When Cultures Move: Measuring and Improving Multicultural Text-to-Video Generation

Shuowei Li **Yuming Zhao** **Parth Bhalerao** **Oana Ignat**
Department of Computer Science and Engineering
Santa Clara University, Santa Clara, CA, USA
{yzhao4,pbhalerao,oignat}@scu.edu

Abstract

Text-to-video (T2V) generation has achieved strong visual fidelity, yet its ability to represent multiple cultures within a single scene remains largely unexplored, a gap with real consequences as generative video becomes a mainstream medium for cultural storytelling. We introduce multicultural text-to-video generation as a new task and present the first benchmark designed to study this setting: 243 culturally grounded prompts and 972 videos, spanning three cultures (Chinese, American, Romanian), three action categories, and both mono-cultural and cross-cultural scenarios. Using this benchmark, we analyze how current T2V models represent culture with CLIP-based metrics, VLM-as-judge assessments, and video quality measures. As one strategy for improving cultural fidelity, we explore MAVEN, which decomposes prompts into *person*, *action*, and *location* dimensions, each handled by a culturally specialized agent in parallel or sequentially. In our controlled setting, parallel specialization achieves the strongest cultural relevance while maintaining visual quality and temporal consistency, suggesting that structured prompt refinement is a promising direction for multicultural T2V generation. We release our dataset and code at: https://github.com/AIM-SCU/MAVEN.

1 Introduction

Text-to-video (T2V) generation has rapidly advanced (Yang et al., 2025; Sun et al., 2024; OpenAI, 2024), shifting the central challenge from visual realism alone to *semantic faithfulness*. Among these dimensions, *cultural grounding*, how people, actions, and places are represented with respect to specific cultures, remains both critically important and insufficiently understood. While prior work has documented systematic cultural gaps in language and image models (Li et al., 2024b; Liu et al., 2025; Kannen et al., 2024), the video setting, where cultural content must remain coherent across both space and time, has received almost no dedicated study. To our knowledge, this work is among the first to systematically frame *multicultural T2V generation* as a problem in its own right.

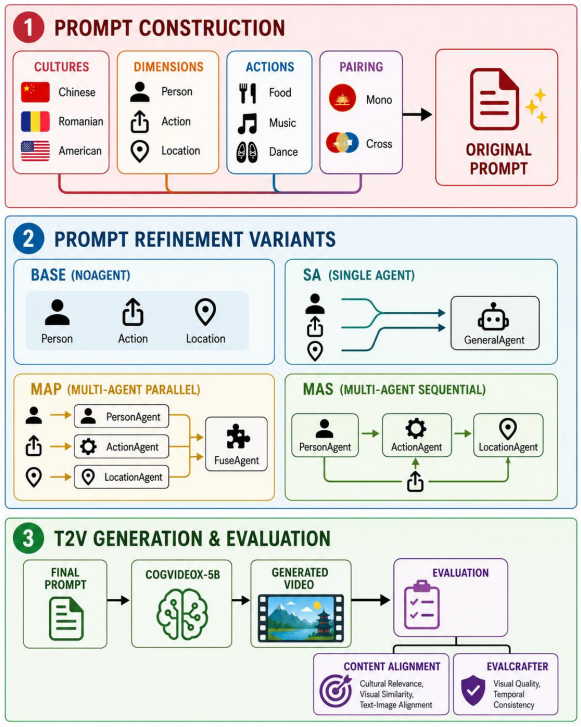


Figure 1: Most T2V systems assume a single culture per scene, but real scenes are often multicultural, combining person, action, and location from different backgrounds. We explore agent-based prompt refinement, shown here, as one strategy for this.

Culture is inherently compositional: a person from one cultural background may perform an action associated with another culture in a location tied to a third. Such *cross-cultural* scenarios parallel the composition problem studied in image transcreation (Khanuja et al., 2025) and matter for downstream uses such as inclusive content creation and accessibility. Yet current T2V benchmarks overwhelmingly assume mono-cultural prompts (Huang et al., 2026; Chen et al., 2025), implic-

Proceedings of the Fifth Workshop on NLP for Positive Impact (NLP4PI), pages 575–591
October 28, 2026 ©2026 Association for Computational Linguistics

itly treating culture as a single uniform attribute, and most T2V systems convey all cultural information through a single prompt or refinement agent, even though person appearance, action execution, and location depiction each draw on distinct forms of cultural expertise. Building on recent multi-agent T2V frameworks (Yuan et al., 2024; Wang et al., 2026), we introduce MAVEN (Multi-Agent Video Enrichment for cultural Narrative), which decomposes cultural grounding into three core dimensions, *person*, *action*, and *location*, each handled by a culturally specialized agent that enriches prompts before video generation. We evaluate on a benchmark of 243 prompts and 972 videos spanning three cultures, three action categories, and both mono-cultural and cross-cultural configurations (Section 3), using CLIP-based metrics (Radford et al., 2021), VLM judgments of cultural relevance (Liu et al., 2025), and video quality assessments (Liu et al., 2024).

Beyond model performance, this problem carries real **societal stakes**: as T2V becomes a mainstream medium for storytelling, education, and marketing, faithfully representing cultures matters for avoiding stereotyping, erasure of underrepresented traditions, and misrepresentation, motivating our focus on *measuring* cultural fidelity alongside improving it. Concretely, we organize our investigation around three research questions:

- **RQ1.** Can multi-agent prompt refinement substantively improve cultural fidelity in T2V generation, and does it close the gap between mono-cultural and cross-cultural prompts?
- **RQ2.** How does the choice of agent communication structure, specifically parallel specialization versus sequential refinement, affect cultural fidelity, visual quality, and temporal consistency?
- **RQ3.** To what extent do CLIP-based automatic metrics agree with VLM-based judgments of cultural relevance, and what does any disagreement reveal about the limits of current T2V evaluation?

2 Related Work

Cultural Understanding in LLMs and VLMs. Cultural gaps have been documented in both LLMs (Li et al., 2024b,a) and vision–language models (Kannen et al., 2024; Liu et al., 2025; Nayak et al., 2025), pointing to systematic knowledge gaps driven by sparse geographic coverage in training data. Culturally grounded evaluation has been extended to multilingual visual question answering (Romero et al., 2024) and to multicultural caption and image generation, where multi-agent frameworks with cultural personas have been shown to improve cross-cultural depiction in static images (Bai et al., 2025; Bhalerao et al., 2026). Our work shares these concerns but identifies two further limitations. First, cultures are studied independently, with no existing work modeling cross-cultural *composition*, i.e., combining person, action, and location from distinct cultural sources within a single generation. Second, all prior work focuses on static image understanding or generation, leaving the temporal consistency challenges unique to video entirely unaddressed.

Multi-Agent Text-to-Video Generation. Multi-agent frameworks for T2V generation decompose generation into specialized subtasks, including end-to-end agent coordination (Yuan et al., 2024), film-crew simulation (Xie et al., 2024; Wu et al., 2025), cross-shot protagonist consistency (Hu et al., 2024), and script-writing alignment (Wang et al., 2026). A related line of work pursues LLM-guided prompt refinement for T2V along non-cultural axes (Xue et al., 2025; Ji et al., 2025; Gao et al., 2025; Yang et al., 2026). We follow these frameworks in decomposing generation into specialized subtasks, but reorient agent roles from *production* (director, editor, keyframe artist) to *cultural dimensions* (person, action, location), enabling systematic study of both mono- and cross-cultural fidelity within a single prompt.

Cultural Text-to-Video Benchmarks. Existing benchmarks for cultural T2V evaluation cover scenario-level cultural fairness (Huang et al., 2026; Wang et al., 2024), geo-cultural bias in city landscapes (Caliskan et al., 2025), attribute-object-action compositionality (Sun et al., 2025), and deep cultural knowledge integration (Chen et al., 2025). However, none address *cross-cultural* composition, where person, action, and location originate from distinct cultures in a single system. Our benchmark directly targets this gap.

3 Dataset

A central contribution of this paper is a new benchmark for culturally grounded T2V evaluation. The benchmark comprises 243 unique prompts (Table 1). Passing each through the four refinement

Component	Count
Cultures (Chinese, American, Romanian)	3
Action categories (food, music, dance)	3
Actions per (culture × category)	3
Distinct actions	27
Landmarks per culture	3
Distinct landmarks	9
Mono-cultural prompts	81
Cross-cultural prompts (novel)	**162**
Total unique prompts	**243**
Refinement pipelines (Base, SA, MAS, MAP)	4
Total generated videos	**972**
Duration per video (seconds)	5
Total video duration (minutes)	**81**

Table 1: MAVEN benchmark statistics.

pipelines (Base, SA, MAS, MAP; Section 4) yields 972 generated videos. The cross-cultural split is intentionally twice the mono-cultural split, since cross-cultural composition is the central novelty of our benchmark and the regime where current T2V models fail most clearly.

Cultures and actions. We span three cultures (**Chinese**, **American**, **Romanian**) chosen for geographic and linguistic distinctness, and three action categories (**food**, **music**, **dance**) chosen for visual distinctiveness and the existence of clearly filmable cultural variants. Each (culture × category) pair contributes 3 actions and each culture contributes 3 landmarks. The full list is given in Table 2.

Prompt construction. Prompts follow the template "P A at L" (person, action, location). **Mono-cultural** prompts fix all three roles to a single culture, yielding $3 \times 3 \times 3 \times 3 = 81$ prompts. **Cross-cultural** prompts draw P, A, L from three *distinct* cultures ($c_p \neq c_a \neq c_l$), yielding $3! \times (3 \times 3) \times 3 = 162$ prompts. A representative example is *"an American person eating dumplings at Bran Castle"* (P=American, A=Chinese, L=Romanian), which requires rendering three distinct cultural sources within a single video.

4 Methodology

Our methodology covers three components: (1) agent-based prompt refinement pipelines, (2) a fixed T2V generation model to ensure fair comparison across pipelines, and (3) implementation details for reproducibility. Figure 1 provides an overview.

Category	Items
Chinese	
Food	Peking duck, mooncakes, dumplings
Music	guzheng, erhu, dizi
Dance	fan dance, ribbon dance, umbrella dance
Locations	Forbidden City, West Lake, Potala Palace
American	
Food	hot dogs, burgers, pizza slice
Music	banjo, electric guitar, saxophone
Dance	hip-hop, moonwalk, tap dance
Locations	Statue of Liberty, Grand Canyon, Mount Rushmore
Romanian	
Food	sarmale, mici, mămăligă
Music	nai, cobză, ţambal
Dance	Hora, Sârba, Brâul
Locations	Bran Castle, Palace of Parliament, Wooden Churches of Maramureş

Table 2: Full list of cultural items in the MAVEN benchmark. Actions in prompts are formed by prepending the canonical verb (*eating*, *playing*, *dancing*) to each Food/Music/Dance item.

4.1 Agent-Based Prompt Refinement

While base prompts include explicit cultural markers, they often lack the fine-grained visual and contextual details needed for culturally faithful video generation. We therefore introduce agent-based prompt refinement, guided by the hypothesis that cultural knowledge for person appearance, action execution, and location depiction belongs to distinct domains.

Each agent is assigned a culture-specific persona corresponding to the dimension it refines: for example, an ActionAgent refining a Chinese food action is instructed as a culturally grounded observer of Chinese dining practices. Each agent's system prompt combines a dynamically generated cultural persona with a dimension-specific instruction; concrete prompts are provided in Appendix C.

Pipeline designs. We compare four prompt refinement pipelines, all taking an original prompt Pr_{orig} and producing a final prompt Pr_{final}.

Base (No-Agent).

$$\text{Pr}_{\text{final}} = \text{Pr}_{\text{orig}}$$

Single-Agent (SA). A single general-purpose agent jointly refines all prompt dimensions using a unified instruction and persona:

$$\text{Pr}_{\text{final}} = \text{SingleAgent}(\text{Pr}_{\text{orig}}).$$

Multi-Agent Parallel (MAP). Three specialist agents independently refine the prompt in parallel,

each targeting a single dimension:

$$\begin{aligned}\text{Pr}_P &= \text{PersonAgent}(\text{Pr}_{\text{orig}}),\\ \text{Pr}_A &= \text{ActionAgent}(\text{Pr}_{\text{orig}}),\\ \text{Pr}_L &= \text{LocationAgent}(\text{Pr}_{\text{orig}}).\end{aligned}$$

A fusion agent then merges their outputs into a coherent prompt:

$$\text{Pr}_{\text{final}} = \text{FuseAgent}([\text{Pr}_P, \text{Pr}_A, \text{Pr}_L]).$$

Multi-Agent Sequential (MAS). The same three specialist agents refine the prompt sequentially, each operating on the previous agent's output:

$$\begin{aligned}\text{Pr}_P &= \text{PersonAgent}(\text{Pr}_{\text{orig}}),\\ \text{Pr}_A &= \text{ActionAgent}(\text{Pr}_P),\\ \text{Pr}_{\text{final}} &= \text{LocationAgent}(\text{Pr}_A).\end{aligned}$$

These pipelines allow us to compare general versus specialized refinement, as well as parallel versus sequential agent coordination.

4.2 Text-to-Video Model

All videos are generated using CogVideoX-5B (Yang et al., 2025), an open-source diffusion-based text-to-video model. We use a fixed generation setup across all experiments (5-second videos at 720×480 resolution, 8 fps, 50 inference steps, guidance scale 6, fixed seed), ensuring that observed differences are attributable solely to prompt refinement strategies rather than model variation. CogVideoX-5B offers a strong balance between generation quality, computational cost, and reproducibility, and directly supports text-only prompts, making it well suited to our framework.

4.3 Implementation Details

Prompt refinement is implemented using a unified agent interface, with all agent calls logged for reproducibility.

All pipelines are executed asynchronously, enabling parallel agent execution where applicable. All agents are instantiated using LLaMA-3.1-70B (Grattafiori et al., 2024), served locally via Ollama and orchestrated through the AutoGen multi-agent framework (Wu et al., 2024).

5 Evaluation and Results

5.1 Evaluation Metrics

We evaluate generated videos along three complementary dimensions: cultural relevance, visual similarity, and text–image alignment. These dimensions are chosen to capture both whether cultural content is faithfully represented and whether prompt refinement introduces meaningful visual change or simply rewrites surface form.

Cultural Relevance. We define a Cultural Relevance Score (CRS) using CLIP (Radford et al., 2021) embeddings. For each prompt, we construct four cultural grounding statements (CGS) targeting distinct dimensions: **OCGS** ("This image belongs to {country}."), **PCGS** ("This image shows {person description}."), **ACGS** ("This image depicts {action}, a practice associated with {culture} culture."), and **LCGS** ("This image shows {landmark} in {country}."). For each video, we uniformly sample 5 frames across the video duration and compute frame-level cosine similarity with each CGS using CLIP:

$$\text{CRS}(v, \text{CGS}) = \frac{1}{5}\sum_{i=1}^{5} \text{sim}(f_i, \text{CGS}).$$

The four dimension scores — OCRS (overall), PCRS (person), ACRS (action), and LCRS (location) — are averaged to yield the overall CRS.

Visual Similarity. We compute a Visual Similarity Score (VSS) to measure how much refinement alters visual content. For each prompt, we pair the base video v_{base} with an agent-refined video v_{agent}, uniformly sample 5 corresponding frames across each video, and compute frame-wise CLIP similarity:

$$\text{VSS}(v) = \frac{1}{5}\sum_{i=1}^{5} \text{sim}(f_{\text{base},i},\ f_{\text{agent},i}).$$

A high VSS indicates that refinement enriches cultural and semantic details without substantially altering the overall scene layout, which we verify in the results below.

Text–Image Alignment. We compute a frame-level CLIP alignment score $\text{AS}(v, t)$ between the frames of a generated video v and a text description t, where t is the original prompt Pr_{orig}, the refined prompt Pr_{final}, or a cultural grounding statement. Two derived metrics quantify the effect of refinement: **cultural enrichment**, the gain in alignment with the *original* prompt for the agent video versus the base video; and **cultural relevance improvement**, the gain in alignment with the cultural

grounding statements. However, because agent refinement also produces longer and more detailed prompts, our current design does not fully isolate the effect of cultural specialization from prompt elaboration; a detail-matched non-cultural refinement baseline would help disentangle these effects.

5.2 VLM-Based Evaluation

To complement automatic metrics, we use a vision–language model (Gemini 2.5 Pro) (Team, 2025) as a judge (Liu et al., 2025). The VLM evaluates cultural relevance, visual similarity, and text–image alignment using the middle frame of the five uniformly sampled frames from each video, providing a consistent frame-level evaluation across all videos. Scores range from 1 (not culturally relevant) to 5 (highly culturally relevant) and are aggregated across dimensions to obtain VLM-based counterparts to CRS and VSS. Since the VLM observes a single representative frame, these scores assess frame-level cultural representation rather than temporal consistency across the full video. Exact evaluation prompts are provided in Appendix D.

5.3 Cultural Relevance Results

Multi-agent pipelines outperform both the base and single-agent (SA) baselines on the Cultural Relevance Score (CRS), with the parallel variant (MAP) achieving the highest overall CRS (Figure 2).

Overall performance. MAP improves CRS by +4.6% over the base pipeline and +1.6% over SA, with MAS following closely. These results suggest that distributing refinement across specialized agents can yield stronger cultural grounding than a single general agent in our setting.

Dimension-level analysis. Improvements are not uniform across dimensions. Person-related relevance improves consistently, though appearance traits occupy a narrow visual footprint and are less reliably captured by frame-level embedding similarity. Action relevance stays stable, as action cues are inherently more ambiguous for CLIP and richer prompts introduce competing visual elements. Location shows the largest gains, with multi-agent pipelines improving LCRS by over 12% relative to Base, more than double SA's gain, since architectural features translate more directly into CLIP-measurable signals. Overall, MAP achieves the most balanced improvements, supporting the potential benefit of dimension-specific specialization for cultural fidelity in our setting.

5.4 Visual Similarity

Despite improving cultural relevance, agent-based pipelines preserve high visual similarity to base videos (VSS > 0.68 for all pipelines; differences among SA, MAS, and MAP are $<1.2\%$). This indicates that refinement primarily enriches semantic and cultural details rather than altering overall scene layout, which is the intended behavior: culturally faithful generation should deepen the cultural grounding of a scene, not reconstruct it from scratch.

5.5 Text–Image Alignment Results

Agent-based refinement improves alignment with both the original and refined prompts (Figure 3).

Alignment with original prompts. MAP achieves the strongest cultural enrichment, improving alignment by +4.9% over base. This indicates that refinement helps T2V models better express the cultural intent already present in the original prompt, rather than drifting from it.

Alignment with refined prompts. All agent pipelines dramatically improve alignment with refined prompts (+29% to +38%), reflecting the increased specificity of enriched prompts. SA achieves the highest alignment here, likely due to greater stylistic coherence from a single-agent rewrite. However, alignment with the original prompt, which is more indicative of cultural fidelity, remains highest for MAP, reinforcing its practical advantage.

5.6 Mono vs. Cross-Cultural Results

Cross-cultural prompts are consistently more challenging than mono-cultural ones, with the average CRS decreasing by approximately 4 to 6% across all pipelines. The one exception is OCRS, which scores higher for cross-cultural prompts because it rewards explicit cultural fusion, whereas the per-dimension scores (PCRS, ACRS, LCRS) penalize the visual incoherence that arises when multiple cultures coexist within a single generated frame. Agent-based refinement narrows this gap, with MAP achieving the smallest per-dimension performance drop, suggesting that parallel specialization is particularly effective in cross-cultural scenarios.

5.7 VLM-Judged Cultural Relevance Results

VLM-based evaluation (Gemini 2.5 Pro) independently corroborates CLIP-based findings (Figure 4).

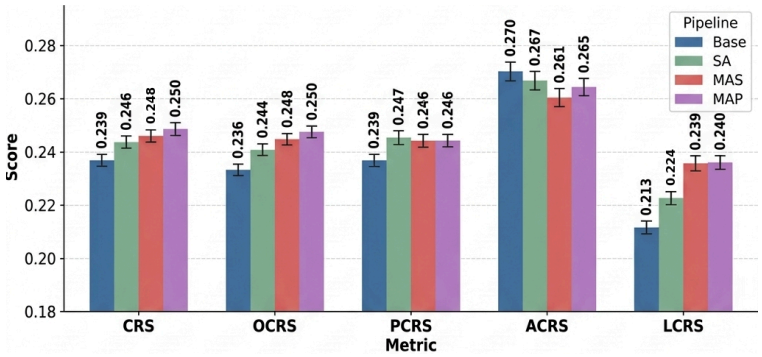


Figure 2: CRS and dimension-specific scores (OCRS, PCRS, ACRS, LCRS) for all four pipelines. Error bars show 95% CIs. MAP achieves the highest overall CRS (0.250), with statistically significant gains over Base (no-agent baseline) on CRS and LCRS.

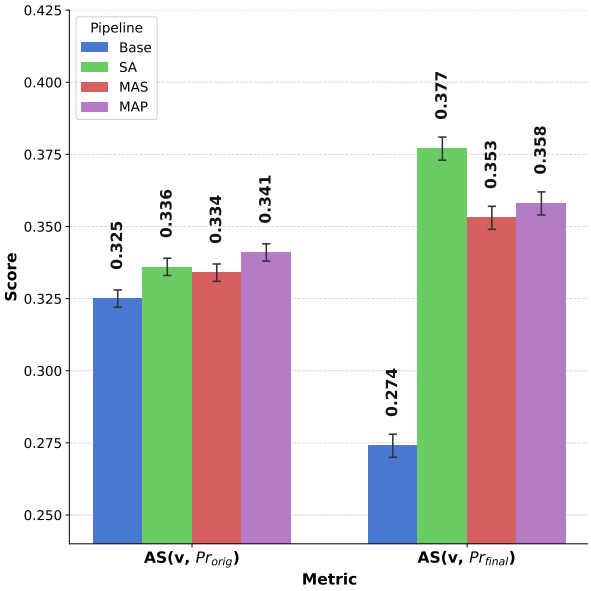


Figure 3: Alignment scores for all four pipelines with the original prompt Pr_{orig} and refined prompt Pr_{final}. Error bars show 95% CIs. All proposed methods (SA, MAS, MAP) improve significantly over Base (no-agent baseline) on refined prompt alignment, with MAP achieving the highest original prompt alignment (0.341).

MAP achieves the highest VLM-judged cultural relevance (VLM_CRS = 3.61), improving over base by +38% and over SA by +5.6%. Location remains the strongest dimension (VLM_LCRS: base 2.86 $\rightarrow$ MAP 4.43, +54.9%), while person-related cues remain the most challenging (VLM_PCRS: base 2.08 $\rightarrow$ MAP 3.12), indicating persistent difficulty in modeling culturally specific appearance.

5.8 VLM-Judged Visual Similarity Results

VLM-judged visual similarity scores (2.05–2.09 on a 1–5 scale) reveal a divergence from CLIP-based VSS: while CLIP reports high similarity (>0.68), VLM assigns substantially lower scores. This gap indicates that agent refinement introduces culturally meaningful visual changes that are perceptually salient to a semantic judge but not captured by embedding distance. All three pipelines score similarly, suggesting that the degree of visual change is not strongly dependent on the specific agent architecture in our setting.

5.9 VLM-Judged Text–Image Alignment Results

VLM-judged alignment further confirms MAP's advantage in enriching original prompt meaning, with a +61.3% improvement over base (MAP: 3.00 vs. base: 1.86). Alignment with refined prompts is highest for SA (2.88), again reflecting stylistic coherence from single-agent rewriting rather than cultural depth. The large gap between alignment with the original prompt (MAP: 3.00) and the refined prompt (base: 1.04) confirms that enriched prompts contain substantially richer cultural content than base videos represent.

5.10 Metric Correlation Results

CLIP-based (Radford et al., 2021) and VLM-judged metrics show moderate to strong Pearson correlation on person, action, and location, validating CLIP as a useful proxy: PCRS is most consistent (r = 0.61–0.66), followed by ACRS (r = 0.40–0.53) and LCRS (r = 0.40–0.71). Correlation is near zero for the overall dimension (OCRS), likely due to its abstract country-level framing. The base pipeline reaches the highest correlation on most dimensions, while agent-refined pipelines drop substantially: on LCRS, base reaches r = 0.71 versus 0.40–0.52 for agent pipelines, a 27–43% relative decrease. This suggests refinement introduces nuanced cultural cues better captured by VLM rea-

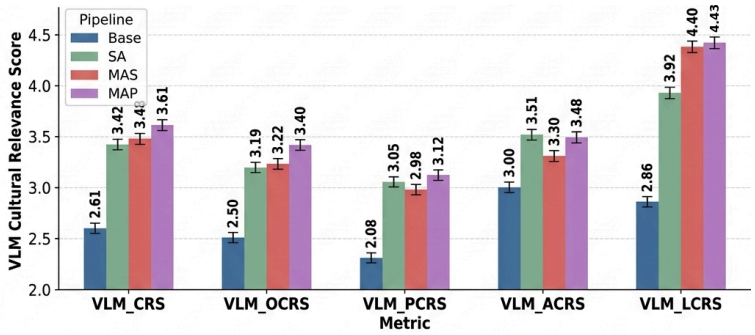


Figure 4: VLM-judged cultural relevance scores (scored 1–5) for all four pipelines across overall and dimension-specific metrics. Error bars show 95% CIs. MAP achieves the highest VLM_CRS (3.61), with statistically significant improvements over Base (no-agent baseline) across all dimensions, most notably on VLM_LCRS (Base: 2.86 → MAP: 4.43).

soning than embedding similarity.

5.11 Video Quality and Temporal Consistency Results

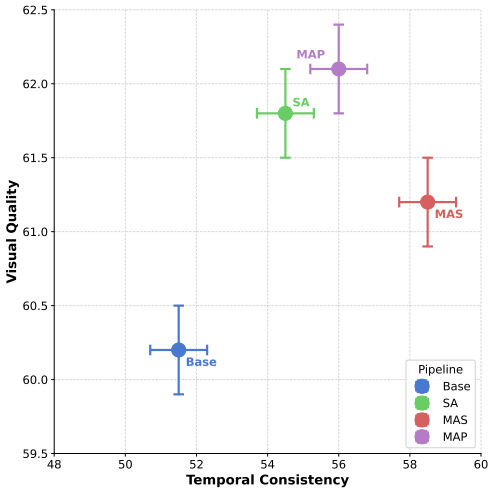


Figure 5: Visual Quality vs. Temporal Consistency for all four pipelines. Each point represents one pipeline; error bars show 95% CIs. All proposed methods improve over Base (no-agent baseline) on both metrics. SA leads on visual quality, MAS on temporal consistency, and MAP achieves the most balanced improvement across both.

Agent-based refinement improves both visual quality and temporal consistency (Figure 5). While SA maximizes visual quality and MAS maximizes temporal consistency, MAP achieves the most balanced improvement across both metrics. This balance, combined with its superior cultural relevance, makes MAP the most robust overall refinement strategy.

5.12 Qualitative Analysis

Qualitative results are consistent with quantitative trends. Figure 6 shows a comparison across all four pipelines for the mono-cultural example "a Chinese person playing guzheng at the Potala Palace". Figure 7 further presents representative cross-cultural examples spanning different combinations of person, action, and background across Chinese, American, and Romanian cultures. The full five-frame version of the example in Figure 6 is provided in Appendix B Figure 8, and further examples can be explored in our dataset.

Pipeline progression and sources of improvement. The four pipelines show a clear progression of cultural detail. **Base** uses the original prompt with no enrichment (CRS = 0.237). **SA** adds basic cues across all dimensions – traditional hair styling, a Tibetan rug, and the Palace's golden roofs (CRS = 0.242, +2.1%). **MAS** deepens all dimensions further: a qipao-inspired updo, specific body posture and hand technique, and a comprehensive architectural description (CRS = 0.249, +5.1%). **MAP** achieves the most comprehensive enrichment across all three dimensions simultaneously (CRS = 0.271, +14.3%), owing to *dimension-specific depth* (each specialized agent brings focused expertise a single agent cannot match) and *cross-dimensional balance* (parallel processing gives all three dimensions equally deep enrichment).

6 Lessons Learned

Parallel specialization outperforms sequential and single-agent refinement. Across a controlled benchmark spanning three cultures and multiple action types, parallel multi-agent refinement (MAP) achieves the highest cultural relevance, improving CRS by 4.6% over Base and 1.6% over SA, and VLM_CRS by 38.3% over Base and 5.6% over

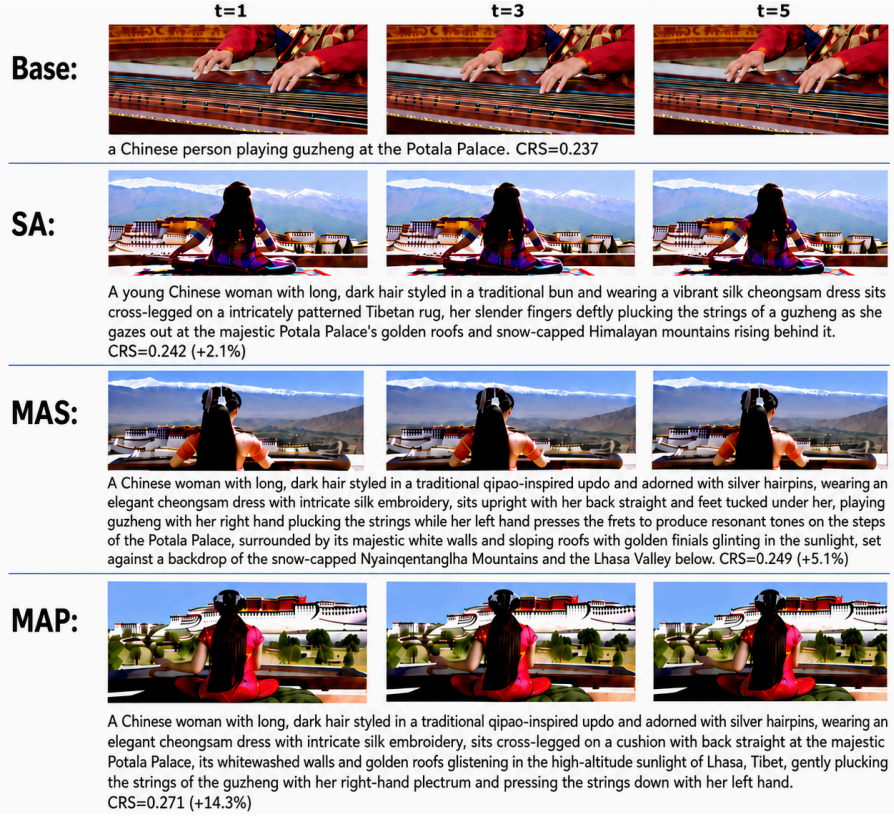


Figure 6: Qualitative comparison across all four pipelines (Base, SA, MAS, MAP) for the prompt *"a Chinese person playing guzheng at the Potala Palace"* (frames $t = 1$, $t = 3$, $t = 5$). Each pipeline shows a clear progression of cultural enrichment, with MAP achieving the most balanced and detailed rendering across person, action, and location dimensions (CRS: Base 0.237 $\rightarrow$ SA 0.242 $\rightarrow$ MAS 0.249 $\rightarrow$ MAP 0.271, +14.3%).

SA. The largest gain is on location, where LCRS improves by 12.7% over Base and VLM_LCRS rises from 2.86 to 4.43 (+54.9%). MAP also delivers the most balanced gains in visual quality (VQ) and temporal consistency (TC), reaching VQ 61.37 and TC 58.18 versus 60.69 and 52.82 for Base: SA leads on VQ (61.99) and MAS leads on TC (58.81), but neither matches MAP on both.

Cross-cultural generation benefits most from explicit decomposition. Cross-cultural prompts remain substantially more challenging than monocultural ones, yet benefit disproportionately from agent-based refinement. This indicates that explicit decomposition and recomposition of cultural dimensions is a promising strategy for cross-cultural generation, and suggests that future systems combining multiple cultural sources within a single scene may benefit from similarly compositional approaches.

Automatic metrics underestimate culturally nuanced improvements. While CLIP-based metrics correlate reasonably with VLM judgments, they systematically underestimate fine-grained cultural nuances introduced by agent refinement, pointing to the need for richer semantic reasoning in both T2V systems and their evaluation.

Structured specialization is beneficial in our controlled setting. Distributing refinement across culturally specialized agents (MAP) outperforms a single general-purpose agent (SA) on cultural relevance under the same underlying model. These results suggest that dimension-specific decomposition can benefit cultural prompt refinement in our experimental setting, while evaluation across additional LLMs and T2V models is needed before drawing broader conclusions about architecture or model scale.

7 Conclusion

In this paper, we introduce multicultural text-to-video generation as a new task, evaluating how models depict people, actions, and locations from different cultural backgrounds within a single video. We release the first benchmark for this setting: 243 prompts and 972 videos spanning

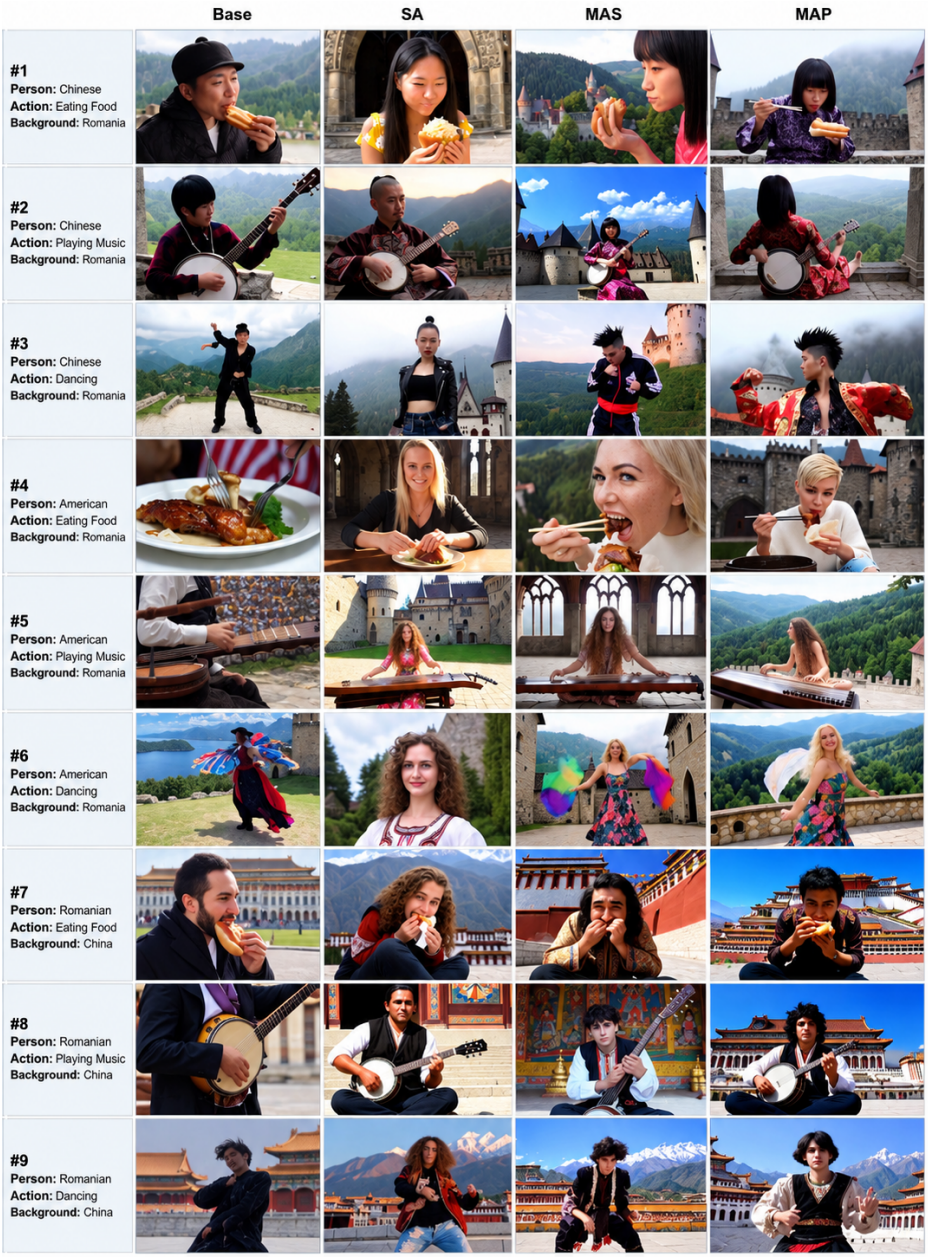


Figure 7: Representative cross-cultural qualitative examples across all four pipelines (Base, SA, MAS, MAP). Each row combines a person, action, and background drawn from different cultural sources, spanning Chinese, American, and Romanian settings. The examples illustrate how the refinement pipelines affect cultural composition across food, music, and dance scenarios.

three cultures, three action categories, and both mono- and cross-cultural scenarios. Our evaluations show that cross-cultural composition remains substantially harder than mono-cultural generation, and that automatic metrics can underestimate the cultural nuance introduced by prompt refinement. As one strategy for improving cultural fidelity, MAVEN decomposes prompts into person, action, and location dimensions, each assigned to a culturally specialized agent, narrowing this gap and yielding more balanced gains than single-agent or sequential refinement. As generative video becomes an increasingly influential medium for cultural storytelling, education, and marketing, we see this work as a step toward systems that measure and reduce the risk of cultural misrepresentation, helping generative video represent the world's diversity more faithfully. Dataset and code: https://github.com/AIM-SCU/MAVEN. We encourage future work extending our findings to additional settings, cultures, and languages.

Limitations

Limited cultural and activity coverage. Our study focuses on three cultures (Chinese, American, Romanian) and three activity categories (food, music, dance), which represent only a small subset of global cultural diversity. These choices were driven by practical considerations, including availability of culturally grounded visual resources and clarity of visual representation. Many cultural dimensions such as gesture, social interaction, spatial organization, color symbolism, and abstract concepts like values or social norms are not explicitly modeled or evaluated. As a result, our findings should not be interpreted as comprehensive across cultures or forms of cultural expression. Future work should expand coverage to a broader range of cultures and activities and incorporate additional cultural dimensions to test the scalability and generality of multi-agent refinement. Such expansion may also surface culture-specific challenges that are not visible in the current setting.

Evaluation on a single text-to-video model. All experiments are conducted using a single open-source T2V model CogVideoX-5B (Yang et al., 2025) to ensure controlled comparison. While this isolates the effect of agent-based prompt refinement, different T2V models may exhibit varying sensitivity to culturally enriched prompts due to differences in training data and generation mechanisms. Evaluating the proposed framework across multiple T2V architectures is an important direction for future work to assess model-agnosticity and identify potential model-specific adaptations.

Model-level generation artifacts. We observe that CogVideoX-5B (Yang et al., 2025) frequently generates subjects from behind or at oblique angles, obscuring culturally specific facial features and clothing details. This tendency limits the model's ability to render person-level cultural cues and likely contributes to the persistently lower PCRS scores observed across all pipelines. Future work should investigate generation strategies that encourage frontal or culturally expressive subject framing.

Human evaluation. Our current evaluation combines CLIP-based metrics with Gemini 2.5 Pro as a VLM judge, providing scalable assessments of cultural relevance across the full benchmark. Human judgments from culturally knowledgeable annotators would provide an important complementary perspective, particularly for nuanced or context-dependent cultural cues that automatic metrics may not fully capture. We therefore interpret our results within the scope of the automatic evaluation framework used in this study and encourage future work to validate these findings with culturally diverse human annotators.

Ethical Considerations

Our goal is to improve cultural fidelity and reduce misrepresentation in generative video systems. However, culturally grounded generation also carries risks, including stereotyping or overgeneralization if cultural signals are treated as fixed or homogeneous. The underlying LLMs and VLMs also inherit strong WEIRD (Mihalcea et al., 2025) biases from their training data (Atari et al., 2023) and can amplify representational harms when coverage of under-represented cultures is sparse (Bender et al., 2021), risks that culturally specialized agents may inherit or magnify. To mitigate these effects, we represent each culture through multiple distinct items rather than a single canonical referent, keep all prompts, refined outputs, and evaluation prompts inspectable, and treat our cultural representations as illustrative rather than exhaustive. We encourage future research to involve broader cultural perspectives and human evaluation to support more responsible and inclusive deployment.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, and the members of the AIM Lab at Santa Clara University for their support and helpful discussions.

References

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, and Joseph Henrich. 2023. Which humans? *PsyArXiv preprint*.

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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the

dangers of stochastic parrots: Can language models be too big? In *FAccT: ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Parth Bhalerao, Oana Ignat, Brian Trinh, and Mounika Yalamarty. 2026. When cultures meet: Multicultural text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 35808–35828, San Diego, California, United States. Association for Computational Linguistics.

Cantay Caliskan, Abylay Iskakov, Yuanzhu Li, Brendan McCollum, and Yi Ren. 2025. SimCityNet: Quantifying Geo-Cultural bias in AI-generated urban videos through interpretable scene embeddings. SSRN. Accessed 2025-10-10.

Yubin Chen, Xuyang Guo, Zhenmei Shi, Zhao Song, and Jiahao Zhang. 2025. T2vworldbench: A benchmark for evaluating world knowledge in text-to-video generation. *CoRR*, abs/2507.18107.

Bingjie Gao, Qianli Ma, Xiaoxue Wu, Shuai Yang, Guanzhou Lan, Haonan Zhao, Jiaxuan Chen, Qingyang Liu, Yu Qiao, Xinyuan Chen, Yaohui Wang, and Li Niu. 2025. RAPO++: cross-stage prompt optimization for text-to-video generation via data alignment and test-time scaling. *CoRR*, abs/2510.20206.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. 2024. Storyagent: Customized storytelling video generation via multi-agent collaboration. *CoRR*, abs/2411.04925.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2026. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285.

Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian Ge, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, Weilin Huang, and Ping Luo. 2025. Prompt-a-video: Prompt your video diffusion model via preference-aligned LLM. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18725–18735.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Simran Khanuja, Vivek Iyer, Xiaoyu He, and Graham Neubig. 2025. Towards automatic evaluation for image transcreation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7034–7047. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Huihan Li, Liwei Jiang, Jena D. Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. CULTURE-GEN: revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling (COLM)*.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *CoRR*, abs/2501.01282.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22139–22149. IEEE.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28657–28670.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd Van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. 2025. 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, Suzhou, China. Association for Computational Linguistics.

OpenAI. 2024. Sora system card. Accessed 2025-10-09.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D' Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Naome Etori, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. 2025. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 8406–8416. Computer Vision Foundation / IEEE.

Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. 2024. From sora what we can see: A survey of text-to-video generation. *CoRR*, abs/2405.10674.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. 2026. MAViS: A multi-agent framework for long-sequence video storytelling. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2273–2295, Rabat, Morocco. Association for Computational Linguistics.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversation. In *First Conference on Language Modeling (COLM)*.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. 2025. Automated movie generation via multi-agent cot planning. *CoRR*, abs/2503.07314.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F. Bissyand, and Saad Ezzini. 2024. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *CoRR*, abs/2408.11788.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2025. PhyT2V: LLM-guided iterative self-refinement for physics-grounded text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18826–18836.

Chengyi Yang, Pengzhen Li, Jiayin Qi, Aimin Zhou, Ji Wu, and Ji Liu. 2026. SCMAPR: self-correcting multi-agent prompt refinement for complex-scenario text-to-video generation. *CoRR*, abs/2604.05489.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. 2025. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. 2024. Mora: Enabling generalist video generation via A multi-agent framework. *CoRR*, abs/2403.13248.

A Reproducibility Details: Compute and Runtime

Experiments are conducted on NVIDIA H100 GPUs. Generating a single 5-second video takes approximately 3 minutes. SA and MAP pipelines require one additional minute for prompt refinement, while MAS requires approximately 2 additional minutes due to its sequential design. Across all prompts and pipelines, total runtime is approximately 65 GPU hours. The MAP runtime is comparable to SA because the specialist agents are executed concurrently.

B Qualitative Analysis

C Agent System Prompts and Personas

C.1 Persona Prompts

Table 3 presents the persona prompts dynamically generated for each agent based on the cultural dimension and culture source.

C.2 Instruction Prompts for Single-Agent Pipeline

Table 4 shows the complete instruction prompt used by the SingleShotAgent for refining all dimensions (person, action, location) simultaneously.

C.3 Instruction Prompts for Multi-Agent Pipelines

Tables 5, 6, 7, and 8 present the instruction prompts for the specialized agents used in multi-agent pipelines (MAS and MAP).

D VLM Evaluation Prompts

This appendix gives the verbatim prompts used for VLM-as-judge evaluation (Section 5.2). Tables 9, 10, and 11 present the prompts for cultural relevance, visual similarity, and text–image alignment, respectively.

Agent Type	Persona Template
PersonAgent	You are a {culture} individual who understands typical appearance traits from this culture.
ActionAgent	You are a {culture} observer skilled at describing how people typically eat, play music, and dance.
LocationAgent	You are a {culture} tour guide who knows how to visually describe iconic landmarks.
SingleShotAgent (mono-cultural)	You are someone familiar with {culture} cultural settings and can enrich any part of the scene.
SingleShotAgent (cross-cultural)	You are someone familiar with {culture1}, {culture2} and {culture3} cultural settings and can enrich any part of the scene.
FuseAgent	No culture-specific persona (fusion task only)

Table 3: Persona prompts for different agent types. The {culture} placeholder is dynamically replaced with the specific culture (Chinese, American, or Romanian) based on the dimension being refined. For cross-cultural prompts in SingleShotAgent, all three cultures are listed.

SingleShotAgent Instruction Prompt
Your task is to improve the full prompt by refining the person, action, and location. Focus on: • appearance details (e.g., facial features, hairstyle, clothing) • how the action is performed (e.g., body posture, hand movement, visible objects) • iconic aspects of the place (e.g., landscape, architecture, lighting, environmental elements) Do not introduce new people, actions, or locations. Return JSON with: "refined_prompt": the updated sentence "justification": a brief explanation of how your changes improve the visual clarity or cultural specificity of the scene.

Table 4: Instruction prompt for SingleShotAgent (SA) pipeline. The agent refines all three dimensions in a single pass.

PersonAgent Instruction Prompt
Your task is to improve only the appearance of the person in the prompt. Focus on visual traits like facial features, hairstyle, or clothing. Do not change the action or the location. Return JSON with: "refined_prompt": the updated sentence "justification": a brief explanation of how your addition makes the person more visually distinctive or culturally aligned.

Table 5: Instruction prompt for PersonAgent. This agent refines only the person dimension.

ActionAgent Instruction Prompt
Your task is to improve only the action portion of the prompt. Focus on how the action is performed: body posture, hand movement, or any visible objects. Do not change the person or the location. Return JSON with: "refined_prompt": the updated sentence "justification": a brief explanation of why your added details enhance the clarity, vividness, or cultural alignment of the action.

Table 6: Instruction prompt for ActionAgent. This agent refines only the action dimension.



Figure 8: Qualitative comparison for a mono-cultural example (“a Chinese person playing guzheng at the Potala Palace”). Five temporal frames are shown for each pipeline (base, SA, MAS, MAP). MAP achieves the highest cultural relevance (CRS: base 0.237, SA 0.242, MAS 0.249, MAP 0.271).

LocationAgent Instruction Prompt
Your task is to improve only the location part of the prompt. Add iconic visual details that make the place recognizable (e.g., landscape, architecture, lighting, environmental elements). Do not change the person or the action. Return JSON with: "refined_prompt": the updated sentence "justification": a brief explanation of how your addition makes the location more vivid or culturally recognizable.

Table 7: Instruction prompt for LocationAgent. This agent refines only the location dimension.

FuseAgent Instruction Prompt

Your task is to merge three versions of the same scene into one sentence.

Keep all appearance, action, and location details. Make the result vivid, coherent, and natural-sounding.

Do not add new people, actions, or places.

Return JSON with:
"refined_prompt": the fused sentence
"justification": a brief explanation of how you combined the three inputs (e.g., merged phrasing, reordered elements, kept best details).

Table 8: Instruction prompt for FuseAgent. This agent fuses the three independently refined prompts from PersonAgent, ActionAgent, and LocationAgent into one coherent prompt.

VLM Cultural Relevance Evaluation Prompt

You will be given one image and four culturally annotated sentences describing different aspects of it (overall scene, person, action, and location).

Your task is to evaluate how culturally aligned the image is with respect to each sentence — that is, whether the visual features in the image reflect the cultural cues or identities expressed in the text.

Specifically, consider appearance, clothing, architecture, symbols, traditions, or any other visual cues linked to a specific country or culture.

For each sentence, reason step by step and assign a score between 1 and 5, where:

1 = The image is not culturally relevant to the description
5 = The image is highly culturally relevant to the description

image: [image_path]
text_overall: "[overall cultural grounding statement]"
text_person: "[person cultural grounding statement]"
text_action: "[action cultural grounding statement]"
text_location: "[location cultural grounding statement]"

The output should be a JSON object ONLY with the following format:

```
{
  "overall_reasoning": "...",
  "overall_score": number,
  "person_reasoning": "...",
  "person_score": number,
  "action_reasoning": "...",
  "action_score": number,
  "location_reasoning": "...",
  "location_score": number
}
```

Table 9: VLM evaluation prompt for cultural relevance. The VLM evaluates how culturally aligned the video frame is with respect to four cultural grounding statements.

VLM Visual Similarity Evaluation Prompt

You will be shown two images. Your task is to evaluate how visually similar the second image is compared to the first image.

Consider all visible aspects, including the person (appearance, pose, clothing), the action (what is happening), the location (environment, background), and any cultural or stylistic details.

Please explain your reasoning step by step, and then assign a score between 1 and 5:

1 = Very different visually
5 = Nearly identical visually

image_1: (reference image)
image_2: (comparison image)

The output should be a JSON object ONLY with the following format:

```
{
  "reasoning": "...",
  "score": number
}
```

Table 10: VLM evaluation prompt for visual similarity. The VLM compares the visual similarity between base video frames and agent-refined video frames.

VLM Text-Image Alignment Evaluation Prompt

You will be shown one image and multiple sentences that describe what the image is expected to show.

Your task is to evaluate how well the visual content of the image aligns with each description.

Focus on whether the image clearly reflects the elements described — such as people, actions, locations, objects, or any culturally or visually specific features.

For each sentence, please explain your reasoning step by step, and assign a score between 1 and 5, where:

1 = The image does not match the description
5 = The image clearly and fully matches the description

Please evaluate all sentences and return your response as a JSON object with this exact format:

```
{
  "text1_reasoning": "...",
  "text1_score": number,
  "text2_reasoning": "...",
  "text2_score": number,
  "text3_reasoning": "...",
  "text3_score": number,
  "text4_reasoning": "...",
  "text4_score": number,
  "text5_reasoning": "...",
  "text5_score": number,
  "text6_reasoning": "...",
  "text6_score": number
}
```

Note: text1-text4 are the four cultural grounding statements (overall, person, action, location), text5 is the original prompt, and text6 is the final refined prompt.

Table 11: VLM evaluation prompt for text-image alignment. The VLM evaluates the alignment between video frames and multiple text descriptions including cultural grounding statements and prompts.