Beyond Translation: Cross-Cultural Meme Transcreation with Vision-Language Models

Anonymous ACL submission

Abstract

Memes are a pervasive form of online communication, yet their cultural specificity poses significant challenges for cross-cultural adaptation. We study *cross-cultural meme transcreation*, a multimodal generation task that aims to preserve communicative intent and humor while adapting culture-specific references. We propose a hybrid transcreation framework based on vision–language models and introduce a large-scale bidirectional dataset of Chinese and US memes. Using both human judgments and automated evaluation, we analyze 6,315 meme pairs and assess transcreation quality across cultural directions. Our results show that current vision–language models can perform cross-cultural meme transcreation to a limited extent, but exhibit clear directional asymmetries: US→Chinese transcreation consistently achieves higher quality than Chinese→US. We further identify which aspects of humor and visual–textual design transfer across cultures and which remain challenging, and propose an evaluation framework for assessing cross-cultural multimodal generation. Our code and dataset are publicly available at https://anonymous.4open.science/r/MemeXGen/.

1 Introduction

Memes are a dominant form of online communication, yet they are difficult to adapt across cultural contexts. A meme that resonates with US audiences may fail among Chinese users—not due to linguistic errors, but because its humor, symbolism, or visual style does not translate culturally. While literal translation preserves surface meaning, it often fails to preserve what makes a meme effective: its intent, humor, and cultural resonance.

This challenge is often described as *transcreation* (Khanuja et al., 2024b): the process of adapting content across cultures by preserving communicative intent rather than literal form. For memes, transcreation requires coordinated adaptation of both text and images, grounded in culturally specific norms, references, and aesthetic preferences. As such, *meme transcreation poses a fundamental multimodal and cultural generation challenge that goes beyond standard translation or captioning*.

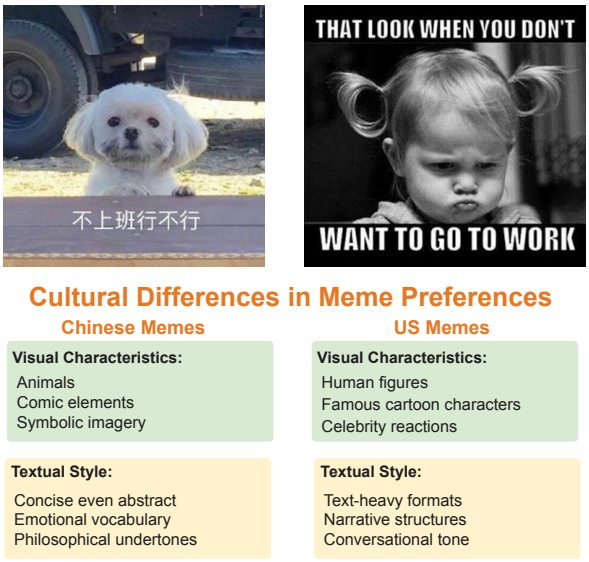


Figure 1: Examples of cultural differences in meme preferences across Chinese and US contexts. Cultural preferences shape humor and visual style, creating challenges for cross-cultural meme transcreation.

Aspect	US	Chinese
Visual	Human, celebrity	Animal, symbolic
Text	Narrative, detailed	Concise, philosophical
Emotion	Situational	Universal
Humor	Sarcasm, relatable	Wordplay, cuteness

Table 1: Cross-Cultural Meme Characteristics

Prior work has primarily studied memes from a recognition and analysis perspective (Hazman et al., 2025; Cao et al., 2023; Zhao et al., 2025). In parallel, recent vision–language models demonstrate strong multimodal understanding capabilities. However, systematic frameworks, datasets, and evaluations for *cross-cultural meme generation* remain limited. In particular, it is unclear how well current models can perform culturally grounded

transcreation, how performance varies across cultural directions, and how such systems should be evaluated.

To address this gap, we empirically study cross-cultural meme transcreation between Chinese and US cultures through a hybrid framework that adapts memes while preserving communicative intent. We introduce a large-scale bidirectional dataset of original and transcreated memes and evaluate transcreation quality using both human judgments and automated evaluation. Our analysis highlights systematic directional effects and cultural factors that shape the success and limitations of current vision–language models in cross-cultural meme transcreation. Table 1 summarizes key cross-cultural distinctions considered in this work.

This paper addresses three research questions:

RQ1: How effectively can vision–language models perform cross-cultural meme transcreation while preserving intent, humor, and cultural nuance?

RQ2: Does transcreation direction introduce systematic asymmetries between US→Chinese and Chinese→US adaptations?

RQ3: How should multimodal meme transcreation be evaluated, and how do human judgments compare with automated evaluations?

Our Contributions. **First, we propose a hybrid framework for meme transcreation** that balances intent preservation with cultural adaptation, offering practical guidance for culturally grounded meme adaptation. **Second, we introduce the first bidirectional meme transcreation dataset**, containing 6,315 original memes and 6,315 corresponding transcreated memes across both Chinese→US and US→Chinese directions. **Third, we present a bidirectional empirical study of Chinese and US meme transcreation**, showing that transcreation quality depends on direction and that specific aspects of internet humor—such as imagery, text style, and emotional expression—transfer unevenly across cultures.

2 Related Work

Cultural Gaps in AI. Despite advances in large language models and vision–language models (VLMs), cultural gaps remain a persistent challenge (Mihalcea et al., 2025; Adilazuarda et al., 2024). Prior work shows that NLP systems often fail to account for cross-cultural variation (Hershcovich et al., 2022), while text-to-image models tend to default to Western-centric representations (Kannen et al., 2024). These biases manifest as systematic performance disparities across languages and cultures (Field et al., 2021; Naous et al., 2023). Recent benchmarks such as GlobalRG (Bhatia et al., 2024) further highlight significant drops in VLM performance on local cultural concepts. Our work contributes to this line of research by studying explicit cultural adaptation in a generative setting, focusing on bidirectional cross-cultural transcreation.

Meme Understanding and Analysis. Most prior research on memes has focused on discriminative tasks, such as classification and detection. For example, PromptHate (Cao et al., 2023) addresses hateful meme detection (Kiela et al., 2021; Kumar and Nandakumar, 2022; Sharma et al., 2023), while MGMCF (Zheng et al., 2024) models fine-grained persuasive features (García-Díaz et al., 2024). Other studies document systematic cultural differences in online humor (Mutheu, 2023; Nissenbaum and Shifman, 2018; Tanaka et al., 2022), analyze the sentiment of memes (Sharma et al., 2020), and show that annotators' cultural backgrounds influence interpretation. In contrast, comparatively little work has explored *generative* meme tasks, particularly those requiring culturally grounded adaptation rather than classification.

Cross-Cultural Transcreation. Transcreation has recently emerged as a framework for adapting content across cultures beyond literal translation. Khanuja et al. (2024b) introduce image transcreation and show that models struggle to balance semantic preservation with cultural appropriateness, motivating dedicated evaluation metrics (Khanuja et al., 2024a). While meme datasets such as MemeCap (Hwang and Shwartz, 2023) or MET-Meme (Xu et al., 2022) provide large-scale meme captioning resources, they lack cross-cultural pairs required for transcreation. Our work extends transcreation to memes, which require tight visual–textual coupling and humor preservation, and introduces a bidirectional benchmark spanning US and Chinese cultures.

Generative Vision–Language Models. Recent VLMs demonstrate strong multimodal understanding and reasoning capabilities, including LLaVA (Liu et al., 2023, 2024), GPT-4V (Lin et al.,

2025), and Qwen-VL (Bai et al., 2023; Wang et al., 2024). Parallel advances in image generation models enable increasingly faithful prompt-based visual synthesis (Black Forest Labs, 2024; Verdú and Martín, 2024). While these models provide the foundation for multimodal generation, their effectiveness for culturally grounded creative adaptation remains underexplored—a gap our study aims to address.

Evaluation of Multimodal Generation. Standard metrics such as CLIPScore (Hessel et al., 2021) and TIFA (Hu et al., 2023) focus on text–image alignment but are not designed to capture cultural fit or humor preservation. Existing cross-cultural benchmarks, including CVQA (Romero et al., 2024), GlobalRG (Bhatia et al., 2024), and WorldCuisines (Winata et al., 2025), primarily address visual question answering rather than generative tasks (Bai et al., 2025; Bhalerao et al., 2025). In this work, we evaluate meme transcreation using both human judgments and automated LLM-based evaluation across multiple quality dimensions, enabling a systematic comparison of human and automated assessments in a cross-cultural setting.

3 Hybrid Transcreation Framework

We introduce a hybrid framework for cross-cultural meme transcreation that balances preservation of communicative intent with culturally appropriate adaptation. Rather than framing memes as a translation task, *our approach explicitly separates culture-invariant elements from those that must change to ensure cultural authenticity*. This section outlines the guiding principles of the framework and the three-stage pipeline that implements them.

In practice, memes combine universal and culture-specific components: literal translation often preserves surface meaning but fails culturally, while full recreation risks drifting from the original intent. To address this trade-off, our hybrid strategy is grounded in three principles.

Preserve universal elements. We retain transferable aspects such as core humor mechanisms (e.g., irony, exaggeration), high-level emotional intent, and common meme formats.

Adapt culture-specific elements. We identify and replace culturally grounded references—such as pop culture, idioms, visual symbols, and stylistic conventions—with culturally appropriate alternatives rather than literal translations.

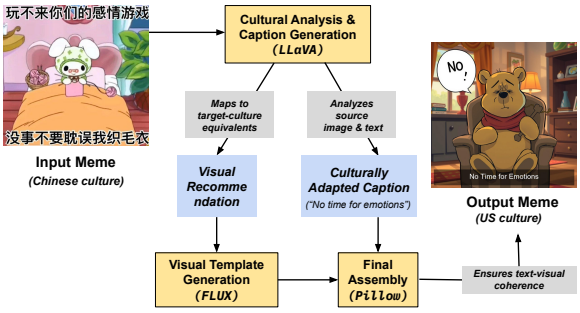


Figure 2: Overview of our three-stage meme transcreation pipeline. (1) A VLM analyzes the *original/input* meme, identifies cultural references and intent, and generates a culturally adapted caption. (2) A diffusion model produces a meme-style visual template aligned with the target culture. (3) A text overlay module assembles the output *transcreated* meme.

Maintain intent consistency. Across all stages, we preserve the meme's communicative goal—*what it aims to express or satirize*—even when textual and visual inputs change.

3.1 Meme Transcreation Pipeline

Figure 2 illustrates our modular three-stage meme transcreation pipeline: *cultural reasoning*, *visual generation*, and *final assembly*, enabling independent control and analysis across cultural directions.

Stage 1: Cultural Analysis and Caption Generation. We use LLaVA 1.6 (13B) (Liu et al., 2024) as the core vision-language model for cultural analysis and caption generation. The model takes the original meme image as input and is prompted to (1) analyze cultural references and humor mechanisms, (2) extract the underlying intent, (3) map source-culture concepts to target-culture equivalents, and (4) generate a culturally appropriate caption in the target language.

We employ LLaVA 1.6 because it offers strong vision–language alignment, robust Chinese and English multilingual performance, and open-source reproducibility. Importantly, we do not fine-tune the model, focusing instead on prompt-based control to isolate the effects of cultural reasoning without introducing task-specific training bias. This stage outputs both a transcreated caption and high-level recommendations for adapting the visual content (i.e., mood, background - examples in Appendix A)

Stage 2: Visual Template Generation. Using the visual recommendations from Stage 1, we generate culturally adapted meme templates with FLUX.1 Schnell (Black Forest Labs, 2024; Verdú and Martín, 2024), a diffusion-based image generation model designed for strong prompt adherence

and fast iteration. At this stage, the goal is not photorealism but meme-appropriate visuals that support the intended humor and allow for clear and readable text overlay. The generated images preserve universal meme structures (e.g., reaction shots, simple backgrounds) while adapting culture-specific elements. For example, Western celebrity figures are often replaced with symbolic or animal-based representations that are more common in Chinese meme culture. Emotional tone is conveyed through facial expressions, posture, and visual metaphors that align with conventions in the target culture.

Stage 3: Final Assembly. In the final stage, we automatically combine transcreated captions with the generated visual templates using Pillow, an open-source image processing library.[1] This step handles font selection, text placement, and layout conventions appropriate for the target culture (e.g., denser layouts for Chinese text and more spaced layouts for English captions), following common social media meme practices. We apply dynamic text wrapping, semi-transparent background overlays for readability, and center-aligned multi-line captions positioned near the image bottom. Final manual quality checks verify readability, visual–text coherence, and that captions do not obscure key visual elements.

Implementation Details. All models are used in their pre-trained form, with prompt engineering (Appendix F) and decoding parameter tuning (e.g., temperature, top-p) to balance creativity and consistency. This modular design supports reproducibility, scalability across cultures, and controlled analysis of cross-cultural meme transcreation.

4 MemeXGen Dataset

To study cross-cultural meme transcreation in a controlled and realistic setting, we introduce **MemeXGen**, a multilingual and multicultural dataset of Chinese and US meme pairs. The dataset consists of 6,315 *original memes* collected from authentic social media platforms and 6,315 *transcreated memes* produced by our pipeline. For each original meme, we generate a corresponding transcreated version in the opposite cultural context, resulting in a total of *6,315 bidirectional meme pairs: 3,165 Chinese→US and 3,150 US→Chinese*. This paired structure enables direct comparison of transcreation quality across directions.

[1] https://python-pillow.org/

4.1 Data Sources

MemeXGen is designed to support systematic evaluation and analysis, with an emphasis on cultural authenticity, diversity of humor styles, and balanced coverage across Chinese and US cultures.

Chinese Memes. *Original* Chinese memes are sourced from the publicly available *Chinese Meme Description Dataset*[2], which aggregates content from two major Chinese social media platforms: **Xiaohongshu** and **Weibo**. Xiaohongshu contributes lifestyle- and emotion-focused memes, while Weibo provides fast-paced, commentary-driven content reflecting mainstream Chinese internet culture.

US Memes. *Original* US memes are drawn from the *MemeCap* dataset[3], which collects memes from popular Reddit communities such as `r/memes` and `r/dankmemes`. These memes reflect dominant US humor styles, including sarcasm, irony, pop culture references, and situational storytelling.

These sources offer complementary views of meme culture in two distinct cultural contexts, enabling systematic bidirectional transcreation.

4.2 Filtering and Dataset Composition

During data inspection, we observe that some *original* memes contain potentially sensitive content (e.g., political references) that could interfere with fair evaluation or raise ethical concerns. To address this, we manually filter the *original* memes to ensure responsible use and reliable evaluation. Specifically, we remove memes that are offensive, contain low-quality or corrupted images, rely heavily on mixed languages, or exhibit weak visual–text integration. After filtering, the dataset contains 6,315 *original* memes, equally split across US and Chinese. We notice that the retention rate is substantially higher for the Chinese subset (97.6%) than for the US subset (55.0%), reflecting stricter content moderation on Chinese platforms compared to the more permissive nature of Reddit.

4.3 Annotation and Evaluation Split

To support emotion analysis and human evaluation, we annotate two disjoint subsets of the data.

Emotion annotations subset. We annotate ≈10% of the *original* memes data (628 memes, equally split between US and Chinese memes) with emotion labels, including emotion category (*Joy, Anger,*

[2] https://github.com/THUDM/chinese-meme-description-dataset

[3] https://github.com/hwang1996/MemeCap

Sadness, Fear, Disgust, Surprise) and emotion intensity on a 1–5 Likert scale. Annotation guidelines follow recent multilingual emotion classification work, as described in BRIGHTER (Muhammad et al., 2025). Three expert annotators perform the annotations independently, achieving strong agreement (Fleiss' (Fleiss, 1971) $\kappa = 0.869$ for emotion category and $\kappa = 0.536$ for intensity).

Human evaluation subset. We reserve a separate 10% subset (628 *original* memes) as the test set for transcreation experiments. This split includes 313 Chinese→US and 315 US→Chinese *original-transcreated* meme pairs and is used exclusively for human evaluation of meme transcreation. Evaluation details are provided in Section 5.

4.4 Dataset Characteristics

To better understand the cultural makeup of the *original* memes, we analyze topic and emotion distributions using Qwen-VL-Max (Bai et al., 2023; Wang et al., 2024), finetuned on the human annotated emotions. To validate the reliability of the predicted labels, an expert annotator manually reviews a random 10% subset and confirms that over 95% of the topic and emotion labels are correct.

Topic Distribution. Both cultures are dominated by themes related to *Internet Culture* (CN 61.0%, US 52.4%) and *Technology/Digital Life* (CN 10.6%, US 15.1%). Beyond these shared themes, clear differences emerge. **US memes more often frame education, family, and everyday experiences as relatable, narrative-driven humor** (e.g., *Education*: 7.8%; *Family*: 4.9%), whereas **Chinese memes emphasize symbolic expression and social pressure**, with lower prevalence of these topics (*Education*: 2.1%; *Family*: 1.9%). *Gaming*-related humor appears among the top US topics (2.7%) but is largely absent from the Chinese top ranks, reflecting differing leisure and achievement orientations.

Emotion Distribution. Automated emotion classification shows that ***Joy* dominates in both cultures** (CN 69.3%, US 73.8%), consistent with memes' primary entertainment role. However, Chinese memes exhibit higher levels of *Anger* (8.3%) and *Sadness* (8.2%), suggesting more frequent **social critique and melancholic expression**. In contrast, US memes show relatively higher *Fear* (7.0%) and *Disgust* (4.4%), aligning with **anxiety-driven and cringe-based humor styles**.

These systematic differences motivate our hybrid transcreation approach and provide context for interpreting performance asymmetries in later experiments. Further data analysis is provided in Appendix B.1.

5 Evaluation Methodology

We evaluate our meme transcreation framework using both human and VLM-based evaluation.

5.1 Metric Definitions

Our evaluation captures not only text and image quality and their interaction, as commonly assessed in prior image generation work (Hu et al., 2023), but also *task-specific* aspects that are critical for cross-cultural transcreation, namely cultural fit and intent preservation. All quantitative metrics are rated on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

We evaluate each transcreated meme along six dimensions: **Caption Quality**, measuring clarity, tone, and meme-appropriate language; **Image Quality**, assessing visual clarity, composition, and recognizability; **Synergy**, capturing how well image and text work together to convey humor or emotion; **Cultural Fit**, evaluating appropriateness and relatability for the target culture; **Intent Preservation**, measuring retention of the original meme's message and emotional effect; and an **Overall Score**, computed as the average across all dimensions. Detailed dimension definitions are provided in Appendix C.

5.2 Human Evaluation

We evaluate meme transcreation on the human evaluation subset, with each meme independently assessed by three human evaluators. Because meme generation is inherently subjective and prior work highlights the importance of modeling annotator perspectives rather than simple aggregation (Deng et al., 2023), we report results separately for each evaluator. All evaluators are bilingual and bicultural, with deep familiarity with both Chinese and US meme cultures. Additional details on annotator backgrounds are provided in Appendix D.

Inter-annotator Agreement. Inter-annotator Pearson correlations indicate moderate to strong agreement ($r = 0.58$–0.81), reflecting reliable yet stylistically distinct evaluation perspectives.

5.3 Automatic Evaluation

To assess the feasibility of automated evaluation, we use six state-of-the-art VLMs on all the data:

Qwen-VL-Max (Bai et al., 2023), LLaVA-v1.6-Vicuna-13B (Liu et al., 2024), InternVL3-8B and InternVL3-14B (Zhu et al., 2025), Qwen3-VL-8B (Wang et al., 2024), and Aya-vision-8b (Dash et al., 2025). These models are selected for their multilingual Chinese–English support, ability to process multiple images, and public availability, enabling reproducible automatic evaluation.

VLM-Human Agreement. We asses VLM evaluation effectiveness by computing the Pearson correlation between each human evaluator and VLM across each dimension (results in Section 6.3).

6 Evaluation Results

We report evaluation results addressing our research questions, using human and automatic metrics to assess cross-cultural performance, directional effects, and evaluation reliability.

6.1 RQ1: Cross-Cultural Performance

Human Evaluation. Table 2 summarizes results from three human evaluators and six LLM evaluators across both transcreation directions. Human evaluators differ in strictness and focus: Evaluator 1 prioritizes entertainment value (mean: 4.42), Evaluator 2 adopts a balanced perspective (mean: 4.09), and Evaluator 3 applies stricter quality standards (mean: 3.31). The resulting 1.11-point spread highlights the inherent subjectivity of cross-cultural meme transcreation evaluation. **Overall, the mean human score of 4.07/5.0 indicates that the proposed pipeline produces effective and generally well-received transcreations.**

Dimension-Level Analysis. Across dimensions, *Caption Quality* receives the highest ratings (mean: 4.20), suggesting effective cross-cultural adaptation of meme text. *Image Quality* is also strong (mean: 4.05), supporting the reliability of FLUX.1 for meme-style visual generation. ***Synergy* is consistently high (mean: 4.23)**, indicating that captions and visuals work well together in most outputs. *Cultural Fit* shows the widest variation across evaluators (range: 3.39–4.57), reflecting the subjective and culturally grounded nature of authenticity judgments. *Intent Preservation* is rated favorably overall (mean: 3.84), though scores may be partially influenced by the dominance of Joy-related memes (69–74% of the data).

VLM Evaluation. Among automated evaluators, **Qwen-VL-Max performs best**, achieving the highest overall scores (3.95 for Chinese→US and 3.71 for US→Chinese) and showing **strong alignment with human judgments** (mean Pearson $r = 0.926$, all $p < 0.001$). Other LLM evaluators exhibit much weaker correlations with humans ($r \leq 0.33$), suggesting that most open-source models struggle to reliably evaluate creative, culturally grounded outputs. **On average, LLM scores are 0.54 points lower than human scores**, indicating a systematic tendency toward conservative scoring.

6.2 RQ2: Directional Asymmetries

We observe a clear directional asymmetry: **US→Chinese meme transcreations significantly outperform Chinese→US** (4.48 vs. 3.93 out of 5.0, $\delta = +0.55$, $p < 0.001$). This gap likely reflects several factors. First, US memes often rely on globally recognizable characters and themes that are easier to localize, whereas Chinese memes frequently depend on context-specific wordplay and implicit cultural knowledge. Second, current VLMs are more strongly exposed to US-centric data during training. Third, evaluators apply stricter authenticity expectations to Chinese→US outputs, where cultural mismatches are more salient to native speakers.

6.3 RQ3: Evaluation Framework Analysis

Table 3 reports correlations between human and LLM evaluators. **Qwen-VL-Max shows consistently strong alignment with all three human evaluators** (Evaluator 1: $r = 0.921$, Evaluator 2: $r = 0.964$, Evaluator 3: $r = 0.894$), with especially high agreement on *Intent Preservation* ($r = 0.993$). In contrast, other models exhibit substantially weaker correlations (e.g., InternVL3-14B: $r = 0.263$, Qwen3-VL-8B: $r = 0.252$, LLaVA-v1.6: $r = 0.005$). **These results suggest that evaluating creative, cross-cultural multimodal content requires deeper multilingual and multicultural grounding than most current open-source VLMs provide.**

6.4 Qualitative Analysis

To complement quantitative metrics, we present representative transcreation examples from both directions and our data analysis observations. Figure 3a shows two successful transcreation samples from both directions (Overall Score: 5.0/5.0), while Figure 3b illustrates several failed transcreation samples (Overall Score: 1.4/5.0).

Success Patterns (30% of outputs scoring ≥4.5/5.0). High quality meme transcreations con-

	Chinese→US						US→Chinese					
Evaluator	**Cap.**	**Img.**	**Syn.**	**Cult.**	**Intent**	**Overall**	**Cap.**	**Img.**	**Syn.**	**Cult.**	**Intent**	**Overall**
Evaluator 1 (H)	4.78	4.51	4.66	4.57	4.24	4.55	4.82	4.22	4.31	4.18	3.89	4.28
Evaluator 2 (H)	4.35	4.11	4.45	4.14	4.00	4.21	4.41	3.84	4.12	3.76	3.67	3.96
Evaluator 3 (H)	3.46	3.52	3.59	3.39	3.29	3.45	3.52	3.19	3.24	2.98	2.93	3.17
Qwen-VL-Max	**4.13**	3.86	**4.20**	**3.74**	3.72	**3.95**	**4.21**	3.58	**3.89**	**3.41**	3.44	**3.71**
LLaVA-v1.6	4.00	4.00	3.81	3.00	**4.00**	3.79	4.05	3.72	3.52	2.67	**3.71**	3.53
InternVL3-8B	3.78	3.84	3.48	3.46	3.97	3.69	3.84	3.55	3.16	3.12	3.68	3.47
InternVL3-14B	3.21	**4.39**	3.16	3.36	3.34	3.53	3.28	**4.11**	2.84	3.02	3.01	3.25
Qwen3-VL-8B	2.83	3.70	2.74	3.59	2.56	3.15	2.91	3.42	2.41	3.21	2.18	2.83
Aya-vision-8b	3.18	4.17	2.83	2.90	2.72	3.10	3.25	3.89	2.51	2.56	2.39	2.92

Note: Best VLM results per column are shown in **bold**, second-best VLM results are underlined. All dimensions rated 1–5 (higher = better). (H) = Human evaluator. Cap.=Caption Quality, Img.=Image Quality, Syn.=Synergy, Cult.=Cultural Fit, Intent=Intent Preservation.

Table 2: Evaluation Results for Chinese→US and US→Chinese Meme Transcreation

Human	VLM	Cap.	Img.	Syn.	Cult.	Intent	Overall
Eval. 1	Qwen-VL-Max	0.961	0.986	0.987	0.901	0.993	0.921
	LLaVA-v1.6	-0.039	-0.039	0.082	-0.178	-0.039	-0.026
	InternVL3-8B	-0.128	-0.026	-0.053	-0.106	0.041	-0.086
	InternVL3-14B	0.241	0.363	0.288	0.216	0.275	0.270
	Qwen3-VL-8B	0.219	0.394	0.316	0.215	0.289	0.281
	Aya-vision-8b	-0.117	-0.032	-0.087	-0.077	-0.065	-0.082
Eval. 2	Qwen-VL-Max	0.989	0.988	0.994	0.980	0.995	0.964
	LLaVA-v1.6	-0.016	-0.016	0.038	-0.112	-0.016	-0.027
	InternVL3-8B	-0.099	0.014	-0.033	-0.061	0.041	-0.066
	InternVL3-14B	0.294	0.350	0.351	0.323	0.365	0.329
	Qwen3-VL-8B	0.265	0.340	0.331	0.294	0.327	0.309
	Aya-vision-8b	-0.088	0.022	-0.048	-0.034	-0.031	-0.058
Eval. 3	Qwen-VL-Max	0.950	0.990	0.985	0.938	0.990	0.894
	LLaVA-v1.6	-0.050	-0.050	0.039	-0.161	-0.050	0.029
	InternVL3-8B	-0.065	0.083	0.014	0.008	0.107	-0.013
	InternVL3-14B	0.296	0.348	0.373	0.320	0.371	0.340
	Qwen3-VL-8B	0.193	0.368	0.310	0.235	0.285	0.263
	Aya-vision-8b	-0.024	0.101	0.016	0.050	0.025	-0.057

Note: Cap.=Caption Quality, Img.=Image Quality, Syn.=Synergy, Cult.=Cultural Fit, Intent=Intent Preservation. Cell colors indicate correlation strength.

Table 3: Pearson correlation coefficients (r) between Human and LLM evaluators across evaluation dimensions.

tain the following elements: (1) *Universally applicable character selection*—the use of recognizable archetypes that can be understood across cultures. (2) *Emotion-focused transcreations*—the retention of the original emotional context with the incorporation of cultural specifics. (3) *Use of natural language conventions*—the use of meme-like linguistic conventions associated with the receiving culture. (4) *Visual and textual unity*—careful matching of image and text

Failure Patterns (1.6% of outputs scoring ≤2.0/5.0). Errors that emerge on failed meme transcreations include: (1) *Failure in Captions*—the use of formal speech that dampens the casual meme vibe; (2) *Disconnects in Visuals*—the use of images that don't fit a culture or issues with visual generation; (3) *Failure to Preserve Humor Mechanisms*—the use of a format that is not amenable to joke structure; (4) *Complete synergy breakdown*—caption-image mismatch creating incoherent messaging.

Directional Patterns. US→Chinese transcreations more frequently achieve natural cultural integration, benefiting from globally recognizable US templates. Chinese→US transcreations struggle with context-dependent wordplay and philosophical concepts that lack Western equivalents, often resulting in superficial adaptations. Additional examples in Appendix G.

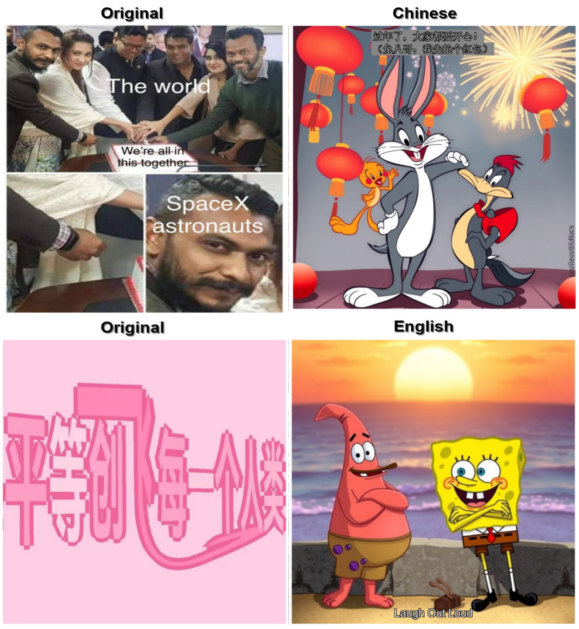
7 Main Takeaways

Effective Cross-Cultural Transcreation. Human evaluations show that the proposed hybrid approach produces high-quality transcreations (mean score: 4.07/5.0), successfully preserving humor and intent while adapting cultural specifics. Strong performance on Caption Quality (4.20) and Image–Text Synergy (4.23) confirms that the three-stage pipeline supports coherent multimodal generation.

Directional Asymmetry Matters. US→Chinese transcreation consistently outperforms Chinese→US (+0.55), reflecting both model exposure biases and deeper cultural differences. In particular, Chinese memes rely more heavily on context-dependent wordplay and implicit meaning, which are harder to adapt than the more visually universal templates common in US meme culture. These results highlight the need for culturally diverse training data in cross-cultural AI systems.

Limits of Automated Evaluation. Qwen-VL-Max shows strong agreement with human judgments ($r = 0.926$), demonstrating that automated evalua-

(a) **Best transcreation example** (**Score: 5.0/5.0**).
Transcreated (Chinese, top right — Bugs Bunny): "Family: Look who's up so early; My mental state after pulling an all-nighter to finish my assignment due."
Original (bottom left — panda meme): "What's wrong with work hours? Doesn't your company expect you to work during work hours?"



(b) **Worst transcreation example** (**Score: 1.4/5.0**).
Transcreated (Chinese, top right — Looney Tunes): "Celebrating the New Year, everyone must be happy! No one left behind, I'm giving you a family photo!"
Original (bottom left — Chinese text): "Attempts at equality satisfy no one."

Figure 3: **Qualitative examples of cross-cultural meme transcreation.** Left: successful US→Chinese adaptation preserving intent, humor, and cultural conventions. Right: failed Chinese→US adaptation illustrating loss of intent and cultural mismatch.

tion of creative, cross-cultural content is feasible. However, weaker correlations from open-source models suggest that reliable automated evaluation remains challenging without extensive multilingual and multicultural grounding.

8 Conclusion

We introduced a hybrid framework for cross-cultural meme transcreation that explicitly separates intent preservation from cultural adaptation, enabling principled analysis of how humor and meaning transfer across cultures. By combining vision–language models with diffusion-based image generation, our approach moves beyond literal translation and treats meme adaptation as a culturally grounded multimodal reasoning problem.

We curated and evaluated a dataset of 6,315 Chinese–U.S. meme pairs, combining authentic social media memes with systematically generated transcreations, and conducted a comprehensive bidirectional evaluation. Our results reveal consistent directional asymmetries in transcreation quality, demonstrating that current models handle certain cultural adaptations more effectively than others. These findings expose concrete limitations in cross-cultural generalization that are not visible in monolingual or translation-based evaluations.

We further show that carefully selected VLM-based evaluators can approximate human judgments on culturally grounded dimensions such as emotion and intent, while most open-source models remain unreliable for assessing intent and cultural fit. Finally, we release MEMEXGEN, the first parallel Chinese–U.S. meme transcreation corpus annotated for emotion and cultural intent, together with evaluation protocols and dataset splits. By open-sourcing data, models, and evaluation metrics, this work establishes a foundation for systematic study of computational humor and cross-cultural multimodal generation, and provides actionable benchmarks for future model development.

Limitations

Scope of Cultural Coverage. This study focuses on meme transcreation between Chinese and U.S. cultures, which differ substantially in language, humor conventions, and visual symbolism. While this contrast enables clear analysis of cultural asymmetries, our findings should not be assumed to generalize uniformly to other cultural pairs. Future work should extend this framework to additional language–culture settings to test the robustness of the observed patterns.

Generality of the Generation Framework. Our transcreation pipeline combines existing vision–language and diffusion models in a modular design intended to support interpretability and controlled analysis rather than architectural novelty. We do not claim optimality of this design, nor do we compare against all possible end-to-end prompting alternatives. Instead, our goal is to provide a transparent framework for studying cultural adaptation. Exploring simpler or fully integrated baselines remains an important direction for future work.

Interpreting Directional Asymmetries. We observe consistent performance differences between US→Chinese and Chinese→US transcreation. While we discuss plausible contributing factors—such as training data exposure, humor structure, and evaluator expectations—these explanations are correlational rather than causal. Disentangling these effects would require controlled experiments that vary data distributions, model pretraining, and evaluation populations independently.

Limits of Automatic Evaluation. Although Qwen-VL-Max shows strong alignment with human judgments in our setting, this result may reflect model-specific strengths in Chinese–English multimodal understanding rather than a general solution to evaluating culturally grounded humor. The weak performance of other open-source evaluators highlights that reliable automated evaluation remains challenging and should be interpreted with caution.

Dataset Composition and Emotion Coverage. Joy dominates the meme distributions in both cultures, reflecting real-world social media trends but limiting stress-testing on negative or socially critical humor. As a result, intent preservation scores may be optimistic for emotionally complex cases. Expanding emotion-balanced datasets is a key area for future research.

Evaluation at Scale and in the Wild. Human evaluation remains inherently subjective, and the observed variation across evaluators highlights the value of modeling diverse perspectives rather than collapsing them into a single score. Expanding the evaluation set and incorporating longitudinal, in-the-wild measurements (e.g., engagement or sharing behavior) would provide deeper insight into real-world cultural impact beyond offline quality ratings.

Broadening Cultural Perspectives. Our annotators are bilingual and bicultural with Chinese–U.S. experience, ensuring informed evaluation of both contexts. Future work can further broaden cultural representation by including evaluators with more localized or region-specific backgrounds, as well as exploring regional variation within Chinese and U.S. meme cultures. Such diversity would deepen understanding of cultural nuance and strengthen the generalizability of cross-cultural evaluation.

Ethical Considerations

Deployment Scope. Our pipeline prioritizes analytical clarity and controlled study of cultural adaptation rather than deployment efficiency. As with any automated cultural generation system, misuse, misinterpretation, or oversimplification of cultural signals remains a risk. We position meme transcreation as a *decision-support tool* intended to assist human creators and analysts, not as a fully autonomous content generator, and strongly recommend human oversight in real-world or sensitive deployments.

Content Safety. We apply stringent manual filtering to exclude offensive or sensitive content, including hate speech, discriminatory media, explicit material, and political content. High *Not Offensive* ratings (92.8% from human evaluators and 96.6% from LLM-based assessment) indicate the effectiveness of these safeguards. However, cultural sensitivity is inherently subjective: content acceptable in one cultural context may still be perceived as offensive in another. Our system therefore cannot guarantee zero harmful outputs and should be used with caution, particularly in public-facing applications.

Cultural Respect and Representation. Automated cultural adaptation risks reinforcing stereotypes or reducing complex cultural practices to surface-level substitutions. While our hybrid framework explicitly separates intent preservation from

cultural adaptation, evaluator feedback reveals occasional cases of shallow cultural mapping (e.g., direct visual substitution without deeper contextual grounding). These limitations highlight the importance of human-in-the-loop workflows, where automated transcreation outputs are treated as drafts rather than finalized content.

Data Privacy and Attribution. All source memes are collected from publicly accessible platforms (Xiaohongshu and Weibo for Chinese memes; Reddit for U.S. memes) and do not contain personal identifying information. We respect the implicit consent associated with public content sharing, though proper attribution remains challenging for viral meme formats with unclear authorship. The dataset is intended strictly for research purposes, and we encourage responsible use consistent with platform norms and community standards.

Misinformation Potential. Meme transcreation tools could be misused to spread culturally-adapted misinformation or propaganda. We emphasize responsible deployment with content verification protocols and transparency about automated generation.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 15763–15784. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *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 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Parth Bhalerao, Mounika Yalamarty, Brian Trinh, and Oana Ignat. 2025. Multi-agent multimodal models for multicultural text to image generation. *ArXiv*, abs/2502.15972.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.

Black Forest Labs. 2024. Flux: Advanced text-to-image generation model. `https://github.com/black-forest-labs/flux`. GitHub Repository.

Rui Cao, Roy Ka-Wei Lee, Tuan-Anh Hoang, Junmo Pang, Kenji Kawaguchi, and Roger Zimmermann. 2023. Prompthate: Prompting for hateful meme classification. *Preprint*, arXiv:2302.04156.

Saurabh Dash, Shivalika Singh, Adrien Morisot, Beyza Ermis, Acyr Locatelli, Sungjin Hong, Arash Ahmadian, Yannis Flet-Berliac, Nathan Grinsztajn, Florian Strub, and 1 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *Preprint*, arXiv:2505.08751.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the ACL*, pages 1905–1925.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

José Antonio García-Díaz and 1 others. 2024. Umuteam at semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. *Proceedings of SemEval-2024*.

Muzhaffar Hazman, Susan McKeever, and Josephine Griffith. 2025. What makes a meme a meme? identifying memes for memetics-aware dataset creation.

Daniel Hershcovich and 1 others. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the ACL*, pages 6997–7013.

Jack Hessel and 1 others. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528.

Yushi Hu and 1 others. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. *Preprint*, arXiv:2305.13703.

Nithish Kannen, Arjun Palani, Vikram V. Ramaswamy, Olga Russakovsky, and Li Fei-Fei. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *Preprint*, arXiv:2407.06863.

Simran Khanuja, Vivek Iyer, Claire He, and Graham Neubig. 2024a. Towards automatic evaluation for image transcreation. *Preprint*, arXiv:2412.13717.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024b. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. *Preprint*, arXiv:2404.01247.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and 6 others. 2021. The hateful memes challenge: Competition report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2025. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *ACM Trans. Intell. Syst. Technol.*

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: towards ai for everyone, with everyone, by everyone. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

Sarah Mutheu. 2023. Cross-cultural differences in online communication patterns. *Journal of Communication*, 4(1):31–42.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Asaf Nissenbaum and Limor Shifman. 2018. Internet memes as contested cultural capital: The case of 4chan's/b/board. *New Media & Society*, 19(4):483–501.

David Romero and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2149–2163, Dubrovnik, Croatia. Association for Computational Linguistics.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Daniel Verdú and Javier Martín. 2024. Flux.1 lite: Distilling flux1.dev for efficient text-to-image generation. Hugging Face Model Hub.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Genta Indra Winata and 1 others. 2025. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 NAACL*, pages 3242–3264.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2887–2899, New York, NY, USA. Association for Computing Machinery.

Zhengyi Zhao, Shubo Zhang, Yuxi Zhang, Yanxi Zhao, Yifan Zhang, Zezhong Wang, Huimin Wang, Yutian Zhao, Bin Liang, Yefeng Zheng, Binyang Li, Kam-Fai Wong, and Xian Wu. 2025. Memereacon: Probing contextual meme understanding in large vision-language models. *Preprint*, arXiv:2505.17433.

Li Zheng, Xiaoming Wang, Yuhan Chen, and Jianwei Liu. 2024. Multi-granular multimodal clue fusion for meme understanding. *Preprint*, arXiv:2503.12560.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Meme Transcreation Framework

Examples of visual recommendations from our Stage 1 LLaVA output to the FLUX.1 model:

1. Create a cartoon image using Tom and Jerry in a detailed pose and expression. Tom, wearing his usual red shirt, is standing behind Jerry, who is dressed in his classic blue sweater. Both characters have a slightly confused look on their faces. Jerry is scratching his head while Tom looks away with a slight frown. Background: An indoor setting that resembles a cocktail party or garden tea event, with blurred figures of people in the background engaged in conversation. Style: Keep Tom and Jerry's traditional animation style with bold lines and solid colors. Mood: Soft focus and warm lighting, suggesting an evening event.

2. Create a cartoon image using Bugs Bunny in a sitting pose, looking upward with a surprised or bewildered expression. - Background: A dimly lit room with a desk cluttered with various toys and a window showing a starry night sky. - Style: Retain the classic animation style with bold lines and vibrant colors. - Mood: Soft, nostalgic lighting with a hint of melancholy.

B Dataset

B.1 Dataset Characteristics

B.2 Topic Distribution Analysis

We applied enhanced weighted topic detection using Qwen-VL-Max (Bai et al., 2023) to conduct a comprehensive topic analysis across all 6,315 filtered memes, revealing fundamental differences in cultural priorities and humor focus.

Chinese Meme Topics. Table 4 presents the top 10 topics in Chinese memes, collectively covering 97.2% of the dataset.

American Meme Topics. Table 5 presents the top 10 topics in American memes, collectively covering 97.7% of the dataset.

Cross-Cultural Topic Comparisons. Table 6 highlights key differences in topic priorities between the two cultures.

Key cultural patterns revealed: (1) *Digital Concentration*—Chinese memes more heavily focused on internet/digital life (71.6% combined vs. 67.5% in US); (2) *Educational Values*—American memes treat education as casual daily experience (7.8%), Chinese memes reflect intense academic pressure (2.1%); (3) *Family Representation*—American memes more frequently feature family humor (4.9%) vs. Chinese hierarchical respect (1.9%); (4) *Leisure vs. Achievement*—American gaming culture prominent (2.7%), absent from Chinese top 10.

B.3 Emotion Distribution Analysis

Using Qwen-based (Bai et al., 2023) automated emotion analysis, we classified all 6,315 memes according to Ekman's six basic emotions, providing insights into cross-cultural emotional expression patterns.

Chinese Meme Emotions. Table 7 presents the emotion distribution in Chinese memes.

American Meme Emotions. Table 8 presents the emotion distribution in American memes.

Cross-Cultural Emotion Comparisons. Table 9 highlights key differences in emotional expression priorities.

Key emotional patterns: (1) *Positive Emphasis*—Both cultures prioritize joy, Americans showing slightly higher positive focus (73.8% vs. 69.3%); (2) *Sadness Acceptance*—Chinese memes

Table 4: Topic Distribution in Chinese Memes (N=3,165)

#	Topic	Count (%)	Cultural Significance
1	Internet Culture	1,931 (61.0%)	Digital lifestyle dominance, social media
2	Technology Digital	337 (10.6%)	Tech adaptation, AI integration
3	Work Career	216 (6.8%)	996 culture, career pressure
4	Social Relationships	148 (4.7%)	Friendships, social dynamics
5	Communication Language	126 (4.0%)	Language barriers, expression styles
6	Personality Psychology	115 (3.6%)	Individual traits, emotional responses
7	Education Learning	65 (2.1%)	Academic pressure, Gaokao system
8	Family Dynamics	61 (1.9%)	Family relationships, generational gaps
9	Animals Pets	46 (1.5%)	Pet culture, cute content
10	Entertainment Media	32 (1.0%)	Movies, shows, celebrity culture

Top 10 Total: 3,077 memes (97.2%)

Table 5: Topic Distribution in American Memes (N=3,150)

#	Topic	Count (%)	Cultural Significance
1	Internet Culture	1,651 (52.4%)	Social media, viral content, online trends
2	Technology Digital	475 (15.1%)	Tech innovation, digital lifestyle
3	Education Learning	247 (7.8%)	School experiences, college culture
4	Work Career	198 (6.3%)	Job market, work-life balance
5	Family Dynamics	155 (4.9%)	Family relationships, parenting
6	Communication Language	87 (2.8%)	Expression styles, conversation humor
7	Gaming Entertainment	85 (2.7%)	Video games, gaming culture, esports
8	Personality Psychology	67 (2.1%)	Individual psychology, personality types
9	Social Relationships	57 (1.8%)	Friendships, social interactions
10	Entertainment Media	54 (1.7%)	Movies, TV shows, celebrity content

Top 10 Total: 3,076 memes (97.7%)

Table 6: Cross-Cultural Topic Priority Comparison

Topic	CN	US	Interpretation
Internet Culture	#1 (61.0%)	#1 (52.4%)	Both dominant; China more concentrated
Technology Digital	#2 (10.6%)	#2 (15.1%)	US higher tech innovation focus
Education Learning	#7 (2.1%)	#3 (7.8%)	US: daily life; China: high-stakes pressure
Family Dynamics	#8 (1.9%)	#5 (4.9%)	US: frequent topic; China: serious element
Gaming Entertainment	–	#7 (2.7%)	US leisure vs. China work/study priority
Work Career	#3 (6.8%)	#4 (6.3%)	Similar priority, different intensity

Table 7: Emotion Distribution in Chinese Memes (N=3,165)

Emotion	Count (%)	Cultural Context
Joy	2,193 (69.3%)	Dominant positive humor expression
Anger	263 (8.3%)	Frustration, social critique
Sadness	258 (8.2%)	Melancholy, disappointment
Surprise	213 (6.7%)	Shock, unexpected situations
Fear	144 (4.5%)	Anxiety, worry
Disgust	94 (3.0%)	Revulsion, distaste

Table 8: Emotion Distribution in American Memes (N=3,150)

Emotion	Count (%)	Cultural Context
Joy	2,325 (73.8%)	Primary emotional expression
Fear	219 (7.0%)	Anxiety, relatable worries
Anger	217 (6.9%)	Frustration, social commentary
Surprise	148 (4.7%)	Unexpected, absurd humor
Disgust	140 (4.4%)	Cringe, distasteful situations
Sadness	101 (3.2%)	Disappointment, darker humor

Table 9: Cross-Cultural Emotion Priority Comparison

Emotion	CN	US	Interpretation
Joy	#1 (69.3%)	#1 (73.8%)	Both dominant; US slightly higher positivity
Anger	#2 (8.3%)	#3 (6.9%)	China: more direct frustration expression
Sadness	#3 (8.2%)	#6 (3.2%)	China: 2.5× higher melancholic acceptance
Surprise	#4 (6.7%)	#4 (4.7%)	Similar priority, China higher absurdist humor
Fear	#5 (4.5%)	#2 (7.0%)	US: anxiety culture, relatable worry themes
Disgust	#6 (3.0%)	#5 (4.4%)	US: higher cringe/distaste expression

express sadness 2.5× more frequently, reflecting cultural acceptance of melancholic humor; (3) *Anxiety Expression*—American memes emphasize fear-based content (7.0% vs. 4.5%), aligning with therapeutic humor trends; (4) *Anger Manifestation*—Chinese memes show higher anger (8.3% vs. 6.9%), possibly reflecting more direct emotional expression; (5) *Cringe Culture*—American memes display higher disgust representation (4.4% vs. 3.0%), consistent with cringe comedy trends.

Impact of Joy Dominance on Experimental Design. The overwhelming dominance of Joy in both datasets (69.3% Chinese, 73.8% American) has important implications for our transcreation experiments and evaluation: (1) *Positive Bias in Evaluation*—Since most transcreated memes will naturally preserve joyful emotions, our system may appear more successful at humor preservation simply due to the high baseline of positive content. This necessitates careful interpretation of intent preservation scores in Chapter **??**; (2) *Limited Negative Emotion Testing*—With less than 30% of memes expressing negative emotions (anger, sadness, fear, disgust), our system receives limited training signals for adapting complex negative emotional tones across cultures, potentially underrepresenting challenges in transcreating emotionally nuanced content; (3) *Generalizability Concerns*—The skewed distribution means our findings may generalize better to lighthearted, positive meme content than to darker, satirical, or critical humor styles; (4) *Cultural Authenticity vs. Emotional Consistency*—The Joy dominance simplifies one aspect of transcreation (emotional tone transfer) while placing greater emphasis on cultural reference adaptation as the primary challenge. Despite this limitation, the Joy-dominant distribution accurately reflects real-world meme ecosystems where positive, shareable content naturally dominates social media platforms—making our experimental conditions ecologically valid even if not emotionally balanced.

C Evaluation Metrics

Quantitative Dimensions (1-5 scale):

Caption Quality Evaluates whether the generated caption works effectively as meme text, considering clarity, readability, appropriate meme language/tone, engaging phrasing, and proper text formatting.

Image Quality Assesses whether the generated image functions effectively as a meme visual, considering visual clarity and quality, appropriate meme composition, recognizable elements/characters, and visual appeal and memorability.

Synergy Measures how well image and caption work together, evaluating coherent message delivery, emotional or humorous impact, logical connection between visual and text, and overall meme effectiveness.

Cultural Fit Evaluates cultural adaptation quality, including alignment with target culture's humor style, appropriate cultural references, target audience relatability, and avoidance of cultural misunderstandings.

Intent Preservation Assesses preservation of the original meme's intent, including message consistency, emotional tone preservation, humorous effect maintenance, and core meaning retention.

Overall Score The average of all dimension scores and reflects overall quality.

D Human Evaluator Profiles

Our evaluation employed three bilingual, bicultural evaluators with a deep understanding of both Chinese and US cultural contexts:

Evaluator 1. Native Chinese speaker with 10+ years US residence, PhD in Communication Studies. Regular engagement with both Weibo/Xiaohongshu and Reddit meme communities. Assessment style: Entertainment-focused, generous scoring emphasizing humor effectiveness over technical perfection. Mean overall score:

4.42/5.0.

Evaluator 2. American-born Chinese with native-level proficiency in Mandarin and US, MA in Comparative Cultural Studies. Active participation in both Chinese and US digital cultures. Assessment style: Balanced and objective, applying consistent standards across dimensions. Mean overall score: 4.09/5.0. Showed highest correlation with Qwen-VL-Max (F1 = 0.925, $r = 0.964$).

Evaluator 3. Native Chinese speaker with 12+ years of US experience, professional translator with meme localization background. Expertise in cultural adaptation nuances. Assessment style: Critical and quality-focused, emphasizing cultural authenticity and linguistic precision. Mean overall score: 3.31/5.0.

All evaluators received identical structured prompts specifying six evaluation dimensions, worked independently without access to others' ratings, and maintained consistency through detailed scoring rubrics. Inter-evaluator correlations demonstrate moderate to strong agreement: Evaluator 1-2 ($r = 0.72$), Evaluator 1-3 ($r = 0.58$), Evaluator 2-3 ($r = 0.81$), confirming reliable yet stylistically distinct evaluation perspectives.

E VLM Evaluator Details

Six VLMs served as automated evaluators, selected for multilingual (Chinese-English) capability, multi-image processing, and reproducibility:

- **Qwen-VL-Max** (Alibaba Cloud): Commercial API with extensive Chinese-English training, demonstrated exceptional human correlation ($r = 0.926$).
- **LLaVA-v1.6-Vicuna-13B**: Same architecture as transcreation system, showed no meaningful correlation ($r = 0.005$).
- **InternVL3-8B/14B**: Recent open-source models with strong vision capabilities, achieved weak positive correlation (8B: $r = -0.049$, 14B: $r = 0.263$).
- **Qwen3-VL-8B-Instruct**: Smaller Qwen variant, weak correlation ($r = 0.252$).
- **Aya-vision-8b**: Massively multilingual model, slight negative correlation ($r = -0.043$).

All LLMs received identical prompts specifying evaluation dimensions and rating scales. Temperature set to 0.7 for balanced consistency-creativity tradeoff.

F Transcreation Prompts

Stage 1 (LLaVA 1.6) - Cultural Analysis Prompt Example:

> *You are a cultural adaptation expert. Analyze this [SOURCE CULTURE] meme and create a transcreated version for [TARGET CULTURE] audiences. Your response should include:*
>
> *1. Cultural Context Analysis: Identify culture-specific elements (references, idioms, visual symbols, humor mechanisms)*
>
> *2. Intent Extraction: What is the core message/emotion/joke?*
>
> *3. Target Culture Mapping: Find equivalent concepts in [TARGET CULTURE]*
>
> *4. Transcreated Caption: Generate a new caption preserving intent while using [TARGET CULTURE] appropriate references and style*
>
> *5. Visual Recommendations: Describe ideal visual template (characters, setting, composition) culturally appropriate for [TARGET CULTURE]*

Stage 2 (FLUX.1) - Visual Generation Prompt Example:

> *Create a meme-style image: [LLaVA's visual recommendations]. Style: internet meme, high contrast, recognizable characters, clear composition suitable for text overlay. [TARGET CULTURE]-appropriate visual elements. Resolution: 1024x1024px.*

Full prompt templates with examples will be made available in our public repository upon acceptance.

G Example Transcreations

Success Example - US→Chinese:

Source (US): "Nobody: Absolutely nobody: Me at 3 am:" [Image: Person raiding refrigerator]

Transcreated (Chinese): "深夜两点的我:" (Me at 2 am) [Image: Cartoon cat staring at food]

Adaptation rationale: Replaced human figure with animal imagery (preferred in Chinese memes), adjusted time (2 am vs 3 am reflects Chinese sleep patterns), simplified narrative structure for conciseness.

Challenge Example - Chinese→US:

Source (Chinese): "内卷" (involution) concept with study-exhausted imagery

Transcreated (US): "The grind never stops" [Office worker imagery]

Limitation: US lacks a precise equivalent for "内卷" (intensifying competition in zero-sum environments). "Grind culture" captures work intensity but misses the systemic competition aspect, illustrating cultural untranslatability challenges.

Additional examples and failure case analysis are available in the supplementary materials.

Figure 4: Examples of successful Chinese→US meme transcreations.
Original (middle left — dog meme): "Honest smile"
Original (bottom left — angry emoji meme): "You looking for a knuckle sandwich?"

Figure 5: Examples of successful US→Chinese meme transcreations.
Transcreated (top right — spongebob meme): "Dad thinks I'm up early studying, I'm really on my 14th straight gaming session"
Transcreated (middle right — spongebob meme): "Kid: Mom, I'm playing a game
Mom: I'm cooking
Kid: Can you pause it?
Mom: How dare you use my own teachings against me?"
Transcreated (middle right — Spencer Wright meme): "Tiktok be spamming ads upfront, still not gonna pay for premium"