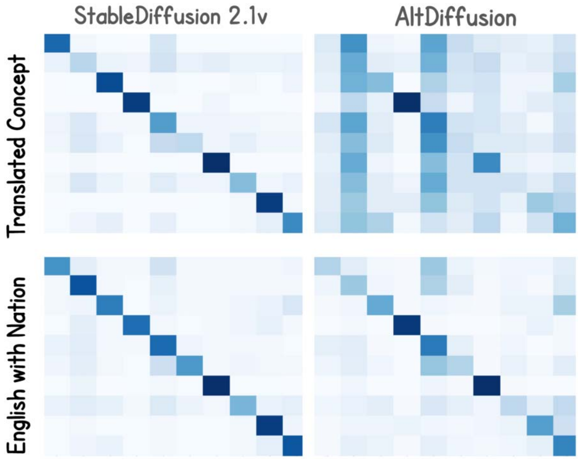


Figure 5: A confusion matrix grid of the NA metric. Prompt Templates[13]: ‘‘Translated Concept’’ (top), ‘‘EN with Nation’’ (bottom). Models: SD (left) and AD (right). Darker colors correspond to higher scores. y-axis: ground-truth languages. x-axis: predicted cultures. For each confusion matrix, we compute the agreement between the predicted and the ground-truth languages (Accuracy, $ACC = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(\text{argmax}(\text{row}_i) = i)$) to measure the cultural encoding strength of a (model, PT) pair. Languages in each grid (top-bottom, left-right): RU, EN, EL, HI, DE, FR, ZH, ES, AR, IW.

2. Cultural Encoding Depends On The Language-Encoding Strategy Of The Prompt And The Model (RQ4). Based on the intrinsic NA results presented in Figure 5, we observe that **models with implicit multilingual encoding (SD) are better cultural encoders than models with explicit multilingual encoding (AD)**. The implied explanation is that explicit encoders bring languages closer together in the embedding space.

Note that the main diagonal is emphasized in both the left column and the bottom row of the heatmap. This indicates that **an effective prompt (‘‘EN with Nation’’) can compensate for the effect of the encoder**. This is reflected by ACC values of 0.8 and 0.5 for explicit encoding compared to 1.0 and 1.0 for implicit encoding, for the translated prompt and the English with Nation prompt, respectively. These patterns resonate with those observed in the other models tested.

Interestingly, AD with the translated prompt is biased towards the American and German cultures (100% of the erroneously predicted cultures are classified as American or German), while the errors of AD with English with Nation are more evenly distributed.[14]

Notably, since language acts as a proxy for multiple nations (e.g., English is spoken in both the USA and the UK, two different cultures), we provide a second-order analysis (Figure 11 in the Appendix) representing the national association distribution with other nations that primarily speak these languages. This analysis implies inherent biases within the encoding, such as Greek images being more associated with Cyprus than Albania.

3. TTI Models Encode Cultural Dimensions (RQ2). Here we show how TTI models capture cultural dimensions outlined in our ontology. Given the challenges in defining an absolute ground-truth for cultural tendencies, we proceed with caution. We avoid direct comparisons with any such ground-truth, mindful of the potential harm such analyses could incur. Instead, in §9, we carefully examine the correlations between our

[13] The ‘‘Fully translated PT’’ is omitted after initial consistency validation.

[14] The automatic NA metric agrees with the human answers to the cultural origin question in 75.0% of the cases in the human evaluation set.

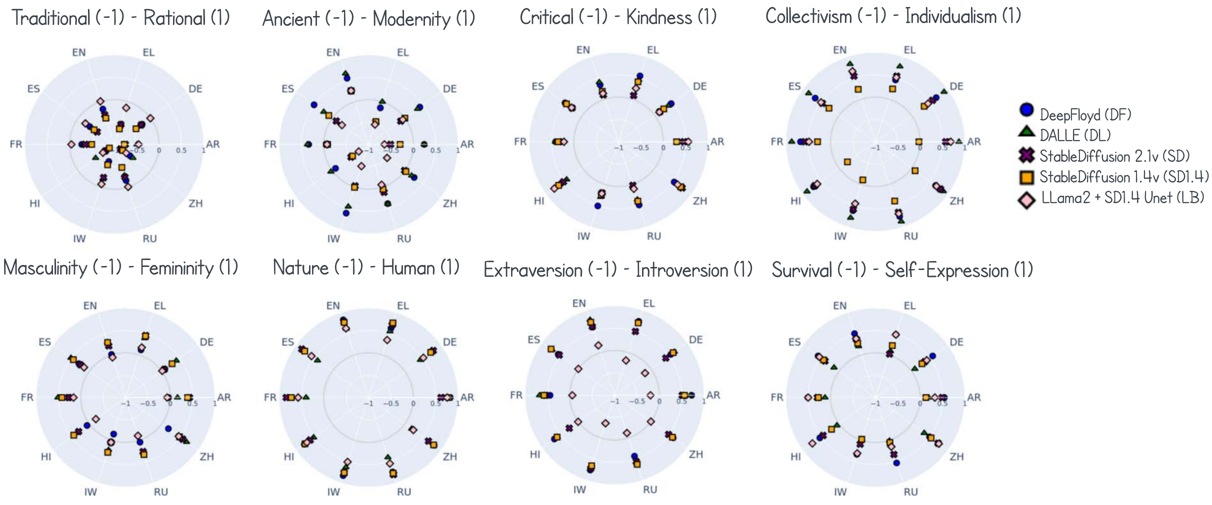


Figure 6: Radar graphs of cultural dimensions as classified by the VQA model. The scores, range [−1, 1], represent the tendency of each culture towards one of the ends of the dimension. For example, 0.6 of the Arabic images were classified as *modern* and 0.3 as *ancient* (0.1 as *can't tell*) and hence the modernity score is 0.3. Circles encode cultural dimensions, markers represent models. Languages appear at different angles on the perimeter. For each dimension, the negative end is at the center of the circle, while the positive end is on the perimeter. For each model, results are averaged across PTs.

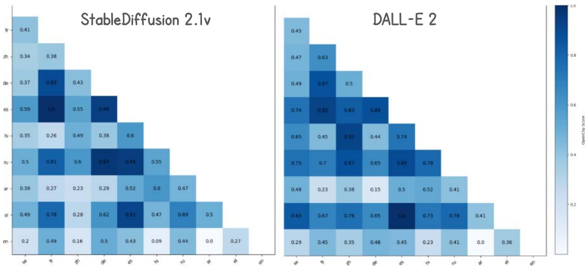


Figure 7: Cross-Cultural Similarity (CCS) analysis for the 'EN with Nation' PT. Darker values note higher similarity. The scores are normalized.

findings on TTI models and social science studies related to the cultures discussed.

The cultural dimension results, presented in Figure 6 (see detailed results in Table 6 in the Appendix), are depicted in radar graphs, each linked to a specific cultural dimension. Utilizing the XDP metric, we analyze the classification of the images generated for each language into the dimensions.

Similarly to the above conclusions, explicit multilingual encoding (the DF model) is less representative of cultural differences, as indicated by the similar dimensional scores it typically assigns to different languages. We hence continue this analysis with SD 2.1v (implicit multilingual encoding) and DL (unknown multilingual encoding).

For the Traditional versus Rational axis, languages such as German, English, Russian, and Hebrew exhibit a predilection for rational aspects over traditional ones. In contrast, Hindi, Chinese, and Arabic lean towards traditional elements, which is also echoed in the other models. These findings echo the Agreeableness axis, in terms of Critical versus Kindness. English, Russian, Hebrew, and German tend to emphasize critical characteristics, whereas Hindi, Chinese, and Arabic exhibit a more pronounced kindness dimension. Likewise, for modernity, Hindi, Arabic, and Greek tend to embody more ancient attributes, while Russian and English images are more modern. In contrast, for some cultural dimensions, for example extroversion-introversion, the models do not reveal significant cross-cultural differences.[15]

4. TTI Models Encode Cultural Differences & Similarities (RQ3). We start with the Cross-Cultural Similarity (CCS) metric analysis (Figure 7; Figure 14 in the Appendix): Similarities

[15] The automatic XDP metric agrees with the human annotators in 74.8% of the images for the modern-ancient dimension, 69.9% for the traditional-rational dimension and 60.6% for the critical-kindness dimension. Notice that the other dimensions are not annotated in the human evaluation set.

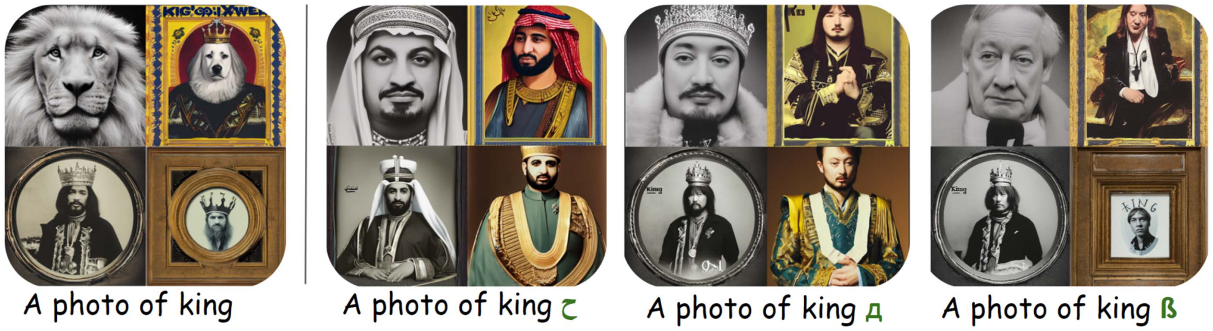


Figure 8: SD Images generated by one letter addition to the prompt ‘a photo of a king’. Left to right: Arabic, Russian, and German letters.

among cultures, computed as the similarities between the images generated by each model for these cultures, when using the ‘‘English with Nation’’ PT. It reveals the extent to which cultural attributes and characteristics are shared across different cultures, as perceived by the different models. Interestingly, all models consistently show the highest similarity scores among German, French, and Spanish. In etymological terms, the images generated by all models demonstrate discernible resemblances between European languages, particularly Romance, while demonstrating distinct disparities when compared to languages with origins in the Indian or Tibetan language families.

We next present the Cultural Distance metric (see §6.1; Figure 13 in the Appendix), measured on the output images from all examined TTI models, indicating the alignment of various cultures with the English culture. Our findings reveal that TTI models encode cultural similarities differently. Translated prompts (‘‘Fully translated’’ and ‘‘Translated concept’’) show the highest cultural distance from the English reference. Particularly, we notice scores higher than the averaged score of the Language PTs for SD as also observed in SD1.4 and LB in Greek (76.16), Arabic (75.75), and Hindi (75.14), for AD in Hebrew (74.57), for DF in German (72.1), and for DL in Chinese (71.4). These findings highlight how different TTI models perceive these cultures differently from the English culture.

5. Alphabet Characters Can Unlock Cultural Features (RQ4). Our empirical results so far suggest that cultural properties can be unlocked through the use of terms from the corresponding language in the prompt. Figure 8 demonstrates that including a single character from the target language in the prompt also results in images with properties of the target culture. We next look more deeply into this phenomenon, asking whether arbitrary strings of letters (Gibberish) in the prompt can serve to unlocking the cultural knowledge in TTI models. Our approach is to optimize the Gibberish sequence so that the generated image represents as much cultural information as possible.

To this end, we adjust a gradient-based discrete prompt optimization method, PEZ (Wen et al., 2023), to suit our requirements: We set the number of *target letters* in the Gibberish term, T, to one of the values in $[1, 2, 3, 5, 10]$, and optimize the term. We initiate the prompt as: ‘‘a photo of <Cultural Concept (CC)> T’’, and for each culture we only utilize the alphabet of its language. The algorithm’s objective is defined as the negative cosine similarity between the embeddings of the resulting prompt (including the inferred letters) and the objective features. We consider two options for objective features: (1) *Textual Objective Features*, where we target the intrinsic features of the NA template (§6.1), i.e., an OpenClip representation of a prompt that adheres to the following template: ‘‘a photo with <national> style’’; and (2) *Visual Objective Features*, the mean image features of a Google search extracted 4-image set, using the ‘‘English with Nation’’ PT as the image query (for the specific CC).

We consider only the SD TTI model, as its implicit multilingual encoding has shown most successful in cultural encoding throughout our above experiments. We run this algorithm for 6 concepts (Appendix B.2.3), 10 languages, and 5 Gibberish string lengths (see above), and hence learn 600 new prompts (300 for each training objective).

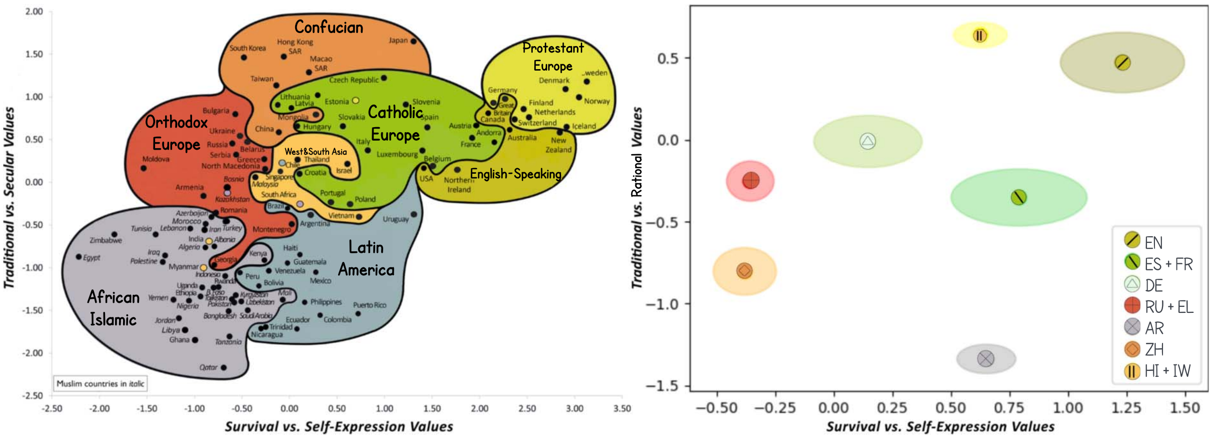


Figure 9: Our cultural dimensions (DP) space (right) inspired by the World Culture Map 2023 by Inglehart-Wazel, i.e., the spread of geographical-cultural values (left). Two dimensions: *Tradition versus Rational* (y-axis) and *Self-expression versus Survival* (x-axis). The axes are the subtraction of the mean scores of the two poles of each dimension. Scores are grouped by region-related languages (as on the left), with std values defining the clusters. Results with 'Fully Translated' PT using SD 2.1.

Images generated from optimized prompts with a textual objective achieve equal accuracy as our best PT, English with Nation, in terms of the NA metric in Chinese (100%), Russian (50%), Arabic (83%), Hebrew (67%), and Hindi (100%). Additionally, they yield equal or superior results compared to Translated PTs in NA and XNA, respectively, in English and French with up to 16% improvement. However, there is no improvement in Spanish and Greek over random Gibberish PTs or the rest of PTs.[16] Notably, the sequence length parameter does not exhibit a clear trend neither for NA nor for XNA. This discovery sparks curiosity about the internal representations of TTI models and, in turn, beckons future research to explore their cultural components.

9 Ablation Analysis

To support our key findings, we analyze potential factors behind the results. We validate cultural dimensions with ground-truth data, propose the characters embedding to explain cultural distance and cross-cultural similarity scores, and test the National Association task with different input data and VQA. Additionally, we conduct a qualitative analysis to reveal underlying representations in the generated images.

Comparing Cultural Dimensions Space to Ground Truth. To examine our findings in relation to social science studies we draw inspiration (Figure 9) from Ingelhart and Wazel's visualization of the contemporary world cultural map (WCM) as a ground-truth.[17] We qualitatively compare them side-by-side with axes representing the subtraction of two poles of dimension, derived from DP scores (see Section 6.1) and grouped by language origins as in WCM. Spanish and French merge to symbolize the Catholic European cultural sphere, while Russian and Greek represent the Orthodox European context. Hindi and Hebrew cluster together, signifying the cultural landscape of Western and South Asian regions.

Despite variations and our distinct scoring methodology, our findings significantly correlate with the original map, indicating that the axes we've identified indeed carry cultural meaning.

Investigating Cultural Distances In The Textual Embedding Space. We propose that the cultural inclination towards European languages (see finding 4 in Section 8) resonates with the encoder's ability to map letters from different languages into distinct and well-defined clusters in the embedding space. This ability likely stems from both the language frequency in the TTI model's pretraining data and the encoder's training objective. Visualizing the characters in the embedding space of CLIP reveals two key findings (see Figure 16 in the Appendix). First,

[16] In the overall analysis, the visual objective shows similar NA metric trends.

[17] https://www.worldvaluessurvey.org/wvs.jsp.

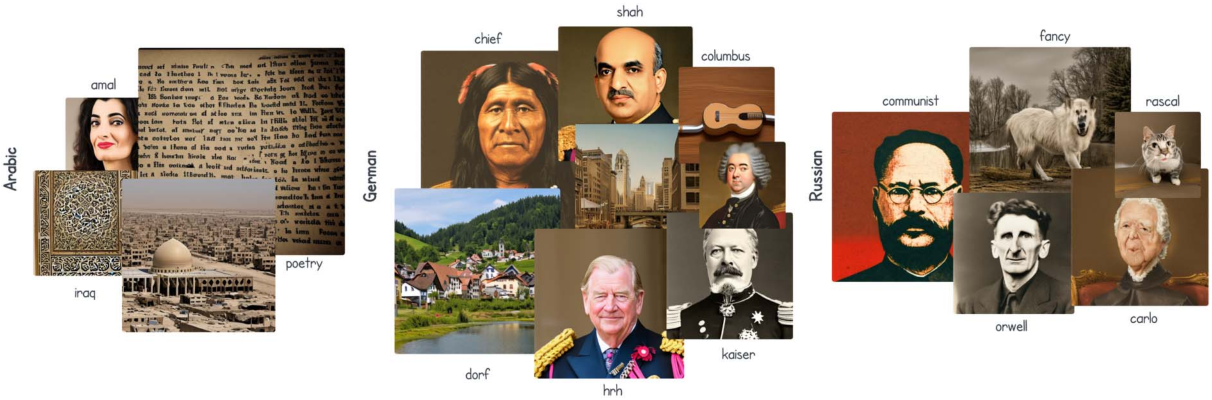


Figure 10: Decomposition of the concept *King* to its unique cultural tokens. Examples with the prompt template Translated Concept in the languages (Left to right): Arabic, Russian, and German.

characters are either mapped to the same embedding (cross-lingual cluster) or have separate embeddings (language-specific clusters). Second, these clusters span linear distances that correlate with CD and CCS scores. Language clusters for ZH, HI, AR, and EL are closer to the cross-lingual cluster, while Latin-specific clusters like FR and ES are closer to the English cluster. Thus, embeddings alone can provide valuable insights into cultural perception, offering a promising direction for future research.

The Expected Performance of National Association. As far as we know, the task of detecting the national origin of an image has not been previously addressed. Thus, to assess the expected performance of *synthetic images* (i.e., TTI-generated images) we evaluate the NA performance on *natural images*. We compare the XNA scores of images generated by all examined TTI models and their corresponding images extracted from the Google Photos search engine (See Figure 15 in the Appendix).[18] The ability to detect the origin of *Natural Images* increases by 25% on average across languages and models, serving as an upper bound for TTI images. This is expected because, first, natural images likely better represent cultures than TTI images, which may contain errors. This shows there is a room to improve TTI models to better encode culture in images. Second, VQA models trained primarily on natural images may perform better within this domain.

Revalidating National Associsation Findings With GPT-4-Vision. To ensure that the high performance of the task is independent of the VQA model (BLIP), we replicate the XNA experiment with GPT-4 Vision on the human annotation subset (due to API usage constraints; see Table 5 in the Appendix). The XNA metric using GPT-Vision brings similar trends with better agreement with the ground truth, with a mean score of 69.36% (over the same 3 PTs and 4 languages), which is higher than the parallel BLIPs performance (45.13%). Notably, GPT-Vision aligns with human answers in 70% of the cases.

Revealing Hidden Cultural Representations of Generated Images. Understanding the factors influencing model generations is challenging, especially from a cultural perspective. To uncover these factors, we conduct a qualitative analysis employing the Conceptor method (Chefer et al., 2023), a recent technique that explains generated images of interest concepts by decomposing them into sets of tokens whose linear combination reconstructs the image. These tokens reflect hidden representations of images generated by translated concepts. For a set of 30 images (3 concepts, 10 languages), we compute the 50 most significant tokens (with the highest weights), manually filter the unique tokens, and generate their images for clearer visualizations. We take the concept *king* as our running example (Figure 10; see additional examples in Figure 17 in the Appendix). In Arabic, the main unique tokens are *poetry, Arabic,*

[18] We experiment with 240 images spanning 6 concepts across our 10 languages while focusing on the English with Nation PT, both in the manual Google search queries (e.g., *Spanish king*) and the images' generation.

Iraq, and *amal*, indicating an emphasis on art. In German, tokens like *chief, shah, kaiser, tenor*, and *dorf* suggest a stronger tendency towards hierarchy and music. In Russian, tokens such as *communist* and *Orwell*[19] imply a more political influence. These inherent tendencies, whether grounded in cultural history or not, ultimately affect the generated images, leaving the door open for future research.

10 Discussion and Future Work

We studied cultural encoding in TTI models, a research problem that, to our knowledge, has not been addressed before. We mapped the quite abstract notion of culture into an ontology of concepts and dimensions, derived prompt templates that can unlock cultural knowledge in TTI models, and developed evaluation measures to evaluate the quality of the cultural content of the resulting images. By doing so we were able to answer *Do*, *What*, *Which*, and *How* research questions about the nature of cultural encoding in TTI models, and to highlight a number of future research directions.

Our study has limitations. We focused on a finite set of cultural concepts and prompt templates. Additionally, our automatic evaluation may struggle with abstract cultural concepts and translation challenges. Nevertheless, we hope this paper will encourage our fellow researchers to further investigate the intersection between culture, multilingual text encoders and TTI models.

Acknowledgments

We would like to thank the action editor and the reviewers, as well as the members of the DDS@Technion NLP group for their valuable feedback and advice. This research was partially funded by VATAT Women in High-Tech Profession Program.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463. https://doi.org/10.1038/s42256-021-00359-2

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *EMNLP (1)*, pages 533–549. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.42

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*. https://doi.org/10.18653/v1/2023.c3nlp-1.12

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, and Joseph Henrich. 2023. Which humans? *PsyArxiv*. https://psyarxiv.com/5b26t.

Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5136–5147. https://doi.org/10.1109/ICCV51070.2023.00474

Charlotte Bird, Eddie L. Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. *arXiv preprint arXiv:2307.05543*.

Michael H. Bond. 1988. Finding universal dimensions of individual variation in multicultural studies of values: The Rokeach and Chinese value surveys. *Journal of Personality and Social Psychology*, 55(6):1009. https://doi.org/10.1037/0022-3514.55.6.1009

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*. https://doi.org/10.18653/v1/2023.c3nlp-1.7

Hila Chefer, Oran Lang, Mor Geva, Volodymyr Polosukhin, Assaf Shocher, Michal Irani, Inbar

[19] Likely referring to George Orwell, the author of the book *1984*.

Mosseri, and Lior Wolf. 2023. The hidden language of diffusion models. *arXiv preprint arXiv:2306.00966*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. https://doi.org/10.1109/CVPR52729.2023.00276

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models. *arXiv preprint arXiv:2202.04053*. https://doi.org/10.1109/ICCV51070.2023.00283

DeepFloyd. 2023. deepfloyd.ai.

Jacob Devlin. 2018. mbert. https://github.com/google-research/bert/blob/master/multilingual.md

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*. https://doi.org/10.18653/v1/2021.repl4nlp-1.4

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey wave 7 (2017–2022) cross-national data-set. *World Values Survey Association*.

G. Hofstede, G. J. Hofstede, and M. Minkov 2010. *Cultures and Organizations: Software of the Mind* (3rd ed.). McGraw-Hill Professional.

Geert Hofstede. 1983. Dimensions of national cultures in fifty countries and three regions. In J. B. Deregowski, S. Dziurawiec, and R. C. Annis (Eds.), *Expiscations in Cross-Cultural Psychology*, pages 335–355.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. 2024. Visage: A global-scale analysis of visual stereotypes in text-to-image generation.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji B. Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models.

Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 919–933. https://doi.org/10.1145/3581641.3584078

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023. On the cultural gap in text-to-image generation. *arXiv preprint arXiv: 2307.02971*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*. https://doi.org/10.1556/084.2021.00009

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, pages 70–87. Springer. https://doi.org/10.1007/978-3-031-19836-6_5

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions.

Robert R. McCrae and Jüri Allik. 2002. *The Five-factor Model of Personality Across Cultures*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4615-0763-5

Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*. https://doi.org/10.1145/3600211.3604711

Tarek Naous, Michael J. Ryan, and Wei Xu. 2023. Having beer after prayer? Measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*. https://doi.org/10.18653/v1/2024.acl-long.862

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *CoRR*, abs/2211.13069.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense QA data? Case study in indonesian and sundanese. *ArXiv*, abs/2402.17302.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*. https://doi.org/10.18653/v1/2023.acl-long.26

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *ArXiv*, abs/2404.12464.

Milton Rokeach. 1967. Rokeach value survey. *The Nature of Human Values*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. https://doi.org/10.1109/CVPR52688.2022.01042

Michael Saxon and William Yang Wang. 2023. Multilingual conceptual coverage in text-to-image models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4831–4848, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.266

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shalom H. Schwartz. 1994. *Beyond individualism/collectivism: New cultural dimensions of values*. SAGE Publications, Inc.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. ‘‘I’m sorry to hear that’’: Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.625

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2022. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*. https://doi.org/10.1613/jair.1.15388

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211. https://doi.org/10.18653/v1/2022.gebnlp-1.21

Harry C. Triandis and Michele J. Gelfand. 1998. Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology*, 74(1):118. https://doi.org/10.1037/0022-3514.74.1.118

Henriikka Vartiainen and Matti Tedre. 2023. Using artificial intelligence in craft education: Crafting with text-to-image generative models. *Digital Creativity*, 34(1):1–21. https://doi.org/10.1080/14626268.2023.2174557

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563. https://doi.org/10.1609/aaai.v37i2.25353

Alan Watts. 1989. On the taboo against knowing who you are. *New York: Randomhouse*.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. 2023. Altdiffusion: A multilingual text-to-image diffusion model. *arXiv preprint arXiv:2308.09991*.

Eunice Yiu, Eliza Kosoy, and Alison Gopnik. 2023. Imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet)?

Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K. Wong. 2024. Bridging different language models and generative vision models for text-to-image generation. *arXiv preprint arXiv:2403.07860*. https://doi.org/10.1007/978-3-031-73004-7_5

A Appendix: Complementary Results

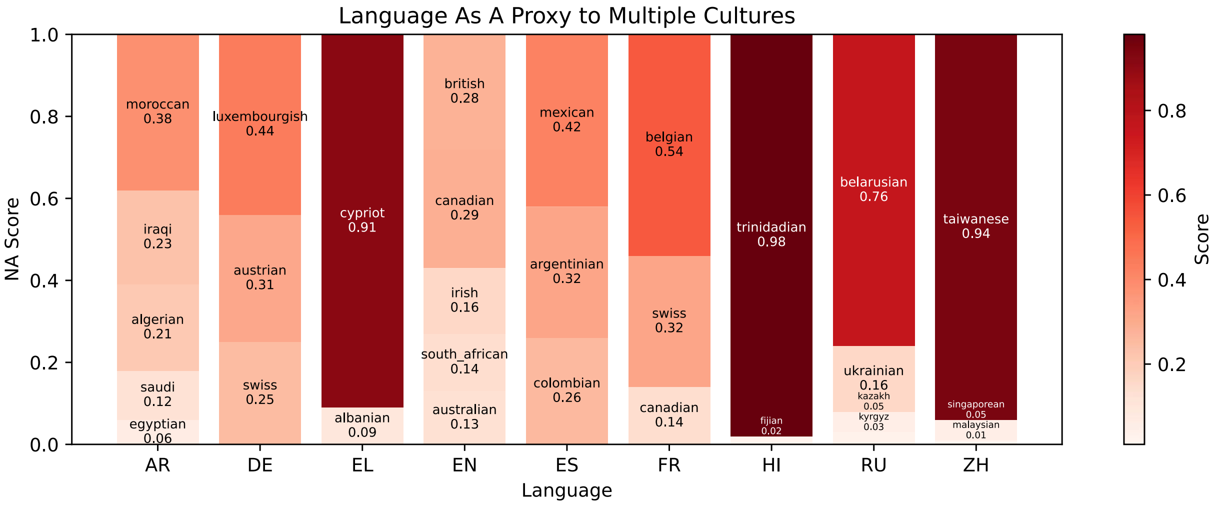


Figure 11: Language as a proxy to multiple cultures (Finding 2 in §8). NA scores (see §6.1) account for additional nationalities that speak these languages, using SD 2.1v images with Translated Concept PT across all concepts. Ten languages (x-axis) with up to 5 other nationalities that primarily speak each language, which are not represented in the paper. Color encode NA score. Blocks represent nations, and the size is relative to the score.

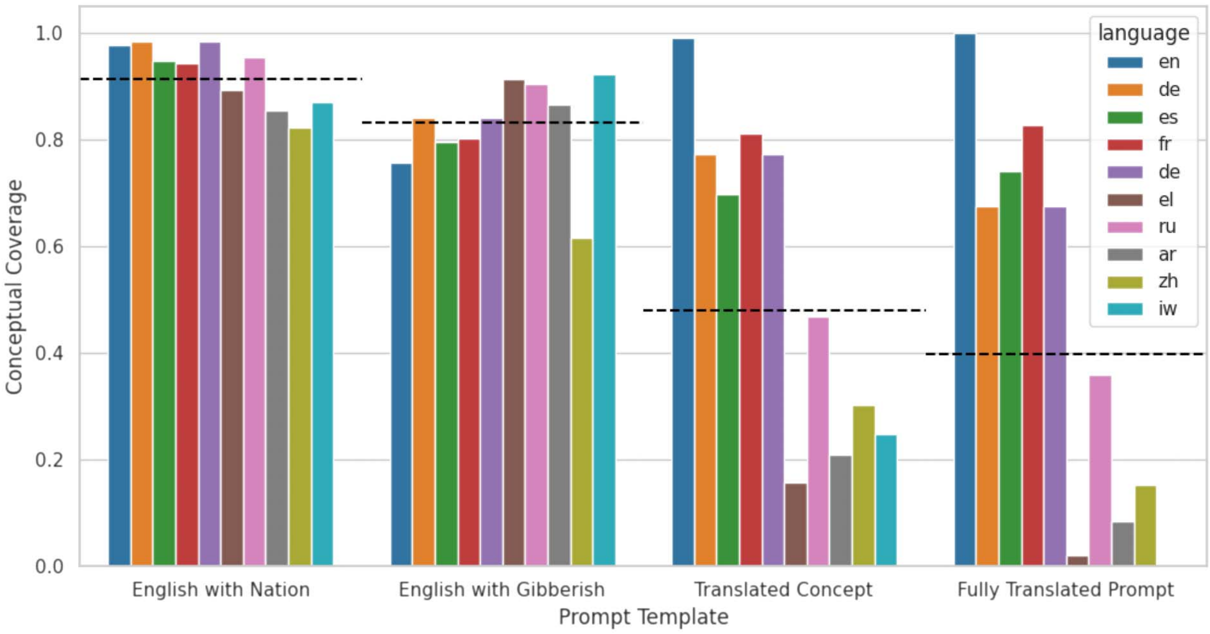


Figure 12: Prompt template impact on conceptual coverage (Prompt Templates in §5). The normalized score is between the CLIP embeddings of BLIP2's image descriptions and their corresponding concept (y-axis) over the prompt templates (x-axis). We experiment with images generated by StableDiffusion 2.1v, covering over 40 tangible cultural concepts. Color encodes language.

Language	**AD**	**DL**	**DF**	**LB**	**SD1.4**	**SD2.1**	**Mean per Language**
AR	0.06447	0.42130	**0.71053**	0.37441	0.49131	0.46919	0.422
DE	0.38389	0.37654	0.37018	0.10269	0.38705	0.35071	0.329
EL	0.09277	0.43519	0.49474	0.17062	**0.52449**	**0.50869**	0.371
EN	0.12638	0.14506	0.22807	0.06161	0.20379	0.20063	0.161
ES	0.09479	0.38272	0.27544	0.10111	0.26856	0.38231	0.251
FR	0.45912	0.36728	0.34561	0.09795	0.42654	0.43444	0.355
HI	**0.61478**	**0.62500**	**0.73684**	**0.74250**	**0.71564**	**0.72512**	**0.693**
IW	0.03145	0.40741	**0.66842**	0.02370	0.42496	0.42812	0.331
RU	0.42453	**0.73148**	**0.70526**	**0.90679**	**0.78357**	**0.85940**	**0.735**
ZH	0.22484	0.31481	**0.76842**	0.27014	**0.50711**	**0.57820**	0.444
Mean per Model	0.252	0.421	**0.530**	0.285	0.473	0.494	0.409

Table 4: Identifying cultural origin (XNA): Full results (Finding 1 in §8). XNA mean scores over translated PTs. Scores above 0.5 are in bold.

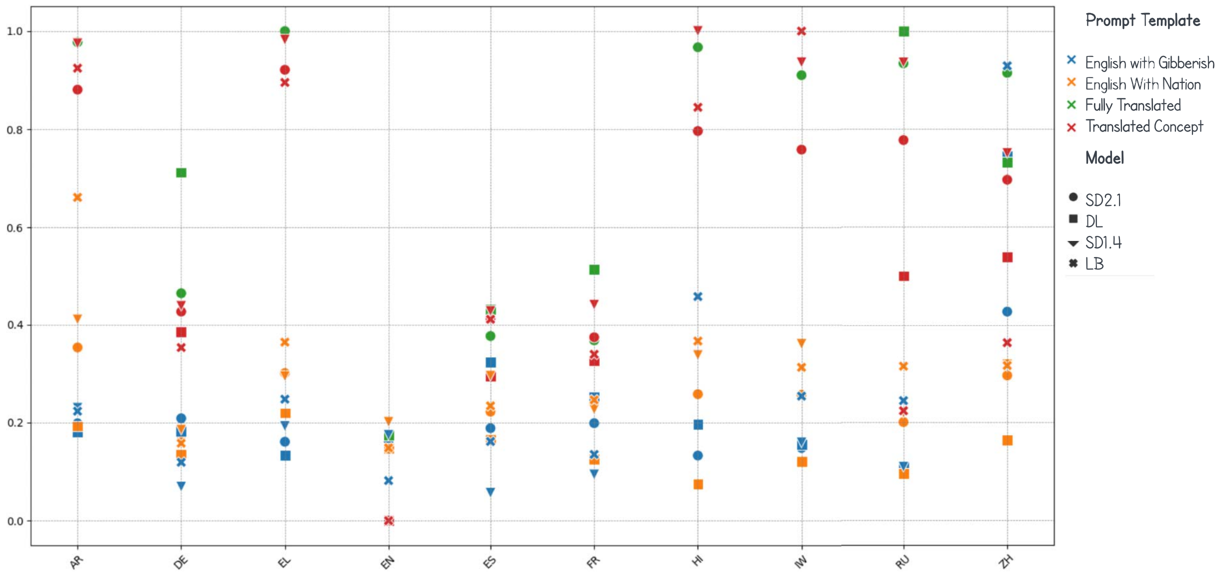


Figure 13: Cultural distance (CD): Full results (Finding 4 in §8). CD of TTI models with an implicit text encoder across 4 Prompt Templates. Languages are shown on the x-axis. Color notes PT and shape notes model. Normalized scores are presented. The more translated parts in the prompt, the greater the distance, especially in non-Latin languages, which show variations across models.

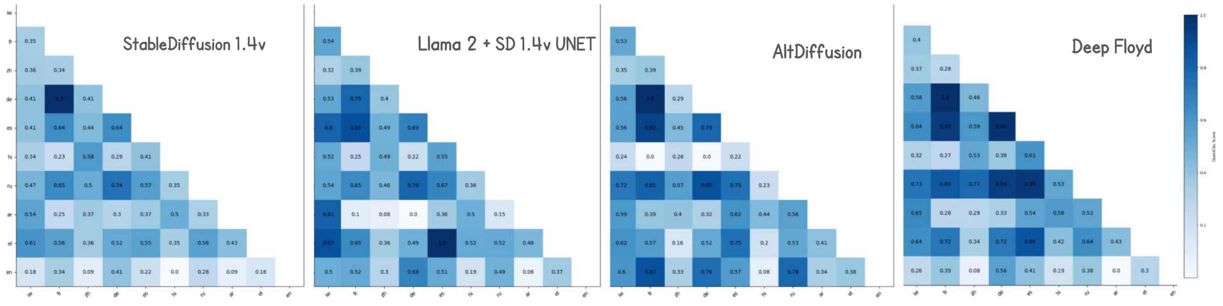


Figure 14: Cross-cultural similarity (CCS): Full results (Finding 4 in §8). CCS metric analysis for the ‘EN with Nation’ PT. Higher values represent higher similarity. Normalized scores are presented.

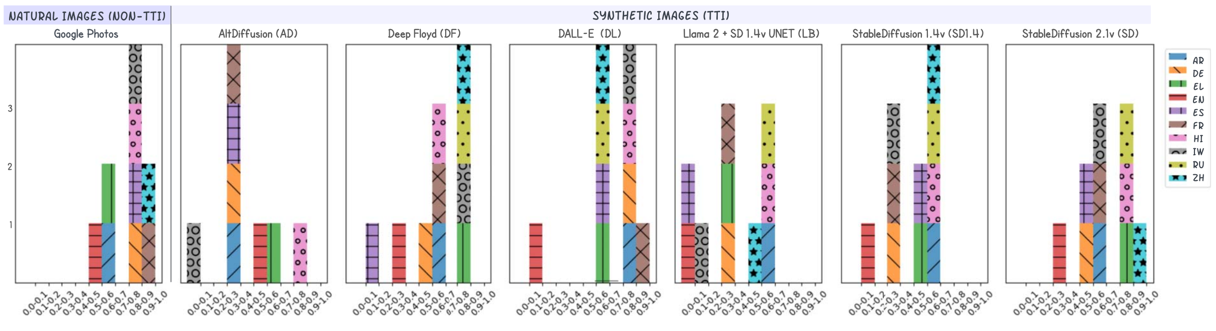


Figure 15: Expected performance of XNA: Natural images vs. synthetic (Analysis 3 in §9). Scores are presented as histograms. Results with English with Nation PT across 6 concepts.

Language/Model	**SD**	**DL**	**AD**
de	66.67% / 27.78%	58.33% / 47.22%	52.78% / 33.33%
es	63.89% / 47.22%	47.22% / 38.89%	41.67% / 8.33%
ru	91.67% / 91.67%	77.78% / 72.22%	55.56% / 44.4%
zh	100.00% / 61.11%	91.67% / 47.22%	83.33% / 22.22%
Overall	81.00% / 56.94%	68.75% / 51.38%	58.33% / 27.08%

Table 5: Revalidating national association findings with GPT-4-Vision (Analysis 4 in §9). Extrinsic national association (XNA) mean scores, presented as an average success ratio over the languages, models, concepts and PTs as in the human evaluation set. Left: GPT-Vision score, right: BLIP2 score.

Model	Language	Modernity - Ancient	Femininity - Masculinity	Rationality - Tradition	Self-Expression - Survival	Extraversion - Introversion	Kindness - Critical	Individualism - Collectivism	Human - Nature
SD2.1v	AR	−0.50616	0.43033	−0.75355	0.49384	0.5071	0.45687	0.56967	0.61611
	DE	−0.1148	0.0702	−0.25712	0.36148	0.41271	0.13852	0.5759	0.78748
	EL	−0.47393	0.45877	−0.53175	0.02275	0.49858	0.08815	0.54218	0.70332
	EN	0.25593	0.21327	−0.30427	0.36967	0.61043	0.10236	0.54408	0.7801
	ES	−0.10901	0.2436	−0.52606	0.38862	0.52702	0.28341	0.45971	0.8436
	FR	−0.06161	0.27014	−0.34881	0.29573	0.53649	0.25213	0.57346	0.83886
	HI	−0.51659	0.29478	−0.74882	0.44076	0.25119	0.52322	0.66919	0.61043
	IW	−0.02464	0.05308	−0.2218	0.09289	0.72227	0.1981	0.58293	0.76209
	RU	0.12606	0.26824	−0.23223	0.3346	0.55545	0.02464	0.65687	0.80853
	ZH	−0.00853	0.58199	−0.82843	0.73176	0.48815	0.75356	0.7962	0.65118
DL	AR	0.24285	0.16191	−0.49523	0.19048	0.69524	0.35714	0.86667	0.68095
	DE	0.35127	0.4019	−0.31962	0.08861	0.64873	0.37342	0.89557	0.43038
	EL	0	0.39436	−0.52113	0.0892	0.723	0.4554	0.76996	0.55399
	EN	0.66197	0.30047	−0.39906	0.20658	0.76995	0.39906	0.82629	0.61503
	ES	0.23676	0.50779	−0.38007	0.14019	0.72897	0.45795	0.84112	0.4081
	FR	0.36448	0.5109	−0.31464	0.12461	0.67601	0.41433	0.86916	0.39564
	HI	0.03792	0.34597	−0.48341	0.06162	0.58294	0.43602	0.88626	0.51659
	IW	0.46262	0.22897	−0.18224	−0.00935	0.66356	0.2944	0.78504	0.63551
	RU	0.40125	0.31662	−0.26959	0.01254	0.60502	0.35737	0.87148	0.40439
	ZH	0.18809	0.68652	−0.4859	0.34797	0.59561	0.60815	0.87461	0.18808
DF	AR	0.24211	−0.03684	−0.76842	0.54211	0.72105	0.52105	0.65263	0.82632
	DE	0.41579	0.08737	−0.23053	0.58105	0.55053	0.45053	0.67895	0.76105
	EL	−0.1421	0.12106	−0.6	0.47895	0.73684	0.54737	0.45263	0.62632
	EN	0.55158	0.01369	−0.17158	0.50211	0.59894	0.3379	0.64947	0.82632
	ES	0.50421	0.28843	−0.32316	0.58105	0.55684	0.48	0.66	0.80632
	FR	0.33685	0.26105	−0.26948	0.50526	0.44211	0.40842	0.67157	0.77369
	HI	−0.08947	0.06315	−0.78948	0.53158	0.66316	0.6	0.70526	0.77368
	IW	0.61579	−0.14211	−0.6	0.33158	0.7579	0.50527	0.61579	0.83158
	RU	0.39158	0.04736	−0.15158	0.53473	0.43053	0.48316	0.75369	0.74527
	ZH	0.25789	0.18421	−0.56843	0.49474	0.61579	0.53158	0.66315	0.81053
LB	AR	−0.64455	−0.06793	0.96366	−0.45655	0.32859	−0.20064	0.55608	0.68405
	DE	−0.05687	0.01422	0.87836	−0.00474	0.41864	−0.16114	0.12638	0.4534
	EL	−0.53239	0.17693	0.84992	−0.14218	0.47868	−0.18483	0.25592	0.51343
	EN	0.259702	−0.05687	0.87046	0.00948	0.38546	0.17061	0.13428	0.65719
	ES	−0.26541	0.14692	0.89732	−0.20221	0.54502	0	0.29858	0.49921
	FR	−0.07267	0.16271	0.8831	−0.04108	0.50079	−0.0553	0.22274	0.48025
	HI	−0.73143	−0.1801	0.98104	−0.68404	0.73775	0.00632	0.77567	0.60347
	IW	−0.48973	0.05055	0.90521	−0.06635	0.30806	−0.34123	0.21485	0.45182
	RU	−0.05213	−0.10269	0.90679	−0.00158	0.15166	−0.11374	0.10426	0.68563
	ZH	−0.48657	0.47077	0.9605	−0.67299	0.75039	0	0.62717	0.68563
SD1.4v	AR	−0.29747	0.38765	0.75475	−0.77848	0.12816	0.56013	0.2943	−0.02848
	DE	0	0.28481	0.82753	−0.4019	0.28639	0.61392	0.24842	0.34969
	EL	−0.46361	0.44621	0.91455	−0.63607	0.2231	0.69779	0.40981	0.24051
	EN	0.25515	0.29319	0.85103	−0.38194	0.27893	0.71474	0.18701	0.23297
	ES	0.0981	0.44146	0.90032	−0.43987	0.43038	0.75316	0.44146	0.27848
	FR	−0.08386	0.42089	0.85601	−0.49525	0.27215	0.57911	0.35443	0.28956
	HI	−0.55854	0.43038	0.79747	−0.86392	0.20728	0.58861	0.65665	−0.23259
	IW	0.00949	0.28639	0.72469	−0.51898	−0.02057	0.65664	0.24684	−0.10285
	RU	0.0538	0.33702	0.90348	−0.45886	0.0538	0.62342	0.3924	0.39082
	ZH	0.03006	0.49209	0.94621	−0.77057	0.53798	0.59493	0.71044	0.0981

Table 6: Depicting cultural dimensions (XDP): Full results (Finding 3 in §8). Full results of the XDP scores across all cultural dimensions and models.

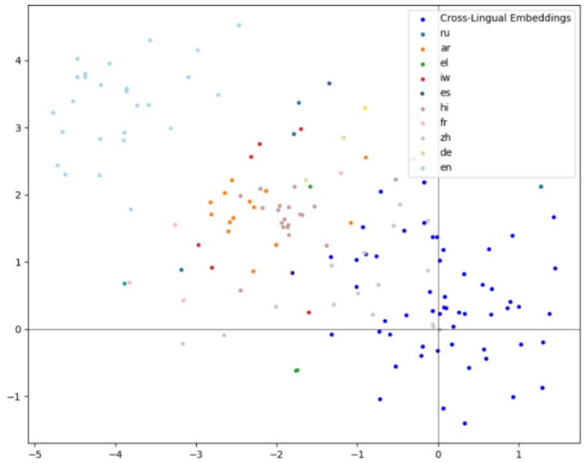


Figure 16: Investigating cultural distances in the textual embedding space (Analysis 2 in §9). t-SNE two-dimensional projection of the textual embeddings of the 10 languages characters. The embedding are of SD 1.4V text encoder, CLIP ViT-L-14.

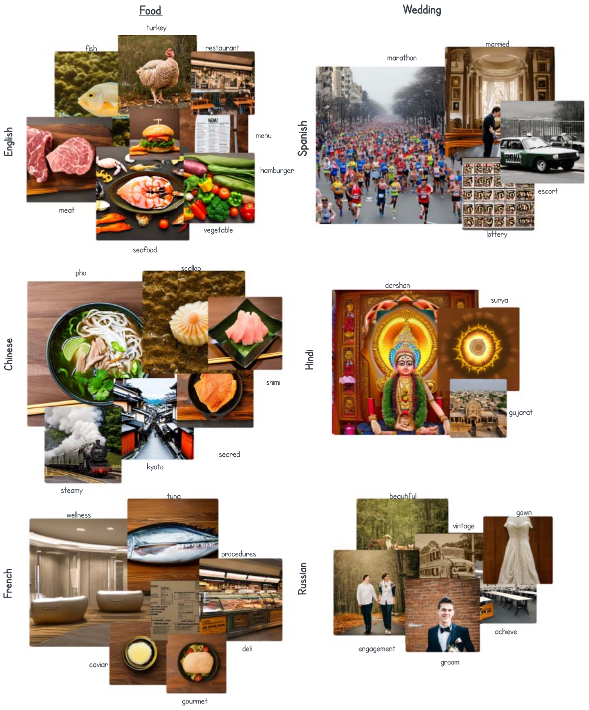


Figure 17: Qualitative results of the Conceptor hidden representations experiment (Analysis 5 in §9). Decomposition of the concept *Food* and *Wedding* to their unique cultural tokens. Examples with prompt template Translated Concept in the languages: Spanish, Hindi, and Russian (top) and English, Chinese, and French (bottom).

B Appendix: Technical Details

In this section, we provide the details required to produce the exact metrics, image generations, and human evaluation.

B.1 TTI Models Technical Details

Model	Version	Scheduler	Inference Steps	Image Size
StableDiffusion	v2.1	EulerDiscrete	50	512 × 512
AltDiffusion	m9	DPMSolverMultistep	50	512 × 512
DeepFloyd	v1.0 I-XL, II-L	–	100	256 × 256
DALL-E	v2	–	–	256 × 256
StableDiffusion	v1.4	–	50	512 × 512
Llavi-Bridge	llama2 + SD 1.4 Unet	–	50	512 × 512

Table 7: Custom technical details of TTI model inference. Default setting is noted with a hyphen.

B.1.1 Technical Details: Conceptor Hidden Representations

The experiment (Analysis 5 in §9) is based on 2 runs for each translated concept: 1) one-step reconstruction to achieve the weights of the linear combination over the tokens in the vocabulary, and 2) single image decomposition to remove the less significant tokens and still conserve a good image reconstruction. The Conceptor utilizes the CLIP encoders (openai/clip-vit-base-patch32). We conducted the experiment using seed number 42.

B.1.2 Conceptual Coverage Formula

To assess the impact of the prompt templates on the conceptual coverage (see Figure 12) we experiment with BLIP 2 to describe the images of 40 tangible concepts (Appendix B.2.4), prompting it with the following prompt: (*Question: What is in the photo? Answer:* (X)). To examine whether the description fits the target concept we compute the cosine similarity score between their CLIP textual embedding: Visual Description $= (\text{VQA}(I_i, \text{X}))_i$ and Conceptual Coverage $= \frac{1}{n}\sum_{i=1}^{n}(\text{visual description}_i \cdot \text{<cc>})$.

Then, we normalize the score by the scores' range for clearer visualization.

B.2 TTI Prompts

B.2.1 Nationalities

In Table 8, the primary column refers to the nationalities used in the National Association (NA) task (§6.1) and in the English with Nation prompt template (§5). The additional nationalities column refers to the second-order NA experiment.

B.2.2 Cultural Concepts (CC) Mapping

Table 9 contains the 200 CCs we defined (§4).

Language (Code)	**Primary Nationality**	**Additional Nationalities**
German (DE)	German	Luxembourgish, Austrian, Swiss]
Greek (EL)	Greek	Cypriot, Albanian
English (EN)	American	British, Canadian, Irish, South-African, Australian
Spanish (ES)	Spanish	Mexican, Argentinian, Colombian
French (FR)	French	Belgian, Swiss, Canadian
Hindi (HI)	Hindi	Trinidadian, Fijian
Arabic (AR)	Arab	Moroccan, Iraqi, Algerian, Saudi, Egyptian
Russian (RU)	Russian	Belarusian, Ukrainian, Kazakh, Kyrgyz
Hebrew (IW)	Israeli	–

Table 8: Language and nationality associations.

B.2.3 Cultural Concepts for Unlocking and Natural Images Experiments

We employ six CCs from the list above for the unlocking experiments (Finding 5 in §8) and the expected performance with natural images (Analysis 3 in §9): *city, food, king, market, nature, car*.

B.2.4 Conceptual Coverage Tangible Concepts

Here we provide the tangible concepts list used in the conceptual coverage analysis (see Figure 12): *university, teacher, school, market, church, wedding, funeral, hospital, doctor, missile, shirt, shoes, jewelry, newspaper, TV, radio, mobile phone, computer, camera, car, plane, leader, bicycle, train, ship, robot, children, flag, king, queen, soldier, cow, dog, cat, fish, horse, bird, snake, baby, elder*.

B.3 Human Assessment

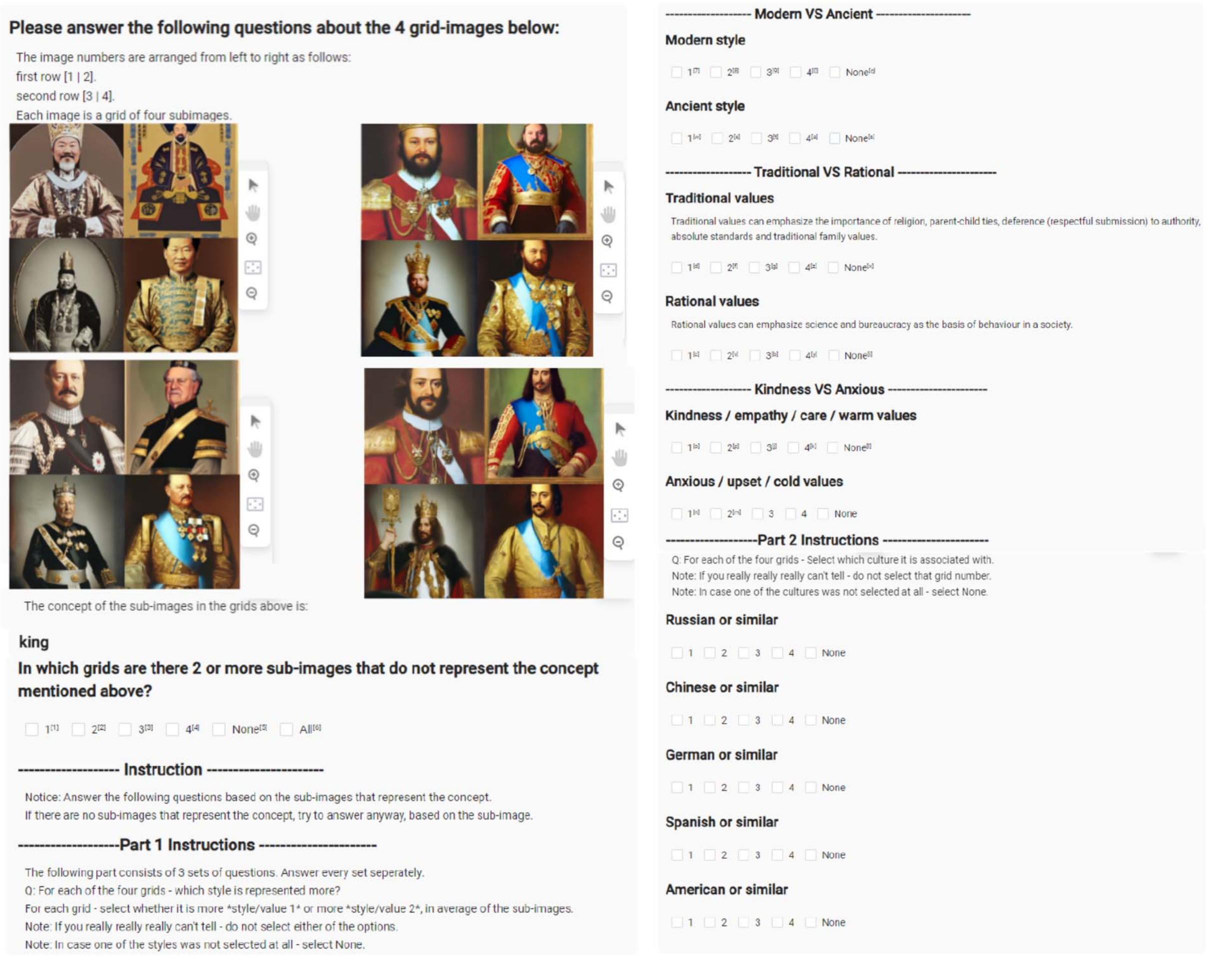
As part of the human evaluation (§6.2), we create the following questionnaire and guidelines (Figures 18, 19).

B.4 Sample System Outputs: Qualitative Examples

We provide image examples (Figures 20, 21) from the CulText2I dataset, that are generated based on our TTI model workflow (§5).

Dear Annotator, Thank you for your invaluable assistance! Your role in this annotation task involves examining grid images, with each image comprising four sub-images. Additionally, you will be required to respond to several follow-up questions. Each task takes around 3 minutes to complete. Please take the time to carefully read the instructions accompanying each question, particularly during your initial run. If you have any suggestions that could contribute to the enhancement of our questionnaire, please don't hesitate to share your insights. A crucial note: The images provided have been generated by Text-To-Image Models. Consequently, some images may contain information that is not relevant or biased. Your dedication to this task is greatly appreciated. Thank you! Best regards

Figure 18: Human questionnaire guidelines.



Please answer the following questions about the 4 grid-images below:

The image numbers are arranged from left to right as follows:
first row [1 | 2].
second row [3 | 4].
Each image is a grid of four subimages.

The concept of the sub-images in the grids above is:

king

In which grids are there 2 or more sub-images that do not represent the concept mentioned above?

□ 1[1] □ 2[2] □ 3[3] □ 4[4] □ None[5] □ All[6]

------------------ Instruction --------------------

Notice: Answer the following questions based on the sub-images that represent the concept.
If there are no sub-images that represent the concept, try to answer anyway, based on the sub-image.

------------------Part 1 Instructions --------------------

The following part consists of 3 sets of questions. Answer every set seperately.
Q: For each of the four grids - which style is represented more?
For each grid - select whether it is more *style/value 1* or more *style/value 2*, in average of the sub-images.
Note: If you really really really can't tell - do not select either of the options.
Note: In case one of the styles was not selected at all - select None.

------------------ Modern VS Ancient ------------------

Modern style

□ 1 □ 2 □ 3 □ 4 □ None

Ancient style

□ 1 □ 2 □ 3 □ 4 □ None

------------------ Traditional VS Rational --------------------

Traditional values

Traditional values can emphasize the importance of religion, parent-child ties, deference (respectful submission) to authority, absolute standards and traditional family values.

□ 1 □ 2 □ 3 □ 4 □ None

Rational values

Rational values can emphasize science and bureaucracy as the basis of behaviour in a society.

□ 1 □ 2 □ 3 □ 4 □ None

------------------ Kindness VS Anxious --------------------

Kindness / empathy / care / warm values

□ 1 □ 2 □ 3 □ 4 □ None

Anxious / upset / cold values

□ 1 □ 2 □ 3 □ 4 □ None

------------------Part 2 Instructions --------------------

Q: For each of the four grids - Select which culture it is associated with.
Note: If you really really really can't tell - do not select that grid number.
Note: In case one of the cultures was not selected at all - select None.

Russian or similar

□ 1 □ 2 □ 3 □ 4 □ None

Chinese or similar

□ 1 □ 2 □ 3 □ 4 □ None

German or similar

□ 1 □ 2 □ 3 □ 4 □ None

Spanish or similar

□ 1 □ 2 □ 3 □ 4 □ None

American or similar

□ 1 □ 2 □ 3 □ 4 □ None

Figure 19: Annotator questionnaire example of the CC: *king*.

Cultural Domain	Cultural Concepts (CCs)	Cultural Reference
Moral Discipline and Social Values	independence, thrift, drug addiction, race, AIDS, immigrants, homosexuality, heavy drinkers, unmarried couples, proud parents, feminism, housewife, cheating, abortion, divorce, sex, suicide, violence, death penalty, surveillance, desire, masculinity, femininity, pleasure, animal (monster)	WVS (Social Values, Ethical Values and Norms) | Hofstede (Masculinity vs Femininity, Indulgence Vs Restrained) | Schwartz (Conformity, Hedonism)
Education	university, teacher, science, school, intelligent person, expert (physics, chemistry, history, biology, engineer, mathematics, literature)	WVS (Corruption) | Hofstede (Power Distance)
Economy	market, industry, cash, bank, economy, boss, job, factory, agriculture, salary, rich person, poor person, money (payroll, mortgage, tax)	WVS (Economic Values, Postmaterialist Index)
Religion	holiday, church, soul, religion, god, death, hell, heaven, wedding, funeral, pray (priest, synagogue, Judaism, Christianity, Islam, Hinduism, Buddhism, cow, snake)	WVS section Religious Values | Hofstede (Power Distance) | Schwartz (Tradition)
Health	mental health, healthcare, hospital, doctor, medicine, treatment (baby, elder, young, teenager, pill, sleep, memory)	WVS (Corruption)
Security	robbery, alcohol consumption, war, civil war, terrorism, crime, jobs vacancy, missile, cyber, unemployment, protection, attack, weapon, peace	WVS (Corruption, Migration, Security) | Schwartz (Security)
Aesthetics	beauty, art, music, drama, dancing, sport organization, food, fashion, beverage (nature, dog, cat, fish, horse, bird, shirt, shoes, jewelry, baseball)	WVS (happiness & Well-being) | Schwartz (Universalism, Harmony)
Material Culture	tool, transportation, power, communication, technology, newspaper, TV, radio, mobile phone, computer, camera, car, plane, bicycle, train, ship, robot (social media)	WVS (Science & Technology, Political Interest and Political Participation)
Personality Characteristics and Emotions (adjective + ‘‘person’’)	neurotic, concerned, shamed, angry, nervous, happy, extravert, introvert, energized, confident, curious, cynical, capable, empathic, obedient, lazy, expressive, friendly, dominant, communicative, proud, polite, truthful, independent, creative	Hofstede (Uncertainty Avoidance, Individualism vs Collectivism, Indulgence Vs Restrained) | Schwartz (Self-Direction) | Personality Across cultures (Big Five) | Rokeach (Instrumental Values))
Social Capital Organizational Membership	family, city, children, father, mother, neighborhood, home, nation, army, grandmother, grandfather, courts, government, political party, police, elections, charity, EU, UN, protest, leader, democracy, human rights, nation, flag, king, queen, soldier (journalist)	WVS (Social Capital, Trust and Organisational Membership, Political Culture and Regimes)| Hofstede (Short Vs Long Term Orientation) | Schwartz (Power, Benevolence)
Countries	(America, China, Russia, Germany, France, Spain, Egypt, India, Arab, Israel)	

Table 9: Cultural Concepts (CCs) and domains table. The CCs are drawn from the definitions and domains in the cultural research reference. The specific value/section in the cultural origin is mentioned in parenthesis. The CCs in parenthesis are additional concepts we added for enrichment with more detailed in-domain concepts. WVS notes World Values Survey.

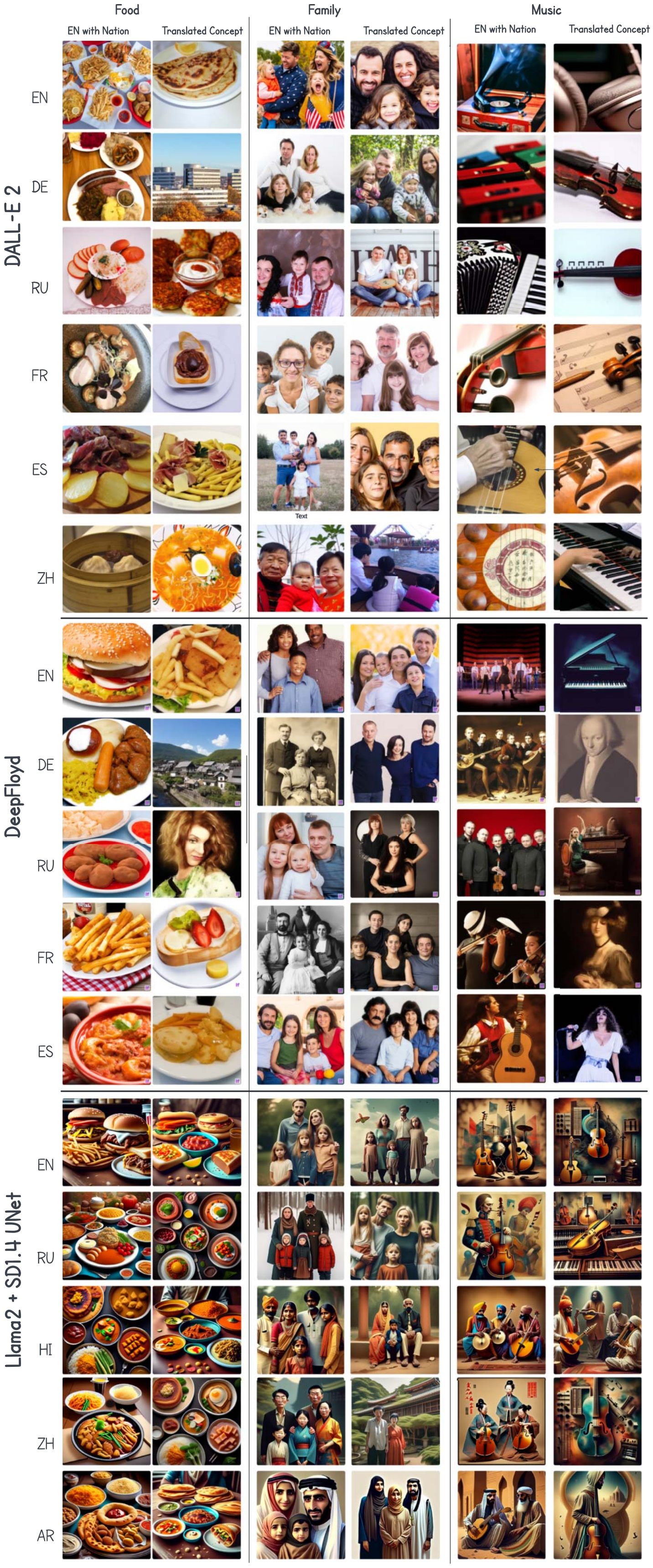


Figure 20: Qualitative examples. Images of Cultural Concepts: *food* (left), *family* (middle) and *music* (right), by **DALL-E** (top) and **DeepFloyd** (middle) and **Llama2 + SD1.4 UNet** (bottom).

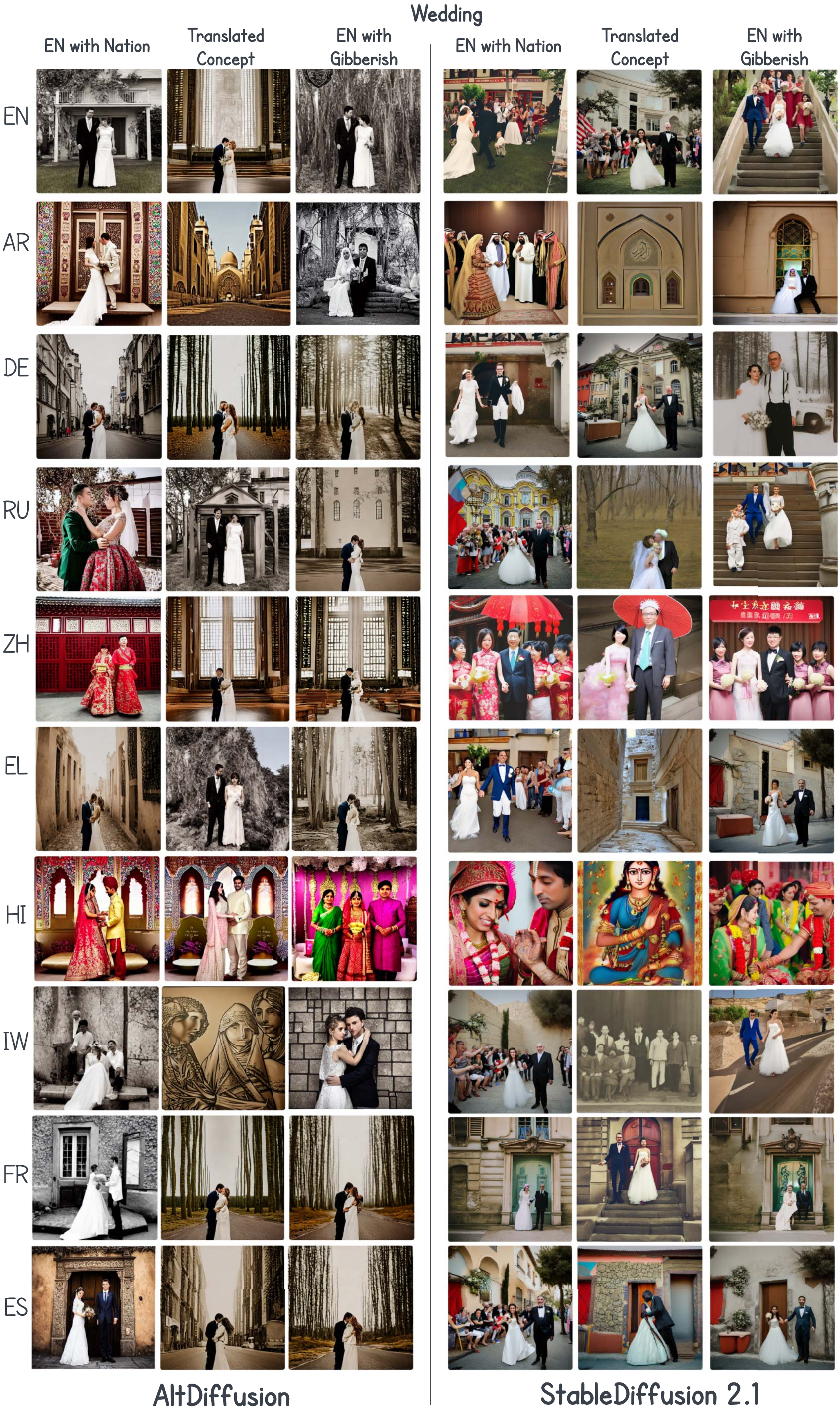


Figure 21: Qualitative examples. Images of Cultural Concept: *wedding*, by **AltDiffusion** (left) and **StableDiffusion** (right).