Learning Cultural Vectors for Cross-Cultural Generation

Sina Malakouti
University of Pittsburgh
sem238@pitt.edu

Deepti Ghadiyaram
Boston University
dghadiya@bu.edu

Boqing Gong
Boston University
bgong@bu.edu

Adriana Kovashka
University of Pittsburgh
kovashka@cs.pitt.edu

https://sinamalakouti.github.io/CulturalVectors

Abstract

Text-to-image diffusion models often struggle to faithfully represent diverse cultural contexts. We study culturally grounded image generation through cultural vectors, defined as the difference between model parameters fine-tuned on culturally grounded data and their pre-trained counterparts. First, we show that scaling cultural vectors at inference time controls the trade-off between cultural alignment, exaggeration, and hallucination. This enables learning cultural vectors from synthetic data while mitigating the exaggerated cultural depictions that may arise from biases in the synthetic images. Second, our analysis reveals that cultural knowledge is primarily encoded in deeper U-Net blocks, while activation-space analysis reveals a strong shared structure across countries. Finally, we investigate the compositionality of cultural vectors. We find that unlike prior work on task arithmetic, naively combining cultural vectors fails to produce meaningful multi-cultural behavior due to strong interference across cultures and degrades generalization to unseen countries. To reduce cultural interference, we introduce Cultural Activation Preservation (CAP), which encourages the preservation of other cultures' representations during training and improves compositionality. To improve generalization to unseen countries, we further introduce Descriptor-Guided Merging (DGM), a culturally aware approach that leverages textual descriptors to estimate cross-cultural similarities and merge cultural vectors. DGM improves both Task Arithmetic and TIES, with TIES+DGM achieving the lowest mean rank in human evaluation and ranking first in 10 of 11 countries.

1 Introduction

Consider the depiction of a simple act of a handshake in Iran. A text-to-image (T2I) model may place a mosque in the background, even when the prompt does not specify a religious setting. Although recent T2I models [10, 27, 4, 20, 30, 9, 34] have fundamentally redefined the boundaries of photorealism and creativity, they remain **unreliable** when generating culturally grounded content. Models frequently depict incorrect actions or objects, stereotypical traditional attire, exaggerated scenes, and religious elements unrelated to the scene and activity. Recent works have primarily focused on benchmarking and evaluating cultural biases in T2I models, revealing systematic limitations in representing diverse cultural concepts, particularly for underrepresented cultures [2, 17, 33, 26, 24]. These findings motivate a broader investigation into how cultural information can be learned and controlled within T2I generative models.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this work, we ask whether diffusion models can learn meaningful *cultural representations*, and if so, can it be *controlled* to improve cultural faithfulness. Answering these questions is important for developing principled approaches to improve cultural faithfulness across diverse cultures while mitigating the amplification of stereotypes in generated images. However, addressing these questions is challenging for two key reasons. First, cultural attributes are expressed through multiple interacting visual elements, such as clothing, objects, and scenes, making them difficult to control during generation. Second, high-quality culturally grounded data is **scarce**, particularly for underrepresented regions, while synthetic alternatives may inherit biases from their source models.

To investigate these questions, we conduct a systematic study, modeling cultural representation as parameter differences between fine-tuned and pretrained models. We refer to these as *cultural vectors*. Inspired by task arithmetic [13], this formulation provides two potential advantages that directly address the challenges above. First, cultural vectors can be scaled at inference time to control the strength of cultural adaptation. Second, vectors learned for different countries can be combined to investigate multi-cultural generation and generalization to countries without target-specific training data. However, while task vectors enable editing and composition in foundation models, this task (cultural generation) is *unique in two ways*: (1) we are interested in generation (via diffusion models), and (2) capturing cultural knowledge in settings where high-quality visual training data is limited, despite the broader availability of cultural knowledge. These challenges motivate a systematic investigation into how cultural representations can be learned, controlled, and composed within diffusion models.

Towards this goal, we first investigate the **properties** of cultural vectors along multiple dimensions: (1) we assess their **validity** by examining whether their similarities reflect cultural relationships; (2) we investigate the **sources** of data for learning them (real or synthetic); (3) we study how inference-time **scaling** affects cultural alignment, exaggeration, and hallucination; and (4) we analyze their **structure** across U-Net layers and in parameter and activation spaces to understand how cultural information is represented. Second, we study **compositionality**, investigating whether combined cultural vectors can exhibit multi-cultural behavior and generalize to unseen countries. We find that naive merging leads to cross-cultural interference, motivating two complementary approaches: (1) We propose *Cultural Activation Preservation* (CAP), which regularizes culture-specific fine-tuning to preserve the pretrained model's internal representations of cultures beyond the target culture, thereby reducing interference during vector composition. (2) We further develop *Descriptor-Guided Merging* (DGM), which leverages textual descriptors to estimate cross-cultural similarities and guide vector composition for unseen countries.

We conduct our analysis and compare the proposed methods on CULTIVate [24] and CulturalFrames [26] benchmarks. Our analysis reveals several key findings: (1) Cultural vectors enable controllable generation, improving cultural alignment while mitigating exaggeration. (2) Synthetic data provides an effective signal to learn these vectors, especially when high quality real data is unavailable. (3) Cultural vectors reflect meaningful cultural similarities, with stronger cultural information concentrated in specific deeper U-Net blocks. Moreover, cultural vectors exhibit a shared structure in activation space. (4) Naive merging suffers from cross-cultural interference, reducing faithfulness to individual cultures, while existing merging methods underperform the pretrained model when generalizing to unseen countries. CAP reduces interference, whereas DGM improves cross-cultural generalization, highlighting the importance of culturally aware composition. Human evaluation further supports DGM, with TIES+DGM achieving the lowest mean rank and ranking first in 10 of 11 countries.

In summary, our contributions are threefold: (1) We introduce *cultural vectors* as a framework for studying cultural representations in diffusion models, showing that they can be learned from synthetic data and controlled through inference-time scaling to improve cultural faithfulness while mitigating exaggeration. (2) We systematically investigate how cultural information is encoded within diffusion models, analyzing cultural similarities, block-wise representations, and the structure of cultural vectors in parameter and activation spaces. (3) We study the compositionality of cultural vectors, revealing cross-cultural interference under naive merging. We introduce CAP and DGM to improve multi-cultural composition and generalization to unseen countries, respectively.

2 Related Works

Cultural bias in T2I models. Recent works [24, 26, 33, 17, 2, 15, 18] have studied cultural bias in T2I models, showing that current models suffer from limited cultural diversity [17, 2] and fail to faithfully generate culturally grounded concepts [33, 18, 26, 24], particularly for underrepresented regions. Specifically, [26, 24] show that models struggle with cultural scenarios that require context (e.g., social activities and social norms), often producing plausible yet culturally incorrect or exaggerated images. However, existing work primarily focuses on evaluating these failures through benchmarks and metrics, with limited understanding of how cultural knowledge is represented and controlled within diffusion models. In this work, we study cultural representations directly in model parameter space through cultural vectors, enabling controllable cultural generation and cross-cultural generalization.

Debiasing diffusion models. Various methods have been proposed to mitigate harmful or biased generation in diffusion models, primarily focusing on gender, racial, and safety biases [35, 16, 21]. In contrast, cultural generation remains heavily underexplored and is more challenging due to its context-dependent and compositional nature. Concretely, [22] improve cultural generation through contrastive finetuning on culturally grounded prompts, while [14] enhances generation through culturally-aware prompt refinement using external cultural knowledge bases. However, these approaches rely on additional training data or external retrieval systems. Differently, we study how cultural knowledge is represented directly in model parameter space through cultural vectors, enabling controllable cultural generation and cross-cultural generalization.

Weight interpolation and task arithmetic. Recent work has shown that independently finetuned models often lie in connected regions of parameter space, enabling interpolation and merging across models without significant performance degradation [37, 7, 8, 32]. Building on this observation, task arithmetic [13] represents task vectors as parameter differences between finetuned and pretrained models, enabling controllable editing, transfer, and multi-task composition through simple vector operations. A body of works [28, 39, 41, 36] further proposed methods for addressing task interference and improving multi-task performance. However, application of task arithmetic and model merging remains understudied in T2I generative models. Diffusion Soup [5] extends [37] to diffusion models, while Deme [23] learns and merges timestep-specific representations for improved diffusion finetuning on the same generation task, and [29] utilizes task vectors for concept unlearning in diffusion models. In contrast, we extend task arithmetic [13] to culturally grounded generation and study whether cultural knowledge can be represented, controlled, and composed in diffusion. To the best of our knowledge, this is the first work to study cultural representations in diffusion model parameter space and their use for cross-cultural generation.

3 Methodology

3.1 Cultural Vectors

Preliminaries. Latent diffusion models (LDMs), e.g., Stable Diffusion [34], generate images by iteratively denoising Gaussian noise over T timesteps. During training, the forward process gradually corrupts a clean latent x_0 by adding Gaussian noise ϵ, yielding a noisy sample x_t:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t$ controls the noise level at timestep t.

The reverse process aims to recover x_0 by iteratively modeling posterior $p_\theta(x_{t-1} \mid x_t, p)$. In practice, a U-Net $\epsilon_\theta(x_t, p, t)$ predicts the noise at timestep t, conditioned on a text prompt p. The model is trained using the denoising objective:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t, x_0, p, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, p, t)\|_2^2\right]. \tag{2}$$

Learning cultural vectors. We investigate how cultural knowledge is encoded in diffusion models and whether it can be controlled to generate culturally faithful images. Inspired by task arithmetic [13], we define *cultural vectors* as the differences between culture-specific fine-tuned and pretrained model parameters. Let θ_0 denote the pretrained U-Net parameters. For each country c, we construct a dataset $\mathcal{D}_c$ of culturally grounded image-caption pairs and fine-tune the model using Eq. 2, obtaining

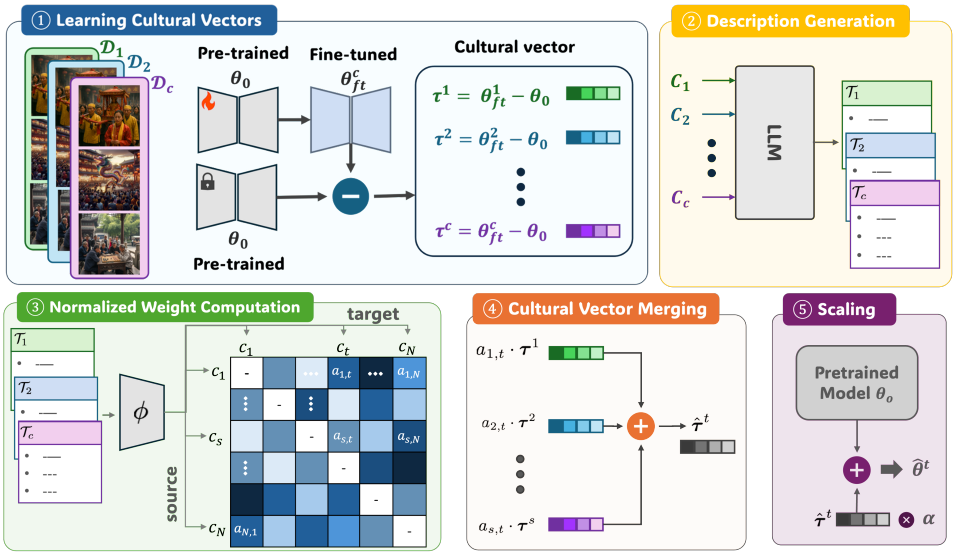


Figure 1: **Overview.** (1) For each country, pre-trained diffusion model is fine-tuned on cultural data to learn cultural vectors. (2) We generate cultural descriptors for each country using an LLM. (3) Descriptor embeddings are used to compute similarities between source and target countries. (4) Descriptor-Guided Merging (DGM) estimates a target cultural vector by weighted merging of culturally similar source vectors. (5) The resulting vector is scaled and applied to the pre-trained model at inference time to control generation.

culture-specific parameters θ^c_{ft}. The corresponding cultural vector is defined as:

$$\boldsymbol{\tau}^c = \theta^c_{ft} - \theta_0. \tag{3}$$

At inference time, we apply the cultural vector to the pretrained model using a scaling coefficient α:

$$\hat{\theta}^c = \theta_0 + \alpha \cdot \boldsymbol{\tau}^c. \tag{4}$$

The coefficient α controls the strength of cultural adaptation, enabling interpolation between the pretrained and the fine-tuned parameters without additional training. Throughout the paper we write $\boldsymbol{\tau} = \theta_{ft} - \theta_0$ and apply $\theta_0 + \alpha \cdot \boldsymbol{\tau}$.

3.2 Sources and Properties of Cultural Vectors

Cultural vectors provide a mechanism for controlling cultural adaptation, but the data sources required to learn them and their properties remain to be understood. We study these properties along the following dimensions (results in Sec. 5.2).

Source of data. For the dataset $\mathcal{D}_c$ in Sec. 3.1, one would most intuitively leverage real data samples from the respective countries. However, real images are difficult to collect, particularly for low-resource countries, and automated web crawling may result in some irrelevant samples being retrieved. Synthetic images provide an alternative but may inherit cultural biases and stereotypes from their source models. We compare both sources and analyze their effects on cultural alignment, exaggeration, and hallucination.

Cultural similarity. To assess whether cultural vectors reflect meaningful cultural relationships, we measure the correlation between pairwise cultural vector similarities and cultural descriptor embedding similarities, using the latter as a proxy for cultural relatedness.

Scaling and controllability. Next, we study how the scaling coefficient α affects generation behavior. By varying α at inference time, we analyze how the strength of the cultural representation influences cultural alignment, exaggeration, and hallucination.

Block-wise cultural representations. To understand how cultural information is distributed across the U-Net, we learn cultural vectors from different blocks and compare their effect on cultural generation metrics from [24] (alignment, exaggeration, and hallucination).

Structure of cultural vectors & activation vectors. To understand the structure of cultural representations, we analyze cultural vectors in both parameter and activation spaces. Following [13], we descriptively check whether cultural vectors concentrate in a shared low-dimensional subspace of parameter space. We perform SVD on $V \in \mathbb{R}^{C\times D}$, where each row is a flattened cultural vector $\boldsymbol{\tau}^c$ and D is the number of parameters, and compute the explained variance ratio and effective rank:

$$\mathrm{EVR}_k = \frac{\sigma_k^2}{\sum_j \sigma_j^2}, \qquad r_{\mathrm{eff}} = \exp\left(-\sum_k q_k \log q_k\right), \quad q_k = \mathrm{EVR}_k, \tag{5}$$

where σ_k denotes the k-th singular value. Similarly, we obtain activation vectors for each U-Net block by averaging the differences between the activations of the culturally adapted model $\theta_0 + \alpha\boldsymbol{\tau}^c$ and the pretrained model θ_0 over a shared set of probe samples. We stack the flattened activation vectors across countries into a matrix A_l, where l denotes the U-Net block, and apply the same SVD analysis. Following [36], we analyze CKA [19] in the appendix.

3.3 Compositionality of Cultural Vectors

We study whether cultural vectors can be composed to enable multi-cultural generation and generalization to unseen countries. Although prior work has shown that models obtained by linearly combining task vectors can exhibit multi-task behavior [13], composing cultural vectors may lead to cross-cultural interference. To improve compositionality, we explore two complementary approaches: preserving pretrained representations of non-target cultures during fine-tuning and leveraging cross-cultural similarities to guide vector merging. We present the results in Sec. 5.3.

Composition of cultural vectors. Given a set of countries $\mathcal{C}$, we define the merged cultural vector as:

$$\boldsymbol{\tau}_{\mathrm{merged}} = \sum_{c\in\mathcal{C}} a_c \boldsymbol{\tau}^c, \tag{6}$$

where a_c is the merging weight for country c. By default, we use uniform averaging ($a_c = 1/|\mathcal{C}|$). We evaluate composition in two settings: (1) merging subsets of countries and evaluating faithfulness on both the merged countries and the full country pool, and (2) leave-one-country-out generalization, where source vectors are combined to estimate a vector for a held-out target country.

Cultural Activation Preservation (CAP). We introduce CAP, an activation-preservation regularization that encourages adaptation to the target culture while preserving the pretrained model's internal representations for non-target cultures. Specifically, when learning $\boldsymbol{\tau}^c$ for country c, we additionally sample image-caption pairs (x_0', p') from countries $\mathcal{C}\setminus\{c\}$ and penalize deviations from the pretrained U-Net's activations:

$$\mathcal{L}_{\mathrm{CAP}} = \mathbb{E}_{t,p',x_0',\epsilon}\left[\frac{1}{L}\sum_{l=1}^{L}\frac{\left\|h_l^{\theta}(x_t', p', t) - h_l^{\theta_0}(x_t', p', t)\right\|^2}{\left\|h_l^{\theta_0}(x_t', p', t)\right\|^2}\right]. \tag{7}$$

h_l^{θ} and $h_l^{\theta_0}$ denote the fine-tuned and frozen pretrained U-Net activations at block l, respectively, over all L down, mid, and up blocks. The final objective is defined as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{LDM}} + \lambda_{\mathrm{CAP}}\mathcal{L}_{\mathrm{CAP}}, \tag{8}$$

where λ_{CAP} controls regularization strength. This encourages each vector to learn its target culture while constraining changes on other countries.

Descriptor-Guided Merging (DGM). We investigate whether cultural similarity can guide composition when no cultural vector is available for a target country t. Given source countries $\mathcal{S} = \mathcal{C}\setminus\{t\}$, we use textual descriptors of cultural attributes, including attire, objects, actions, and scenes, following [24]. For each country c, we average its descriptor embeddings:

$$\mathbf{z}_c = \frac{1}{|\mathcal{T}_c|}\sum_{d\in\mathcal{T}_c}\phi(d), \tag{9}$$

where $\mathcal{T}_c$ is the descriptor set and ϕ is a text encoder. We use normalized, positive mean-centered similarities between the target and source representations as merging weights:

$$\tilde{a}_{s,t} = \max\left(0, \mathrm{sim}(\mathbf{z}_s, \mathbf{z}_t) - \mu_t\right), \qquad a_{s,t} = \frac{\tilde{a}_{s,t}}{\sum_{s'\in\mathcal{S}}\tilde{a}_{s',t}}. \tag{10}$$

Here, $\mu_t = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathrm{sim}(\mathbf{z}_s, \mathbf{z}_t)$ is the mean source similarity. The estimated target vector is

$$\hat{\boldsymbol{\tau}}^t = \sum_{s \in \mathcal{S}} a_{s,t} \cdot \boldsymbol{\tau}^s. \tag{11}$$

Descriptors determine only the merging weights; DGM requires neither target-country training images nor additional fine-tuning.

4 Experimental Setup

4.1 Datasets

CULTIVate [24] covers 576 activities across 9 categories and 16 countries, comprising 5 Global North and 11 Global South countries. Each activity includes culturally grounded descriptors and real images. The benchmark includes 6 T2I models: SD3-Medium [9], FLUX [20], Qwen-Image [38], DALL-E 3 [4], GPT-Image-1 [27], and Gemini 2.5 Flash [10]. We utilized synthetic and real images as training data.

CulturalFrames [26] covers 9 countries (Iran, China, Japan, South Africa, Brazil, Chile, Canada, Poland, Germany). CulturalFrames includes 983 prompts in family, religion, and dates of significance categories, paired with 3,637 images generated by three T2I models (Flux [20], SD3-Large [9], and Imagen3 [1]). These image-text pairs are utilized as our training dataset $\mathcal{D}$ to learn cultural vectors.

4.2 Implementation Details

The models utilize Stable Diffusion v1.5 [34] as backbone, generating images at 512×512 resolution with 30 denoising steps and a classifier-free guidance (CFG) scale [11] of 7.5. Cultural vectors are learned from U-Net parameters, while the VAE and text encoder remain frozen. We fine-tune the U-Net for 1,500 iterations using a learning rate of 10^{-5}, batch size of 4, and mixed precision on a single NVIDIA L40S GPU per country. Inference-time scaling and vector merging require no additional training. For CAP, we set the regularization coefficient $\lambda_{\mathrm{CAP}} = 0.1$. For CULTIVate [24], we use 100 synthetic images per culture generated by three proprietary models, and for CulturalFrames [26], we use 363 images per culture generated by three models. We additionally examine the effect of different image generators, including publicly available models, in the appendix. We use $\alpha = 0.7$ for all methods and baselines to ensure a consistent comparison and study its effect in Tab. 6. All evaluation images are generated with a fixed random seed of 42 for reproducibility. For TIES [39], we follow the original setup, retaining the top $k = 50\%$ of parameter changes and using sum of values to determine the sign.

4.3 Evaluation

Evaluating cultural faithfulness is challenging due to its subjectivity and limited ground truth images. We use two complementary approaches. **Automatic evaluation** follows [24] (best metric on this task and dataset), using GPT-4o [12] to rate cultural alignment (ALIGN), exaggeration (EXAG), and hallucination (HAL) on a 1–5 Likert scale. We compute FAITH $= \frac{1}{3}(\mathrm{ALIGN} + (6 - \mathrm{HAL}) + (6 - \mathrm{EXAG}))$, averaging scores from two evaluations per image. **Human evaluation** uses native annotators recruited through Prolific in two studies: (1) ranking pretrained, TIES, and TIES+DGM generations across 301 prompts from 11 countries, and (2) comparing synthetic and real cultural vectors across 274 prompts from 10 countries. Images are presented in randomized order, with attention checks and at least two annotators per form.

5 Results

5.1 Can we learn meaningful cultural vectors?

Cultural vectors capture cultural knowledge. Tab. 1 shows that cultural vectors learned from real and synthetic data improve cultural faithfulness over the pretrained model and generally outperform full fine-tuning. Synthetic $\boldsymbol{\tau}$ improves FAITH by 18.9% (from 3.75 to 4.46) on CULTIVate and from 4.26 to 4.39 on CulturalFrames, outperforming both the pretrained and full fine-tuning baselines.

Table 1: **Cultural vectors capture cultural knowledge across datasets.** We evaluate on CULTIVate and CulturalFrames. $\alpha = 0.7$, ↑/↓ indicate higher/lower is better and ± is standard deviation. For FAITH, we report relative improvement vs fine-tuning. CulturalFrames lacks real images.

Method	Data	CULTIVate				CulturalFrames			
		ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑
Pre-trained	–	3.44 ±0.33	2.28 ±0.31	1.92 ±0.39	3.75	3.91 ±0.15	1.75 ±0.21	1.38 ±0.15	4.26
τ	Random	3.48±0.33	2.27±0.35	1.89±0.39	3.77	3.90 ±0.12	1.75±0.20	1.70±0.14	4.15
	Synth	**4.56** ±0.22	2.00 ±0.26	**1.18** ±0.10	**4.46** (+4.2%)	4.16 ±0.11	**1.73** ±0.24	**1.25** ±0.10	**4.39** (+0.9%)
	Real	3.89 ±0.25	**1.94** ±0.16	1.57 ±0.20	4.13 (+7.3%)	N/A	N/A	N/A	N/A
Fine-tune	Synth	4.49 ±0.25	2.39 ±0.24	1.26 ±0.17	4.28	**4.19** ±0.08	1.85 ±0.25	1.28 ±0.10	4.35
	Real	3.73 ±0.30	2.40 ±0.22	1.77 ±0.28	3.85	N/A	N/A	N/A	N/A

Table 2: **Generalization to new activities.** Evaluation on 10 held-out activities per country in CULTIVate ($n = 160$, $\alpha = 0.7$). ↑ and ↓ indicate that higher and lower values are better, respectively.

Method	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑
Pre-trained	3.42	**2.35**	1.99	3.69
τ (Synth)	**3.70**	2.44	**1.79**	**3.82**
τ (Real)	3.51	2.42	2.05	3.68
Fine-tune (Synth)	3.62	2.78	1.88	3.65



Figure 2: Inference-time scaling controls cultural alignment and exaggeration.

These gains extend across Global North and Global South countries (Tab. 6 in Appendix), with synthetic τ matching or outperforming full fine-tuning in 15 of 16 countries (Tab. 8 in Appendix) and achieving the highest FAITH across all nine activity categories (Tab. 9 in Appendix). In contrast, a random vector with matched mean and variance yields negligible or negative changes in FAITH ($+0.02$ on CULTIVate and -0.11 on CulturalFrames), suggesting that the improvements depend on the learned cultural vector rather than arbitrary parameter perturbations.

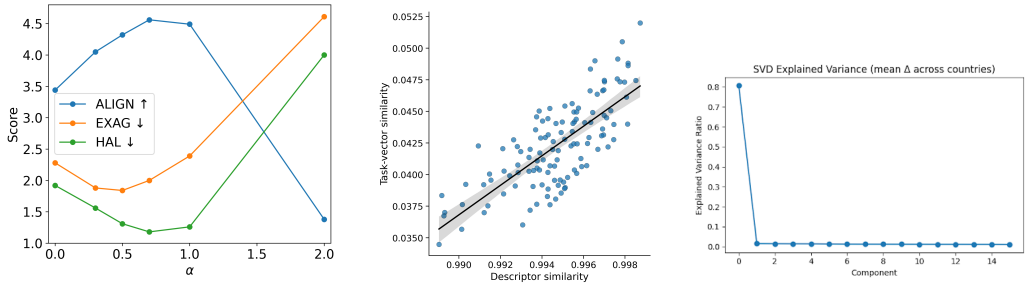
Generalization to unseen activities. We further evaluate whether cultural vectors generalize to unseen activities. We split CULTIVate into train and test sets by randomly selecting 10 activities per country as held-out set. As shown in Tab. 2, consistent with Tab. 1 (in-domain), cultural vectors learned from synthetic data achieve the best performance, outperforming the pretrained model, full fine-tuning, and cultural vectors learned from real data.

Advantages synthetic vs. real training data. Tab. 1 shows that cultural vectors learned from synthetic data outperform those learned from real data in alignment (ALIGN 4.56 vs. 3.89) and hallucination (HAL 1.18 vs. 1.57), while real τ achieves slightly lower exaggeration (EXAG 1.94 vs. 2.00). We further compare real and synthetic vectors through human evaluation (Tab. 16 in Appendix), where native annotators prefer images generated using synthetic τ in 62.1% of head-to-head comparisons. These findings demonstrate the effectiveness of synthetic data for learning cultural vectors, while increasing exaggeration which can be controlled by scaling α (see Sec. 5.2).

5.2 Properties of cultural vectors

Effect of α scaling. Fig. 3a shows that scaling α enables gradual control over cultural adaptation. Cultural alignment and faithfulness generally improve up to $\alpha = 0.7$, while larger values increase exaggeration and hallucination. Fig. 2 illustrates how increasing α can improve cultural relevance and introduce more appropriate cultural details than the pretrained model. However, excessive scaling (e.g., $\alpha = 2$) can produce severely exaggerated and incoherent images (not shown). Notably, scaling cultural vectors enables improving cultural faithfulness and exaggeration simultaneously.

Cultural vectors encode meaningful cultural knowledge. We measure pairwise cosine similarity between descriptor embeddings and cultural vectors across countries, using descriptor similarity (Sec. 3.3) as a proxy for cultural relatedness. Fig. 3b shows strong correlation between the two (Pearson $r = 0.75$, Spearman $\rho = 0.72$), confirming that cultural vectors capture culturally meaningful



(a) Ablation on α. (b) Pairwise similarity correlation. (c) Activation-space spectrum.

Figure 3: **Properties of cultural vectors.** (a) Scaling α controls cultural alignment, exaggeration, and hallucination. (b) Vector similarities correlate strongly with cultural descriptor similarities. (c) Activation shifts are dominated by a single principal direction.

Table 3: **Single-block training of cultural vectors.** Applying cultural vectors to different UNet blocks. Best result per up/down blocks **bolded**.

Model	Layer	HAL ↓	EXAG ↓	ALIGN ↑
τ	down-block-0	1.93	2.38	3.54
	down-block-1	1.77	2.30	3.67
	down-block-2	**1.49**	**2.06**	**4.13**
	down-block-3	1.89	2.43	3.50
	mid-block	1.69	2.37	3.83
	up-block-0	1.87	2.41	3.65
	up-block-1	**1.28**	**1.88**	**4.39**
	up-block-2	1.74	2.13	3.76
	up-block-3	1.88	2.24	3.54

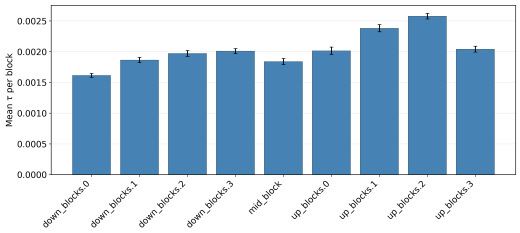


Figure 4: **Per-block shift magnitude.** Cultural vectors produce the largest parameter shifts in up-block-1 and up-block-2.

structure. Figs. 8b/ 8c (Appendix) further support this, with culturally proximate countries clustering together in both similarity matrices.

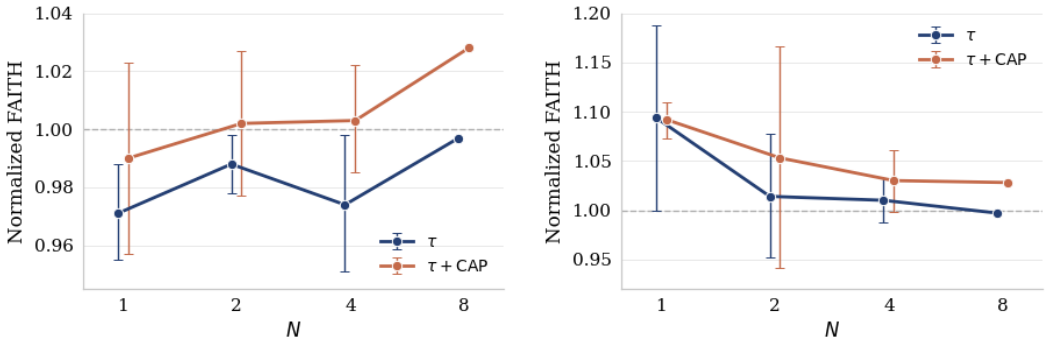
Shared activation-space structure across cultures. Fig. 3c shows that activation vectors across countries exhibit a low-dimensional structure, with an effective rank (see in Sec. 3.2) of 2.93 out of 16 possible directions. CAP increases the effective rank to 4.02, suggesting less concentrated activation shifts across cultures. These findings suggest a potential relationship between shared activation structure and cross-cultural interference. Detailed analysis is provided in Appendix Sec. C.4.

Block-wise application of cultural vectors. Fig. 4 shows that the largest parameter shifts occur in up-block-1 and up-block-2. Tab. 3 shows that up-block-1, a relatively deep decoder block, achieves the highest cultural alignment (ALIGN 4.39) and lowest hallucination (HAL 1.28), while maintaining low exaggeration (EXAG 1.88). In contrast, applying cultural vectors to shallower decoder blocks yields lower alignment and higher exaggeration, suggesting that the effectiveness of cultural adaptation depends on the U-Net block.

5.3 Compositionality of Cultural Vectors

Having established that cultural vectors capture meaningful cultural knowledge, we investigate their compositionality in three settings: (1) multi-cultural generation through merging vectors from multiple countries (Fig. 5); (2) generalization to unseen countries (Tab. 4); and (3) combining synthetic and real data to improve generation when no real target data is available (Tab. 15 in Appendix).

Naive merging reveals cross-cultural interference. In Fig. 5, we evaluate multi-cultural behavior and the compositionality of cultural vectors using Monte Carlo sampling over eight countries from CULTIVate (held-out set activities). For each $N \in \{1, 2, 4, 8\}$, we randomly sample three subsets of N countries (one for $N = 8$) and uniformly average their cultural vectors. We evaluate FAITH normalized by each country's individual full fine-tuning baseline, considering both the full country pool and the countries included in each merge. We report mean and stdev across sampled subsets.



(a) FAITH across all countries.

(b) FAITH on seen countries.

Figure 5: **Multi-cultural behavior of cultural vectors under composition.** We uniformly average N randomly selected cultural vectors and evaluate normalized FAITH on held-out activities. (a) FAITH across all countries generally increases as more vectors are merged. (b) FAITH on merged countries decreases with increasing N due to interference. CAP improves FAITH and compositionality in both settings. Results are averaged over three random subsets per N, except $N = 8$, which has one subset.

Table 4: **DGM improves generalization over naive merging.** DGM improves both TA/TIES. $\alpha = 0.7$; **Bold** denotes best within each method; ↑/↓: higher/lower is better.

Method	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑
Pre-trained	3.44	2.28	1.92	3.75
TA [13]	3.37	2.46	2.23	3.57
TA + DGM	**3.51**	**2.32**	**2.03**	**3.72**
TIES [39]	3.35	2.62	2.42	3.43
TIES + DGM	**3.61**	**2.19**	**1.97**	**3.82**

Table 5: **DGM Human evaluation.** Mean ranks from native annotators across both Global North (GN), Global South (GS) regions. Lower is better. TIES+DGM achieves the lowest mean ranks.

Method	GN	GS	All
Pre-trained	2.08	2.09	2.08
TIES	2.26	2.16	2.20
TIES+DGM	**1.66**	**1.75**	**1.72**

Fig. 5 shows that increasing N generally improves normalized FAITH across the full country pool (Fig. 5a), while reducing faithfulness on the merged countries (Fig. 5b), suggesting interference between cultural adaptations. Notably, CAP achieves better merging results than naive cultural vector composition, with performance improvements that persist as the number of merged countries increases, underscoring its effectiveness to improve robustness to cross-cultural interference.

Leave-one-out generalization. We evaluate whether cultural vectors from 15 source countries can be composed to generalize to a held-out target country on CULTIVate. To investigate our hypothesis that **culturally aware composition improves cross-cultural transfer**, we apply DGM to Task Arithmetic (TA) [13] and TIES [39].

Tab. 4 shows that naive merging degrades all four metrics relative to the pretrained model. In contrast, **DGM improves both merging baselines across all metrics**. TIES+DGM achieves the highest FAITH among merging methods (3.82 vs. 3.43 for TIES), exceeding the pretrained baseline (3.75), although its HAL remains slightly higher. Human evaluation (Tab. 5) further supports these gains: **TIES+DGM achieves the lowest mean rank** (1.72 vs. 2.08 for pretrained and 2.20 for TIES) and ranks first in 10 of 11 countries. These results suggest that descriptor-based cultural similarities provide useful guidance for effective merging.

Qualitative example. Fig. 6 compares τ and TIES+DGM against the pretrained model and TIES [40] merging. For "handshake in China," both τ and TIES+DGM produce more culturally grounded depictions than the baselines. For "dragon boat festival," TIES introduces irrelevant flowers, whereas τ and TIES+DGM better capture the boats associated with the event. Similarly, for "New Year's in the USA," TIES generates an unrelated wedding scene, while τ and TIES+DGM produce images more consistent with the prompt. These examples illustrate improvements in cultural faithfulness for the selected prompts, complementing the quantitative results.



Figure 6: **Qualitative Examples**. We compare Ties+DGM and τ against pre-trained and TIES.

6 Conclusion

In this work, we investigated how cultural knowledge in diffusion models can be represented, controlled, and composed through cultural vectors. We showed that cultural vectors enable controllable culturally grounded generation, with scaling revealing trade-offs between cultural alignment, exaggeration, and hallucination. Our analysis of data sources and internal representations further highlighted the potential of synthetic data for learning cultural representations and shared structure in the activation changes induced by cultural vectors. However, we found that naively merging cultural vectors can lead to cross-cultural interference, limiting their compositionality. To address these challenges, we introduced Cultural Activation Preservation (CAP), a regularization objective that preserves pretrained representations of non-target cultures, and Descriptor-Guided Merging (DGM), which leverages cultural similarities to improve generalization to unseen countries. Overall, our findings provide insights into how cultural knowledge is represented in diffusion models and offer practical approaches toward more controllable and culturally faithful cross-cultural generation.

7 Limitations

First, following prior work [26, 3, 33, 24], we use countries as proxies for cultures, which does not capture within-country diversity. Modeling such diversity requires data that sufficiently captures within cultural differences and expert knowledge, both of which can be difficult to obtain. Although our framework could be extended to finer-grained cultural representations through additional fine-tuning, we leave this direction for future work. Second, in this work we develop DGM as the simple yet first method to motivate the importance culture-aware compositional methods. DGM relies on descriptor-based cultural similarities; [24] has shown these descriptors can effectively capture cultural activities, but they are still prone to overlook important cultural distinction or reflect LLM biases. Nevertheless, descriptors are inexpensive to generate and require no target-country images, offering a scalable approach to extending cultural adaptation to additional countries and cultural groups. Finally, although descriptor-guided merging improves cross-cultural generalization, interference between cultural vectors persists. Developing more effective composition methods that preserve individual cultural representations remains an open challenge.

8 Acknowledgment

This work is supported by NSF Grant No. 2329992. We gratefully acknowledge the support of those who contributed to the human evaluation. We also thank Aysan Aghazadeh and Yufei Fiona Wang for their valuable comments and help throughout the project.

References

[1] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

[2] Abhipsa Basu, Mohana Singh, and Venkatesh Babu Radhakrishnan. Geodiv: Measuring concept diversity of images across geographical regions. In *CVPR 2025 Workshop Vision Language Models For All*, 2025.

[3] Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. Diffusion models through a global lens: Are they culturally inclusive? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31137–31155, 2025.

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[5] Benjamin Biggs, Arjun Seshadri, Yang Zou, Achin Jain, Aditya Golatkar, Yusheng Xie, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Diffusion soup: Model merging for text-to-image diffusion models. In *European Conference on Computer Vision*, pages 257–274. Springer, 2024.

[6] Kyle Buettner, Sina Malakouti, Xiang Lorraine Li, and Adriana Kovashka. Incorporating geo-diverse knowledge into prompting for increased geographical robustness in object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13515–13524, June 2024.

[7] Pietro Buzzega, Riccardo Salami, Angelo Porrello, and Simone Calderara. Rethinking layer-wise model merging through chain of merges. *arXiv preprint arXiv:2508.21421*, 2025.

[8] Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. Fusing finetuned models for better pretraining. *arXiv preprint arXiv:2204.03044*, 2022.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[10] Google DeepMind / Google. Nano banana (gemini 2.5 flash image), 2025. URL `https://ai.google.dev/gemini-api/docs/image-generation`. Image editing / generation model.

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[13] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6t0Kwf8-jrj`.

[14] Suchae Jeong, Inseong Choi, Youngsik Yun, and Jihie Kim. Culture-TRIP: Culturally-aware text-to-image generation with iterative prompt refinement. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9543–9573, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.483. URL `https://aclanthology.org/2025.naacl-long.483/`.

[15] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347, 2024.

[16] Yilei Jiang, Wei-Hong Li, Yiyuan Zhang, Minghong Cai, and Xiangyu Yue. Fairgen: Enhancing fairness in text-to-image diffusion models via self-discovering latent directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18411–18420, October 2025.

[17] Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 13716–13747. Curran Associates, Inc., 2024. doi: 10.52202/079017-0439.

[18] Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.573.

[19] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

[20] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[21] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12006–12016, June 2024.

[22] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. Scoft: Self-contrastive fine-tuning for equitable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10822–10832, June 2024.

[23] Qianli Ma, Xuefei Ning, Dongrui Liu, Li Niu, and Linfeng Zhang. Decouple-then-merge: Finetune diffusion models as multi-task learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23281–23291, 2025.

[24] Sina Malakouti, Boqing Gong, and Adriana Kovashka. Culture in action: Evaluating text-to-image models through social activities. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=opG4m2U0Oo`.

[25] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[26] Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd Van Steenkiste, Yash Goyal, Karolina Stańczak, and Aishwarya Agrawal. Culturalframes: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953, 2025.

[27] OpenAI. Gpt image 1, 2025. URL `https://platform.openai.com/docs/models/gpt-image-1`. Model documentation.

[28] Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754, 2023.

[29] Minh Pham, Kelly O Marshall, Chinmay Hegde, and Niv Cohen. Robust concept erasure using task vectors. *arXiv preprint arXiv:2404.03631*, 2024.

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[31] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15691–15701, October 2023.

[32] Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. Warm: On the benefits of weight averaged reward models, 2024. *URL https://arxiv.org/abs/2401.12187*.

[33] Aniket Rege, Zinnia Nie, Mahesh Ramesh, Unmesh Raskar, Zhuoran Yu, Aditya Kusupati, Yong Jae Lee, and Ramya Korlakai Vinayak. Cure: Cultural gaps in the long tail of text-to-image systems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15680–15691, 2025.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[35] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22522–22531, June 2023.

[36] George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. *ICLR*, 2025.

[37] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[38] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[39] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=xtaX3WyCj1`.

[40] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

[41] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

A Table of Contents

- **Additional Method Details**
  - Cultural descriptor generation
  - Cross-Cultural Vector Analogy (CVA)
- **Additional Results**
  - **Learning and Controlling Cultural Vectors**
    - Cultural faithfulness across countries
    - Cultural faithfulness across activities
    - Effect of cultural vector scaling
  - **Sources and Properties of Cultural Vectors**
    - Effect of cultural vector scaling
    - Parameter-space SVD analysis
    - Cultural similarity analysis
    - Activation-space CKA analysis
  - **Ablations on Learning Cultural Vectors**
    - Impact of synthetic training data diversity
    - Conditional vs. unconditional training
  - **Cross-Cultural Generalization**
    - Progressive cultural vector merging
    - CAP and output-preservation ablations
    - CAP activation-space analysis
    - Generalization to unseen countries with DGM
    - Cross-cultural vector analogy
  - **Human Evaluation**
    - DGM versus pretrained and TIES merging
    - Real versus synthetic cultural vectors
- **Broader Impact**
- **Evaluation Prompts**
  - Cultural alignment (ALIGN)
  - Cultural exaggeration (EXAG)
  - Cultural hallucination (HAL)

B Additional Method Details

Descriptor Generation. Prior work [31, 25] has shown that LLM-generated visual descriptors capture object appearances for zero-shot classification and characterize visual differences across geographic regions [6, 2]. Following [24], we use LLM-generated cultural descriptors across five dimensions: attire, background, objects, interactions, and spatial layout. Descriptors are generated by multiple LLMs and consolidated by a separate LLM to filter inconsistent or incorrect entries. The resulting descriptor sets $\mathcal{T}_c$ are used to compute cultural similarities for Descriptor-Guided Merging (DGM), as described in Sec. 3.3.

Cross-Cultural Vector Analogy (CVA). While synthetic data enables scalable learning of cultural vectors, it may introduce biases inherited from the underlying text-to-image models. To investigate whether differences between real and synthetic cultural vectors can be transferred across countries, we model this difference as an offset in parameter space.

Let $\mathcal{S} \subset \mathcal{C}$ denote the source countries with both real and synthetic training data. We compute the average real-to-synthetic offset across source countries and apply it to the synthetic vector of a target country t:

$$\hat{\boldsymbol{\tau}}^t_{\text{real}} = \boldsymbol{\tau}^t_{\text{synth}} + \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(\boldsymbol{\tau}^s_{\text{real}} - \boldsymbol{\tau}^s_{\text{synth}} \right). \tag{12}$$

We report the results in Tab. 15.

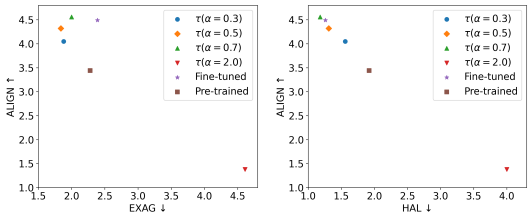
Table 6: **Controlling cultural adaptation via cultural vectors on CULTIVate.** We vary the scaling coefficient α, and all vectors are applied to the full model. ↑/↓ indicate higher/lower is better. Results are mean ± standard deviation (GPT-4o-as-a-judge). Region's average is sample weighted.

Method	Global North				Global South				All			
	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑
Pre-trained	3.67 ±0.21	1.98 ±0.26	1.64 ±0.26	4.02	3.33 ±0.33	2.42 ±0.23	2.04 ±0.39	3.62	3.44 ±0.33	2.28 ±0.31	1.92 ±0.39	3.75
τ ($\alpha = 0.3$)	4.17 ±0.29	1.74 ±0.17	1.50 ±0.19	4.31	3.99 ±0.35	1.95 ±0.24	1.59 ±0.34	4.15	4.05 ±0.33	1.88 ±0.23	1.56 ±0.30	4.20
τ ($\alpha = 0.5$)	4.39 ±0.18	**1.73** ±0.18	1.29 ±0.14	4.46	4.29 ±0.34	**1.89** ±0.21	1.31 ±0.24	4.36	4.32 ±0.30	**1.84** ±0.21	1.31 ±0.21	4.39
τ ($\alpha = 0.7$)	**4.55** ±0.16	1.83 ±0.19	**1.20** ±0.10	**4.51**	**4.56** ±0.25	2.07 ±0.25	**1.17** ±0.10	**4.44**	**4.56** ±0.22	2.00 ±0.26	**1.18** ±0.10	**4.46**
τ ($\alpha = 2.0$)	1.28 ±0.19	4.66 ±0.13	4.09 ±0.48	1.51	1.43 ±0.18	4.58 ±0.17	3.96 ±0.25	1.63	1.38 ±0.19	4.61 ±0.16	4.00 ±0.33	1.59
Fine-tuned	4.50 ±0.10	2.30 ±0.15	1.29 ±0.11	4.30	4.48 ±0.30	2.43 ±0.26	1.25 ±0.19	4.27	4.49 ±0.25	2.39 ±0.24	1.26 ±0.17	4.28

Table 7: **Effect of cultural vector scaling on CulturalFrame.** We report ALIGN, EXAG, HAL, and FAITH across different scaling coefficients α. Results are mean ± standard deviation (GPT-4o-as-a-judge).

Method	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑
Pre-trained	3.91 ±0.15	1.75 ±0.21	1.38 ±0.15	4.26
$\boldsymbol{\tau}$ ($\alpha = 0.3$)	4.05 ±0.11	**1.71** ±0.22	1.33 ±0.13	4.34
$\boldsymbol{\tau}$ ($\alpha = 0.5$)	4.15 ±0.14	1.73 ±0.22	1.29 ±0.11	4.38
$\boldsymbol{\tau}$ ($\alpha = 0.7$)	4.16 ±0.11	1.73 ±0.24	**1.25** ±0.10	**4.39**
$\boldsymbol{\tau}$ ($\alpha = 2.0$)	4.01 ±0.13	2.54 ±0.25	1.36 ±0.17	4.04
Fine-tuned	**4.19** ±0.08	1.85 ±0.25	1.28 ±0.10	4.35

Figure 7: **Scaling trade-offs.** Cultural alignment versus exaggeration and hallucination under different scaling coefficients α.



(a) ALIGN vs. EXAG (b) ALIGN vs. HAL

C Additional Results

C.1 Learning and Controlling Cultural Vectors

Performance across Global South and Global North countries. As shown in Tab. 6, cultural vectors with $\alpha = 0.7$ improve cultural alignment (ALIGN) over full fine-tuning across both Global South (GS) and Global North (GN) countries, while reducing cultural exaggeration (EXAG). These results suggest that scaling cultural vectors provides a more favorable balance between alignment and exaggeration than full fine-tuning.

We further examine country-level cultural faithfulness (FAITH) in Tab. 8. Cultural vectors outperform or match full fine-tuning in 15 of 16 countries, with average relative improvements of 4.0% and 4.6% across GS and GN countries, respectively.

Performance across cultural activities. To examine whether the improvements extend across different cultural scenarios, we compare cultural faithfulness across nine activity categories. As shown in Tab. 9, cultural vectors outperform full fine-tuning in all nine categories, with the largest relative improvement for eating (+8.5%). These results indicate that the improvements are not restricted to particular countries or activity categories.

C.2 Sources and Properties of Cultural Vectors

Impact of scaling α. Tabs. 6 and 7 examine the effect of varying the scaling coefficient. Smaller values of α generally yield lower cultural alignment and exaggeration, while larger values strengthen cultural adaptation. However, excessive scaling (e.g., $\alpha = 2$) can degrade image quality.

Figs. 7a and 7b further illustrate that increasing α tends to improve cultural alignment while also increasing exaggeration and hallucination. Intermediate scaling values provide a more favorable balance between these objectives.

Cultural vectors occupy independent directions in parameter space. As shown in Fig. 8a, the explained variance is distributed nearly uniformly across singular values for up-block-1, yielding effective ranks of 15.83 out of a maximum of 16. This near full-rank structure indicates that cultural vectors span largely independent directions in parameter space, rather than concentrating in a shared low-dimensional subspace.

Table 8: **Cultural faithfulness across Global South (GS) and Global North (GN) countries.** FAITH scores for the pretrained model, full fine-tuning (Full-FT), and cultural vectors ($\alpha = 0.7$). Relative changes over Full-FT are shown in parentheses. Green and red indicate improvements and declines, respectively. Best scores are in **bold**.

Global South (GS)

Country	Pretrained	Full-FT	τ
Brazil	3.60	**4.56**	**4.56** (0.0%)
China	3.97	4.27	**4.43** (+3.7%)
India	4.11	**4.44**	**4.44** (0.0%)
Indonesia	3.48	4.22	**4.56** (+8.1%)
Iran	3.15	3.81	**4.11** (+7.9%)
Mexico	3.87	4.31	**4.62** (+7.2%)
Nepal	3.89	4.29	**4.38** (+2.1%)
Nigeria	3.74	**4.31**	4.27 (-0.9%)
Philippines	3.45	4.56	**4.68** (+2.6%)
South Africa	3.59	4.18	**4.37** (+4.5%)
Turkey	3.30	4.01	**4.40** (+9.7%)
GS Average	3.65	4.27	**4.44** (+4.0%)

Global North (GN)

Country	Pretrained	Full-FT	τ
France	4.17	4.38	**4.40** (+0.5%)
Germany	3.94	4.25	**4.57** (+7.5%)
Italy	4.12	4.27	**4.54** (+6.3%)
Spain	3.64	4.22	**4.48** (+6.2%)
USA	4.23	4.42	**4.56** (+3.2%)
GN Average	4.02	4.31	**4.51** (+4.6%)

Table 9: **Cultural faithfulness across activity categories.** We report FAITH scores for the pretrained model, full fine-tuning (Full-FT), and cultural vectors ($\alpha = 0.7$). Relative improvements over Full-FT are shown in parentheses, with green indicating improvements and red indicating declines. Best scores are in bold.

Model	Celebration	Concert	Dancing	Eating	Funeral	Game	Greeting	Religious	Wedding
Pretrained	3.82	4.39	3.74	3.91	4.24	2.89	4.19	4.03	4.19
Full-FT	4.28	4.27	4.44	4.12	4.07	4.22	4.30	4.27	4.26
τ ($\alpha = 0.7$)	**4.47** (+4.4%)	**4.44** (+4.0%)	**4.50** (+1.4%)	**4.47** (+8.5%)	**4.25** (+4.4%)	**4.44** (+5.2%)	**4.49** (+4.4%)	**4.43** (+3.7%)	**4.43** (+4.0%)

Cultural similarity. Fig. 3b shows a positive correlation between pairwise cultural vector similarities and descriptor-based cultural similarities. The corresponding similarity matrices (Figs. 8b and 8c) exhibit related patterns across countries. These observations suggest that parameter-space similarities reflect some of the relationships captured by cultural descriptors, although descriptor similarity is only a proxy for cultural relatedness.

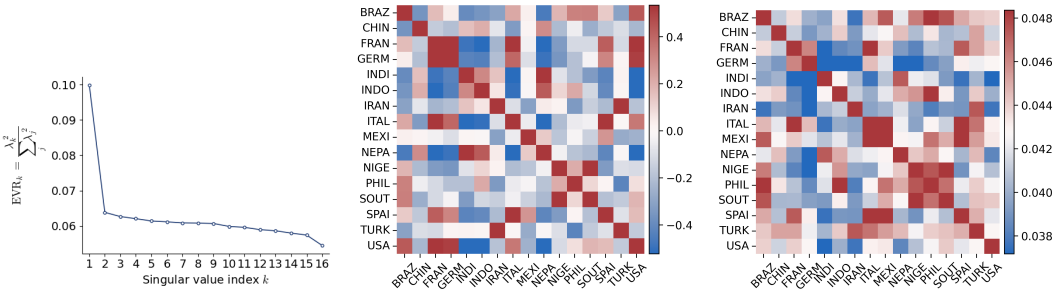
CKA analysis. Following [36], we apply centered kernel alignment (CKA) [19] to compare activations across countries (Figs. 9a, 9b, and 9c). We observe high representation similarity across the analyzed countries, including pairs with different cultural descriptors. The high similarities indicate shared activation structure across vectors.

C.3 Ablations on Learning Cultural Vectors

Impact of the number of T2I models used to learn cultural vectors. In Tab. 10, we investigate the effect of increasing the number of T2I models used to generate synthetic training data. Following the experimental setup in Sec. 4, we additionally include synthetic data generated by three public T2I models (Stable Diffusion 3 Medium, FLUX.1, and Qwen-Image), increasing the number of source models from three to six.

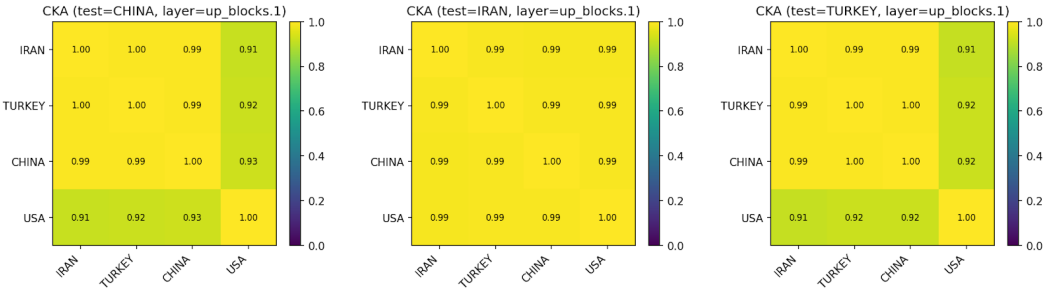
We observe that in both settings (i.e., 3 and 6 T2I model), cultural vectors significantly outperform pre-trained model while improve full finetuned baseline. This further confirms the effectiveness of cultural vectors in steering the generation of models towards more faithful cultural images. However, we observe including more models' and increasing the training data size, performance degrades compared with three proprietary T2I models as training data sources. We hypothesize that this is due to additional noise and more stereotypical or culturally inaccurate images by public models compared to strong proprietary models.

Impact of training strategy. Tab. 11 shows that cultural vectors outperform the pretrained baseline under both conditional and unconditional training. Conditional training yields stronger absolute performance (ALIGN 4.56 vs. 3.60), while the relative gain of cultural vectors over full fine-tuning is larger under unconditional training.



(a) Independence of cultural vectors (b) Descriptor similarity (c) Cultural vector similarity

Figure 8: **Cultural vectors reflect cultural similarity.** (a) Cultural vectors occupy largely independent directions in parameter space, as shown by the distributed singular value spectrum. (b,c) Descriptor and cultural vector similarity matrices, scaled to the 5-95th percentiles to enhance contrast.



(a) CKA China (b) CKA on Iran (c) CKA on Turkey

Figure 9: CKA [19] analysis.

C.4 Cross-Cultural Generalization

Composition of cultural vectors. Tab. 12 examines the effect of progressively merging cultural vectors. The results show that combining additional vectors can introduce trade-offs between performance on different countries. For instance, we observe that adding $\boldsymbol{\tau}_{iran}$ increases FAITH for Iran and Turkey while degrading performance on the USA and China. This indicates that the cross-cultural interference may be related to cultural similarities between countries. Additionally, the high similarity we observed in Fig. 3c may help explain the interference among cultural vectors, which we improve by CAP.

Cultural Activation Preservation (CAP). To address interference during composition, CAP regularizes culture-specific fine-tuning by penalizing changes in the pretrained model's activations on non-target cultural samples. As discussed in Sec. 5.3, CAP improves cultural faithfulness under vector composition. These results suggest that preserving non-target activations can mitigate some of the interference introduced by merging, although it does not eliminate the compositionality challenge. Additionally, we observe that CAP increases the effective rank of the stacked activation matrix (Sec. 3) across three different sources: real images (Fig. 10a), private models that were used to train the model (Fig. 10b), and public T2I models that were unseen during training (Fig. 10c)

Furthermore, Tab. 13 presents a four-country ablation involving Iran, the USA, China, and Turkey. Following the progressive merging setting in Tab. 12, we progressively combine cultural vectors from Iran, the USA, and China, evaluating ALIGN across all four countries, including Turkey, whose vector is not included in the merge. Naive merging does not consistently improve ALIGN over the pretrained model, particularly when all three vectors are combined. We compare Cultural Activation Preservation (CAP) with output preservation, which regularizes the U-Net's predicted noise to remain close to that of the pretrained model on non-target cultures. CAP generally achieves higher ALIGN than output preservation across the evaluated configurations. In particular, $\lambda_{\text{CAP}} = 0.1$ achieves the highest scores in two of the three merging configurations, supporting the effectiveness of activation preservation in reducing cross-cultural interference.

Table 10: **Effect of training data diversity on cultural vectors.** Comparison of cultural vectors ($\boldsymbol{\tau}$) and full fine-tuning using synthetic data from different T2I models. Values denote mean scores.

Training Data	Model	α	HAL ↓	EXAG ↓	ALIGN ↑
Pre-trained	–	–	1.92	2.28	3.44
3 T2I Models (Private)	$\boldsymbol{\tau}$	0.7	1.18	2.00	4.56
	Fine-Tuned	–	1.26	2.39	4.49
6 T2I Models (Private + Public)	$\boldsymbol{\tau}$	0.7	1.40	2.18	4.26
	Fine-Tuned	–	1.39	2.54	4.24

Table 11: **Conditional vs. unconditional cultural adaptation.** Comparison of cultural vectors ($\boldsymbol{\tau}$) and full fine-tuning under conditional and unconditional training. Values denote mean scores.

Training	Model	α	HAL ↓	EXAG ↓	ALIGN ↑
Pre-trained	–	–	1.92	2.28	3.44
Conditional	$\boldsymbol{\tau}$	0.7	1.18	2.00	4.56
	Fine-Tuned	–	1.26	2.39	4.49
Unconditional	$\boldsymbol{\tau}$	0.7	1.80	2.64	3.60
	Fine-Tuned	–	2.39	3.46	2.99

Table 12: **Effect of progressively adding cultural vectors.** Values show FAITH. Values in parentheses indicate change relative to the pretrained model (improvement, degradation). Avg is average. τ_{IR}, τ_{US}, and τ_{CN} denote cultural vectors of Iran, United States, and China, respectively.

Cultural Vectors Added	China	Iran	Turkey	USA
Pre-trained	3.79	3.31	2.64	3.92
$+\,\tau_{IR}$	2.80 (-0.99)	3.89 (+0.58)	3.12 (+0.48)	2.48 (-1.44)
$+\,Avg(\tau_{IR}, \tau_{US})$	3.88 (+0.09)	3.73 (+0.42)	3.16 (+0.52)	4.23 (+0.31)
$+\,Avg(\tau_{IR}, \tau_{US}, \tau_{CN})$	3.82 (+0.03)	3.45 (+0.14)	2.78 (+0.14)	4.30 (+0.38)

Descriptor-Guided Merging (DGM) across regions. Tab. 14 provides the full results of leave-one-country-out generalization on CULTIVate, separately reporting performance for Global North (GN) and Global South (GS) countries. Naive Task Arithmetic (TA) and TIES merging generally degrade performance relative to the pretrained model, whereas Descriptor-Guided Merging (DGM) improves both methods across all metrics in both regions. In particular, TIES+DGM achieves FAITH scores of 4.06 and 3.71 for GN and GS, respectively, while also reducing exaggeration and hallucination. Across all countries, TIES+DGM achieves the highest FAITH among merging methods (3.82), exceeding the pretrained baseline (3.75), although direct cultural vector adaptation achieves higher faithfulness (4.46). These results suggest that descriptor-based weighting improves the composition of existing cultural vectors for unseen countries.

Global North to Global South generalization. We evaluate whether cultural vectors can generalize to Global South countries using only Global North real data, simulating a setting where real data is available for higher-resource regions but limited for underrepresented ones. We assume access to synthetic data for all countries but restrict real training data to Global North countries.

We compare cultural vectors learned from synthetic target data with Cross-Cultural Vector Analogy (CVA), which adjusts these vectors using a real-to-synthetic offset transferred from Global North source countries.

Tab. 15 shows that cultural vectors learned from synthetic Global South data substantially improve cultural faithfulness over the pretrained model (FAITH 4.44 vs. 3.62). Applying Cross-Cultural Vector Analogy (CVA) further reduces cultural exaggeration (EXAG 1.85 vs. 2.07) while preserving comparable faithfulness (FAITH 4.43 vs. 4.44). These results suggest that transferring real-to-synthetic offsets across cultures can help mitigate exaggeration without compromising cultural faithfulness.

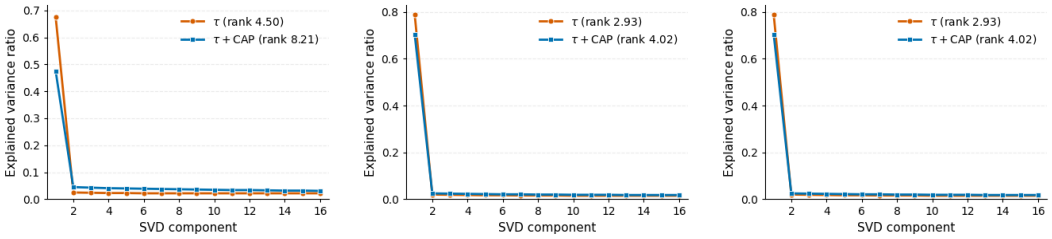
C.5 Human Evaluation

To complement automatic evaluation, we conduct human studies to assess the cultural faithfulness of generated images. We recruit native annotators through Prolific, screening participants by nationality to match the evaluated countries. We randomize image presentation and include attention checks for quality control. Annotators are compensated at a rate exceeding 8 USD per hour, with approximately two annotations collected per prompt.

Real vs synthetic cultural vectors. We conduct a head-to-head comparison of cultural vectors learned from real and synthetic data across 10 countries, including three Global North countries (USA, Germany, and France) and seven Global South countries (Turkey, China, India, Nigeria, Brazil, Mexico, and Iran). Annotators select the more culturally faithful image for each prompt. As shown in Tab. 16, synthetic-data vectors are preferred in 62.1% of judgments overall, including 64.9% in

Table 13: **Ablation of preservation objectives and regularization strength.** We compare CAP and output preservation across different λ values and cultural vector compositions. ALIGN is averaged over four countries (Iran, USA, China, and Turkey). Multiple vectors are uniformly averaged. Parentheses indicate changes from pretrained (improvement, degradation). Best scores are in **bold**.

Method	λ	τ^{IR}	$\text{Avg}(\tau^{IR}, \tau^{US})$	$\text{Avg}(\tau^{IR}, \tau^{US}, \tau^{CN})$
Pretrained	–	3.42	3.42	3.42
τ (No Preservation)	–	2.87 (-0.55)	3.50 (+0.08)	3.41 (-0.01)
Output Preservation	0.01	3.04 (-0.38)	3.40 (-0.02)	3.52 (+0.10)
	0.1	3.12 (-0.30)	3.36 (-0.06)	3.39 (-0.03)
	1.0	3.53 (+0.11)	3.66 (+0.24)	3.80 (+0.38)
CAP	0.01	3.68 (+0.26)	3.69 (+0.27)	3.65 (+0.23)
	0.1	**3.73** (+0.31)	3.75 (+0.33)	**3.85** (+0.43)
	1.0	3.60 (+0.18)	**3.89** (+0.47)	3.77 (+0.35)



(a) real images

(b) Synthetic data from private models (train data)

(c) Synthetic data from public T2I

Figure 10: **Activation-space Spectrum.** SVD analysis (at 400 time-step) show that CAP increases the rank of the activations across different images: real (unseen), private T2I (trained data), and public T2I (unseen).

Global North and 61.0% in Global South countries. At the prompt level, synthetic data vectors receive the majority preference for 138 of 274 prompts (50.4%), compared with 68 prompts (24.8%) for real data vectors, while the remaining 68 prompts (24.8%) are tied. These results provide further human evidence supporting the effectiveness of synthetic data for learning cultural vectors.

Cross-cultural generalization. We evaluate Descriptor-Guided Merging (TIES+DGM) against standard TIES [40] and the pretrained model by asking annotators to rank images based on cultural faithfulness. The study covers 11 countries across Global North (France, Germany, Spain, USA) and Global South (Brazil, China, India, Iran, Mexico, Nigeria, Turkey), comprising 301 prompts. As shown in Tab. 5 (main paper), TIES+DGM achieves the lowest mean rank overall (1.72), outperforming both the pretrained model (2.08) and TIES (2.20). This improvement holds across Global North (1.66 vs. 2.08 and 2.26) and Global South (1.75 vs. 2.09 and 2.16), with TIES+DGM achieving the best mean rank in 10 of 11 countries. Pairwise Wilcoxon signed-rank tests with Holm correction confirm significant improvements over the pretrained model ($p = 2.6 \times 10^{-7}$) and TIES ($p = 5.4 \times 10^{-13}$). Inter-annotator agreement is also substantial (Kendall's $W = 0.71$).

D Broader Impact

This work studies culturally grounded generation in text-to-image models, particularly for underrepresented cultures that may be inaccurately or stereotypically represented in current systems. We hope our findings contribute to more inclusive and controllable generative models and support future research on culturally aware generation.

At the same time, cultural representations learned from country-level datasets may oversimplify the diversity within cultures or reinforce biases present in training images and cultural descriptors. Scaling cultural vectors may also amplify stereotypes or introduce unsupported cultural elements. Therefore, improvements in cultural faithfulness should not be interpreted as eliminating bias or ensuring accurate representation of all communities within a country.

Table 14: **Descriptor-Guided Merging (DGM) improves cross-cultural generalization in the leave-one-out setting.** Source cultural vectors are merged to generate images for a held-out target culture with $\alpha = 0.7$ on CULTIVate (16 countries). Direct-adaptation methods use target-country training data and are included only as references. Bold indicates the best result within each method group; ↑/↓ indicate higher/lower is better.

Group	Method	Global North				Global South				All			
		ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑	ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑
-	Pre-trained	3.67	**1.98**	**1.64**	4.02	3.33	2.42	2.04	3.62	3.44	2.28	1.92	3.75
Merging-based methods	TA [13]	3.59	2.19	2.01	3.80	3.27	2.59	2.33	3.45	3.37	2.46	2.23	3.57
	TA + DGM	**3.64**	**2.10**	**1.88**	**3.89**	**3.44**	**2.41**	**2.10**	**3.64**	**3.51**	**2.32**	**2.03**	**3.72**
	TIES [39]	3.52	2.45	2.29	3.59	3.27	2.69	2.48	3.37	3.35	2.62	2.42	3.43
	TIES + DGM	**3.87**	**1.96**	**1.74**	**4.06**	**3.50**	**2.30**	**2.07**	**3.71**	**3.61**	**2.19**	**1.97**	**3.82**
Direct adaptation	$\boldsymbol{\tau}$	**4.55**	1.83	**1.20**	**4.51**	**4.56**	2.07	**1.17**	**4.44**	**4.56**	2.00	**1.18**	**4.46**
	Fine-tune	4.50	2.30	1.29	4.30	4.48	2.43	1.25	4.27	4.49	2.39	1.26	4.28

Table 15: **Global North to Global South generalization.** ↑/↓ indicate higher/lower is better.

Method	Training data	Global South			
		ALIGN ↑	EXAG ↓	HAL ↓	FAITH ↑
Pre-trained	–	3.33	2.42	2.04	3.62
$\boldsymbol{\tau}$ (Ours)	Synth (GS)	4.56	2.07	1.17	**4.44**
CVA (Ours)	Synth (GS) + Real (GN)	4.42	**1.85**	1.28	4.43
Fine-tuned	Synth (GS)	4.48	2.43	1.25	4.27

Table 16: **Human preference for real vs. synthetic cultural vectors.** Pairwise preferences from native annotators across 10 countries (%). Best results are in **bold**.

Region	Synthetic Preferred	Real Preferred
Global North (GN)	**64.9%**	35.1%
Global South (GS)	**61.0%**	39.0%
Overall	**62.1%**	37.9%

E Evaluation Prompts

This section provides the prompts used to evaluate cultural alignment, exaggeration, and hallucination.

Table 17: ALIGN MLLM-as-a-Judge Prompt

ALIGN Likert Scoring Prompt

System: You are a helpful assistant scoring how well an image depicts a prompt as it would typically appear in the specified country. Provide a rating from 1 to 5 using the guidelines below. Only output the score in the required format.

Scoring Scale:
1 = Not well at all (completely wrong or unrelated)
2 = Slightly relevant
3 = Somewhat accurate
4 = Mostly accurate
5 = Exactly as expected for this country

Task:
PROMPT: `{{prompt}}`
COUNTRY: `{{country}}`

User: Respond with ONLY the score in the exact format below:
score is $< num >$

Do not include any other text, explanation, or formatting.

Table 18: EXAG MLLM-as-a-Judge Prompt

EXAG MLLM-as-a-Judge Prompt

System: You are a helpful assistant scoring how exaggerated an image is in terms of cultural elements (e.g., attire, objects, background, art/design). Rate on a scale from 1 to 5 using the rules below. Only output the score in the required format.

Scoring Scale:
1 = Appropriate and balanced.
2 = Slightly decorative but within cultural norms.
3 = Several features feel idealized or overly emphasized.
4 = Many elements feel exaggerated and unrealistic.
5 = Strongly stereotypical, misleading, or culturally irrelevant.

Task:
PROMPT: `{{prompt}}`
COUNTRY: `{{country}}`

User: Respond with ONLY the score in the exact format below:
score is $< num >$

Do not include any other text, explanation, or formatting.

Table 19: HAL MLLM-as-a-Judge prompt

HAL MLLM-as-a-Judge Prompt

System: You are a helpful assistant scoring whether an image contains incorrect elements with respect to the mentioned activity and country. Rate on a scale from 1 to 5 using the rules below. Only output the score in the required format.

Scoring Scale:
1 = Not at all. All elements seem appropriate.
2 = Slightly. One minor element seems foreign.
3 = Somewhat. A few elements feel unusual.
4 = Mostly. Many elements appear incorrect or from another culture.
5 = Strongly. The image is completely wrong, most elements are wrong, the scene is wrong, or belongs to another country.

Task:
PROMPT: `{{prompt}}`
COUNTRY: `{{country}}`

User: Respond with ONLY the score in the exact format below:
score is $< num >$

Do not include any other text, explanation, or formatting.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper claims to study whether culturally grounded knowledge can be represented and controlled through cultural vectors in diffusion models, and introduces Descriptor-Guided Generalization (DGM) for cultural generation . These claims are directly supported through analyses of controllability, compositionality, and cross-cultural generalization presented in Sec.3 (includes explanations), Sec. 5 (includes results), and additional results in Appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Sec. 7. Additionally, training doesn't have any additional overhead. Also while using LLMs that are not fully deterministic for evaluation (as they are shown to be strongest evaluators in this domain) we have followed various steps to ensure validity and reliability of the results (e.g., setting random seed, temp to zero and averaging multipel runs – see Sec. 4)

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: No theoretical results is included in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes all implementation details are included in Sec. 4. Prompts are provided in Sec. E, and method has been clearly describe in Sec. 3. Also code will be made public upon acceptance and publicly available datasets from prior works are utilized.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The work has utilized publicly available datasets (e.g., prompts/images) such as [26, 24]. We have provided all the prompts in Sec. E, and described the method in Sec. 3. While we will make the code available upon acceptance, we expect reproducibility as all training details and settings are provided in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the details are described in Sec.4 also models available on huggingface are utilized.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Some of the main experiments/tables include standard deviation/variance (e.g.,Table 1 also human eval shows significant p-value.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For GPU and training time please refer to Sec.4.Additional note: the paper doesn't modify the training of the baseline diffusion model, hence its training/inference doesn't exceed the baseline (i.e., stable diffusion v1).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper does not include any sensitive or private information as it utilizes public datasets. the submission is anonymized, and the paper discuss limitations and reproducibility details. Additionally, for human study Prolific is utilized with fair compensation of more than 8 USD.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader Impact is discussed in Appendix Sec. D

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not release new generative models or scraped datasets, and instead studies cultural representations using existing pretrained diffusion models and publicly available datasets. Therefore, we do not foresee additional misuse risks beyond those already associated with existing text-to-image models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: The public datasets and models used in the paper are properly discussed and cited 4.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This work does not introduce or release new datasets, models, or code assets as part of the submission. All methodological, training, and experimental details are provided in Sec. 4

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Yes, all the details are provided in the Appendix (e.g. exmaple of the study), and the details.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: risks were minimal and study was exempt

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM has been used to for writing purposes. The LLM usage in creating descriptors and evaluation follows prior works and all setting are provided in Sec. 4. LLMs were not used to propose ideas or in ways that will impact scientific rigor or originality

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.