

Persona Prompting in Multimodal Urban Perception: Descriptive Convergence and Interpretive Variation

Neemias B da Silva^{1,2}, Matt Ratto¹, Myriam Delgado², Rodrigo Minetto²,
Daniel Silver¹, Thiago H Silva^{1,2}

¹ University of Toronto, Toronto, Canada

² Universidade Tecnológica Federal do Parana, Curitiba, Brazil

Correspondence: neemias.buceli@mail.utoronto.ca

Abstract

This study examines how persona prompting shapes language generated by two multimodal large language models in urban perception, a setting for examining subjective interpretations of shared visual evidence. We organize outputs into three functional layers: *descriptive grounding* (captions), *intermediate semantic layer* (perception tags), and *interpretive framing* (justifications). Using approximately 60,000 persona-conditioned annotations from each of two MLLMs, Qwen3-VL and Gemma4, we find that captions converge strongly across persona profiles and show only small attribute-associated differences. Justifications vary substantially more: economic status produces the largest difference in both models, with political orientation and personality also prominent. Paired image-level comparisons confirm larger justification than caption differences for these three attributes. For perception tags, personas sharing the same attribute level produce more similar tag sets than personas with different attribute levels, with the largest separation observed for economic status. Exploratory topic analysis further suggests persona-specific evaluative emphasis. Across models, profile-pair similarity patterns are strongly correlated for all three output types, although agreement is lowest for justifications. Overall, persona prompting affects interpretive framing more strongly than descriptive grounding.

1 Introduction

Large language models (LLMs) are increasingly used as *social simulators*, with persona prompts intended to approximate diverse perspectives (Filippas et al., 2024; Aher et al., 2023; Argyle et al., 2023). Applications include social-bias analysis and synthetic population simulation, raising questions about whether persona-conditioned models can represent collective perception and behavior (Park et al., 2024; Li et al., 2025). Although persona prompting may introduce pluralism, differ-

ences among persona-conditioned outputs need not reflect meaningful perspectives; they may instead stem from superficial linguistic variation or stereotypical model associations (Boelaert et al., 2025).

Most prior work evaluates persona prompting through structured outputs, such as labels, choices, or task performance (Hu and Collier, 2024; Beck et al., 2024; da Silva et al., 2026). Less is known about how personas shape generated language, particularly descriptions and explanations. This distinction is especially relevant in multimodal settings, where visually grounded descriptions and subjective interpretations may respond differently to persona conditioning. Urban perception provides a useful setting: observers evaluate qualities such as safety, order, upkeep, beauty, and apparent socioeconomic status from visual cues, and these judgments vary across individuals and social contexts (Shen et al., 2026; Lopes et al., 2023; da Silva et al., 2025).

Despite growing evidence that persona conditioning can alter model behavior, an important question remains: *where* do these effects emerge in multimodal generation? Differences between persona-conditioned outputs may reflect changes in what the model describes, in the semantic concepts it associates with visual cues, or in how it interprets and evaluates the same evidence. Distinguishing these possibilities is important because differences in generated language may arise from description, semantic selection, or interpretation. We therefore examine persona effects across three complementary forms of output: descriptions of observable content, semantic perception tags associated with visual cues, and sentiment-linked justifications that interpret and evaluate the same visual evidence.

In our setting, a model annotates each urban image with a sentiment label on a five-point scale (from Negative to Positive), a caption, a set of perception tags, and a justification. We organize these outputs into three functional layers (Figure 1): *de-*

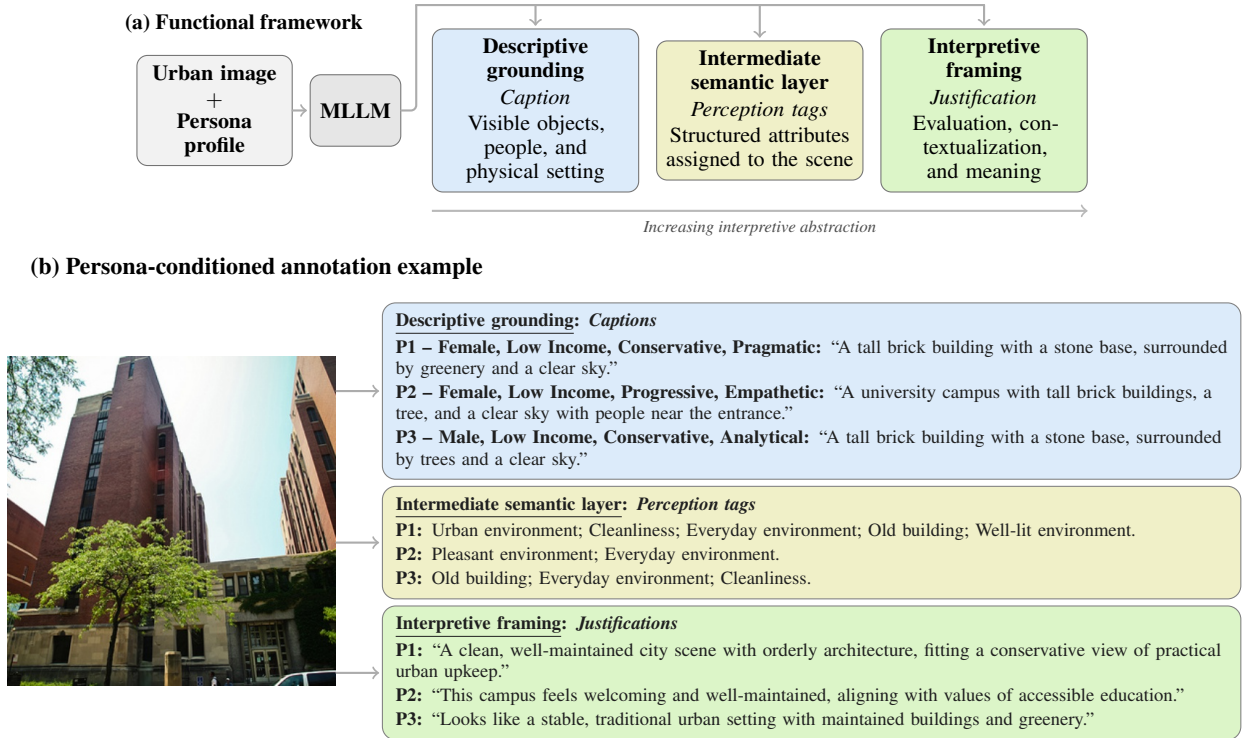


Figure 1: Functional framework and persona-conditioned example. Panel (a) distinguishes descriptive grounding, intermediate semantic selection, and interpretive framing. Panel (b) groups three outputs for the same image by type: captions converge on visible content, whereas perception tags and justifications vary more in semantic emphasis and evaluative framing.

scriptive grounding, represented by captions of visible content; an *intermediate semantic layer*, represented by perception tags; and *interpretive framing*, represented by justifications that explain the assigned sentiment through the evaluation and contextualization of the scene. The three layers leave different room for a persona to shape the output: captions are anchored in visual evidence shared by all observers, perception tags require selecting labels from a controlled vocabulary, and justifications are open-ended and explicitly evaluative. If persona conditioning acts mainly on interpretation, its effects should be more apparent in perception tags and justifications than in captions. Persona profiles may therefore produce similar descriptions of the same visual evidence while emphasizing different interpretations or evaluative aspects. Our central finding is that persona conditioning primarily affects interpretive framing, while descriptive grounding remains comparatively stable.

We examine this distinction using approximately 60,000 persona-conditioned annotations from each of two MLLMs: Qwen3-VL and Gemma4 (Section 3). For each model, 1,200 persona-conditioned agent instances annotate the same urban scenes under matched prompts, personas, and decoding

settings, alongside no-persona baselines.

Our contributions are fourfold: (i) we introduce a functional framework distinguishing descriptive grounding, an intermediate semantic layer, and interpretive framing; (ii) we show that persona-associated variation is substantially stronger in interpretive than in descriptive language, supported by paired image-level comparisons of captions and justifications on a common cosine-similarity scale; (iii) we identify economic status as the most consistent source of interpretive variation, with political orientation, personality, and gender exhibiting smaller or model-dependent patterns; and (iv) we assess cross-model consistency and test whether the economic status pattern persists under deterministic decoding. Code, prompts, configurations, and processed outputs are publicly available¹.

2 Related Work

Prior work shows that LLMs can reproduce stylized experimental findings and simulate decision-making processes (Aher et al., 2023; Filippas et al., 2024), supporting applications in surveys, experiments, and content analysis (Bail, 2024). However,

¹<https://minds-lab-utfpr.github.io/Persona-Interpretive-Analysis>

important limitations remain, including bias, limited validity, and difficulties in representing human diversity (Wang et al., 2024, 2025).

A related line of research investigates LLMs as synthetic respondents and simulated populations. Studies show that LLMs can approximate human-like responses and support large-scale agent simulations (Argyle et al., 2023; Park et al., 2024), while also revealing limitations in modeling belief structures and reasoning processes (Li et al., 2025; Barrie and Cerina, 2026). These findings suggest that evaluation should extend beyond predictive accuracy to include consistency and interpretability.

Persona prompting plays a central role in this context. Conditioning models on demographic or psychological attributes can influence outputs, although effects are often limited or task-dependent (Hu and Collier, 2024; Beck et al., 2024). Most related work evaluates persona effects on structured outputs — labels, ratings, or choices — leaving open the question of whether and how personas shape the language of open-ended responses (Lutz et al., 2025; Malik et al., 2025).

Our work also connects to research on multimodal LLMs and urban perception. Prior studies show that urban image perception is subjective and socially contextualized, with perceptions of safety, disorder, beauty, and wealth varying systematically across observers and environments (He et al., 2026; Balsa-Barreiro et al., 2026; Lopes et al., 2023; Xu et al., 2025; Oliveira et al., 2020). Recent work further shows that urban visual perception and its relationship with sentiment vary across socioeconomic contexts, with differences observed between lower- and higher-income communities (Shen et al., 2026). Multimodal LLMs can partially capture some of these perceptual dimensions (Wu et al., 2026; da Silva et al., 2025). At the same time, recent work reports strong within-persona consistency but limited cross-persona variation in outputs such as sentiment labels (da Silva et al., 2026), raising the question of whether persona effects emerge more clearly in interpretive language than in structured outputs.

We address this gap directly, analyzing captions, justifications, and perception tags generated by MLLMs to examine how persona effects emerge in descriptive grounding and interpretive framing. The key question is whether persona conditioning produces structured variation in how scenes are interpreted, even when descriptive content converges.

3 Data and Methodology

3.1 Dataset and Conditions

We analyze two datasets. The first is a publicly available *Qwen3-VL* annotation corpus (da Silva et al., 2026), comprising 60,208 records generated under persona-conditioned and no-persona settings for 50 urban scene images. Each record contains four outputs: (i) a predicted sentiment label on a five-point scale (Negative, Slightly Negative, Neutral, Slightly Positive, Positive), (ii) a caption describing the image, (iii) one or more perception tags (the prompt requests one to five) selected from a controlled vocabulary of 593 labels, and (iv) a justification explaining the assigned sentiment. Captions and justifications contain 16.9 ± 5.0 and 18.0 ± 4.9 words on average, respectively. Both persona-conditioned and no-persona justifications are written as full prose explanations rather than simple label rationales, enabling linguistic analysis beyond sentiment labels. Figure 1b illustrates one annotation record per persona for the same urban image.

All annotations were generated under matched decoding settings (temperature=0.1, seed=42), following the configuration documented in the accompanying code repository and prior dataset description (da Silva et al., 2026). We conduct a targeted decoding-sensitivity ablation at temperature $T = 0$ for economic status (Appendix B). We use the publicly available dataset and construct a second dataset with *Gemma4* using the same images, prompts, persona profiles, and decoding settings. The prompt templates are given in Appendix H. The two models are served differently: *Qwen3-VL* is *Qwen3-VL-8B* run through Ollama (tag *qwen3-v1:8b*, *Q4_K_M* quantization), whereas *Gemma4* is *google/gemma-4-E4B-it* run through Hugging Face transformers with *bf16* weights. This matched design reduces, though does not eliminate, differences attributable to the experimental configuration, allowing us to compare persona-conditioned patterns across the two models. For each model, we consider three annotation conditions: persona-conditioned generation, no-persona generation with extended reasoning, and no-persona generation without extended reasoning.

- **Persona.** The main corpus encompasses annotations produced using $|P| = 24$ persona profiles obtained from the Cartesian product of $|D| = 4$ demographic and personal-

ity attributes: *gender* (Male, Female), *economic status* (Low, High), *political orientation* (Conservative, Progressive), and *personality* (Pragmatic, Empathetic, Analytical). Each profile is instantiated as $|A| = 50$ persona-conditioned agent instances ($24 \times 50 = 1,200$ in total), and each instance annotates all ($|I| = 50$) images using persona-conditioned prompts with extended reasoning enabled (`think=True`). We use the term *persona-conditioned agent instance* to denote a generation replicate under a fixed persona profile and decoding configuration. These instances do not interact, maintain memory, or act autonomously; each annotation is generated through a separate inference call under the same decoding configuration. The full list of profiles is shown along the rows of Figure A1 (Appendix A).

- **No-Persona.** A no-persona setting, in which the same model annotates the same images using a neutral observer prompt, includes two variants: with extended reasoning (**NPT**, `think=True`) and without (**NPNOT**, `think=False`). The two variants produce five annotations per image (250 each), providing complementary baselines for the persona-generated corpus.

Table 1 summarizes all annotations per model and experiment condition.

Model	Experiment Condition			Missing	Total Observed
	Persona	NPT	NPNOT		
Qwen3-VL	60,000	250	250	292	60,208
Gemma4	60,000	250	250	1	60,499

Table 1: Expected and observed numbers of annotations by model. The expected total is 60,500 annotations per model. Missing records correspond to failed generations relative to this total. No missing records occurred in the NPT or NPNOT cases.

In the following subsections, we evaluate within-persona stability and cross-persona divergence separately for captions, justifications, and perception tags. For each matrix, we compute pairwise similarities between agent annotations, comparing agents that share the same persona profile or attribute level with those that do not, depending on the analysis.

We summarize diagonal separation using the Diagonal Strength Index (DSI) to quantify how much stronger the diagonal elements of a matrix (representing similarities between agents from the

same persona profiles) are compared with the off-diagonal elements (representing agents from different persona profiles): $DSI = \frac{\mu_D - \mu_O}{\mu_D}$, where μ_D denotes the mean of the diagonal elements; μ_O is the mean of the off-diagonal elements.

3.2 Evaluating Within-Persona Stability

We first assess the internal stability of agents sharing the same persona, i.e., the reproducibility of generated outputs when all demographic and personality attributes are held constant. We evaluate stability across three output types:

- **Captions (descriptive grounding) and justifications (interpretive framing):** Captions and justifications are encoded using the Sentence-BERT model `all-MiniLM-L6-v2`, and semantic similarity is computed as the cosine similarity between the resulting sentence embeddings.
- **Perception tags (intermediate semantic layer):** Each annotation contains a set of perception tags drawn from a controlled vocabulary. These tags form an intermediate semantic layer between descriptive grounding and interpretive framing, capturing structured aspects of scene interpretation. Because the tags are discrete labels rather than unconstrained text, we measure exact set overlap using Jaccard similarity, defined as the intersection over the union of the two tag sets. Records with an empty tag list (13 in Qwen3-VL, none in Gemma4) are treated as missing and excluded from perception pairs rather than scored as zero overlap.

For each type (captions, justifications, or perception tags), we compute a distribution of within-persona similarities and summarize it for each profile using the mean. Higher values indicate greater convergence among outputs generated under the same persona profile. Topic modeling was retained only for justifications because captions showed high cross-agent similarity and insufficient semantic variation to support meaningful thematic discrimination.

3.3 Evaluating Cross-Persona Divergence

We evaluate whether outputs differ systematically across four persona dimensions D : gender, economic status, political orientation, and personality.

For each image $i \in I$ and dimension $d \in D$, annotation pairs are partitioned into:

- **Within-group pairs** $\odot$: agents sharing the same attribute level (e.g., two high-income agents), regardless of the remaining persona dimensions.
- **Cross-group pairs** $\otimes$: agents with different attribute levels (e.g., high-income vs. low-income), regardless of the remaining persona dimensions. For personality, which has three levels, this set pools the three possible level contrasts.

As a sensitivity analysis, we also compare matched profiles that differ only on the target dimension (Appendix D).

We compare groups across complementary output types:

- **Captions (descriptive grounding) and justifications (interpretive framing)**: We compute pairwise cosine similarity between sentence embeddings, as described in Section 3.2.
- **Perception tags (intermediate semantic layer)**: We compute Jaccard similarity between tag sets, as described in Section 3.2.
- **Topic distributions (interpretive framing)**: For each image and persona dimension, we construct topic distributions for each attribute level after excluding unassigned outlier documents (i.e., BERTopic topic -1). We compute Jensen–Shannon divergence (JSD) between attribute-level pairs—one pair for each binary dimension and three pairs for personality—and average the values across pairs and images. Significance is assessed using 1,000 within-image label permutations that preserve group sizes. Heatmaps show the most frequent topic proportions across personas.

For captions, justifications, and perception tags, we take the image as the unit of analysis. For each image i and persona dimension d , we compute the mean similarity among annotation pairs sharing the same attribute level, $\bar{s}_{i,d,\odot}$ and among pairs with different attribute levels, $\bar{s}_{i,d,\otimes}$. The paired image-level difference is set as $\Delta_{i,d} = \bar{s}_{i,d,\odot} - \bar{s}_{i,d,\otimes}$. Positive values indicate that annotations sharing the target attribute level are, on average, more similar than annotations with different attribute levels. We

Model	Diagonal Strength Index (%)		
	Captions	Justifications	Perception
Qwen3-VL	2.3	14.9	11.7
Gemma4	3.7	30.0	25.1

Table 2: DSI (%) for both models and output types.

denote $\Delta_d = \frac{1}{|I|} \sum_{i \in I} \Delta_{i,d}$ as the mean difference across images.

Because each image contributes a within-group and a cross-group mean, we conduct a separate paired analysis for each model, output type, and persona dimension. For image i , the paired observations are $\bar{s}_{i,d,\odot}$ and $\bar{s}_{i,d,\otimes}$. Equivalently, we apply a two-sided Wilcoxon signed-rank test to the 50 image-level differences $\{\Delta_{i,d} : i \in I\}$, testing whether their location is zero. We report Δ_d , the mean difference across images, with a 95% bias-corrected and accelerated (BCa) bootstrap confidence interval obtained by resampling images. Benjamini–Hochberg correction is applied across the 12 tests within each model. Because the signed-rank test and Δ_d summarize different aspects of the distribution, interpretation emphasizes Δ_d and its confidence interval, with statistical significance as complementary evidence. Full statistics, effect sizes, sensitivity analyses, and complete-case results appear in Appendix C.

To compare descriptive and interpretive text directly, we conduct a second paired analysis. For each model and persona dimension d , we apply the same procedure to the 50 image-level contrasts $\Delta_{i,d}^{\text{just}} - \Delta_{i,d}^{\text{cap}}$, pairing justification and caption differences from the same image. Benjamini–Hochberg correction is applied across the four persona dimensions within each model. Perception-tag differences use Jaccard similarity and are therefore analyzed separately.

For topic distributions, a larger Jensen–Shannon divergence indicates greater variation in the themes emphasized across persona groups.

4 Results

4.1 Persona Effects from a Profile Perspective

For each profile pair (p_x, p_y) , similarity is computed within each image and then averaged across shared images, yielding a 24×24 inter-profile similarity matrix for each model (Appendix A).

Table 2 shows DSI (%) values for both models and the three types of outputs. It reveals a consistent pattern across the models. Caption generation exhibits low DSI values (2.3–3.7%), in-

Type	Pearson r	Spearman ρ
Captions (cosine)	0.89	0.87
Justifications (cosine)	0.80	0.78
Perception tags (Jaccard)	0.88	0.83

Table 3: Cross-model agreement between Qwen3-VL and Gemma4, measured by correlating the upper triangles of their image-conditioned inter-profile similarity matrices. Higher correlations indicate more similar profile-pair rankings across models.

indicating weak profile-associated separation in descriptive content. In contrast, DSI is higher for perception tags (11.7–25.1%) and highest for justifications (14.9–30.0%), indicating stronger profile-associated structure in interpretive outputs. Across all three output types, Gemma4 produces higher DSI values than Qwen3-VL, indicating stronger separation between within-profile and cross-profile similarity. In both models, captions show the lowest DSI, whereas perception tags and justifications show stronger profile-associated separation.

Table 3 summarizes cross-model agreement between the resulting matrices. It shows strong agreement across all output types. Captions show the highest cross-model agreement, whereas justifications show the lowest.

4.2 Persona Effects from a Dimensional Perspective

We now examine how similarity is associated with individual persona dimensions across three output types: captions (*descriptive grounding*), justifications (*interpretive framing*), and perception tags (*intermediate semantic layer*). Building on the convergence observed in Section 4.1, this analysis asks whether semantic similarity is systematically structured by persona attributes.

Figure 2 compares within-group and cross-group similarity across four persona dimensions. Captions and justifications are measured using cosine similarity, whereas perception tags are measured using Jaccard similarity.

For captions, within-group similarity exceeds cross-group similarity for every persona dimension in both models, but the differences are small: Δ_d ranges from +0.002 for gender in Qwen3-VL to +0.021 for economic status in Gemma4, while mean caption similarities remain approximately 0.87–0.91. Descriptive grounding, therefore, exhibits detectable but limited attribute-associated variation, with captions remaining highly similar across persona groups.

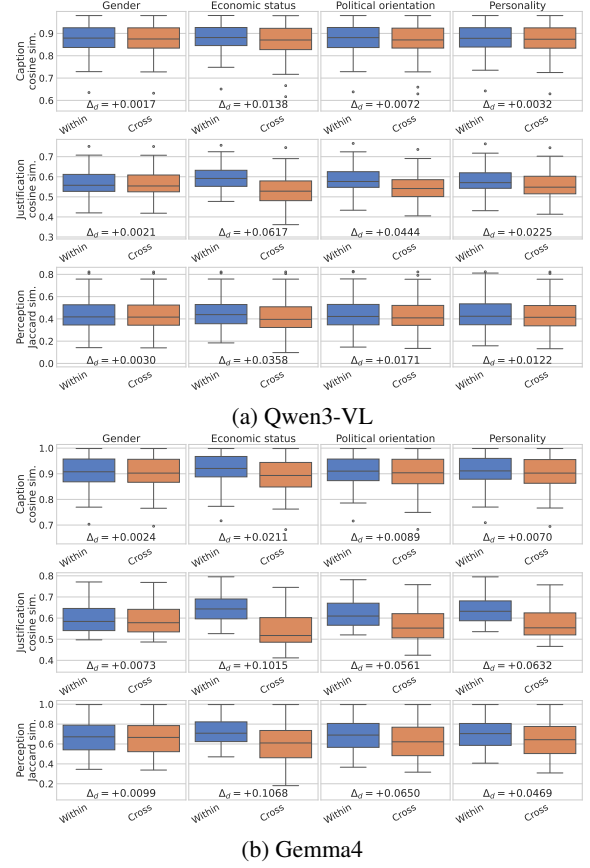


Figure 2: Within-group (⊙) vs. cross-group (⊗) similarity for captions (top row), justifications (middle row), and perception tags (bottom row) across persona dimensions. The mean image-level difference Δ_d is shown beneath each pair of boxplots. Captions and justifications use cosine similarity, whereas perception tags use Jaccard similarity. Full confidence intervals, effect sizes, and adjusted tests are reported in Appendix C.

Justifications show larger attribute-associated differences. Economic status produces the largest difference in both models (Qwen3-VL: $\Delta_d = +0.062$; Gemma4: $\Delta_d = +0.101$), followed by political orientation in Qwen3-VL ($\Delta_d = +0.044$). In Gemma4, personality ($\Delta_d = +0.063$) and political orientation ($\Delta_d = +0.056$) are also prominent. Gender produces the smallest justification difference in both models (Qwen3-VL: $\Delta_d = +0.002$; Gemma4: $\Delta_d = +0.007$).

Direct paired contrasts on the common cosine scale ($\Delta_{i,d}^{\text{just}} - \Delta_{i,d}^{\text{cap}}$) show that justification differences exceed caption differences for economic status, political orientation, and personality. The mean image-level contrasts ($\Delta_d^{\text{just}} - \Delta_d^{\text{cap}}$) range from +0.019 to +0.048 in Qwen3-VL and from +0.047 to +0.080 in Gemma4. Gender again produces the smallest contrast; for Qwen3-VL, its mean con-

trast is +0.0004, and the BCa confidence interval includes zero. Full paired results are reported in Appendix C.

For perception tags, economic status produces the largest within–cross difference in both models (Qwen3-VL: $\Delta_d = +0.036$; Gemma4: $\Delta_d = +0.107$), whereas gender produces the smallest (+0.003 and +0.010, respectively). Political orientation and personality fall between these extremes. Gemma4 shows larger Jaccard differences than Qwen3-VL across all dimensions; these values are interpreted separately from the cosine-based ones.

These results locate persona-associated variation primarily in interpretive language. Justifications exhibit substantially larger within–cross differences than captions, particularly for economic status and political orientation, while caption differences remain small. Perception tags also show greater within-group than cross-group similarity across all dimensions.

Within each output type, Gemma4 generally exhibits larger within–cross differences than Qwen3-VL, although economic status remains the largest dimension and gender the smallest in both models. This pattern reinforces the central distinction of the paper: *persona conditioning affects interpretive framing more strongly than descriptive grounding*. The economic status pattern also persists under deterministic decoding ($T = 0$; Appendix B).

A matched-factorial sensitivity analysis that holds the remaining persona attributes fixed preserves the broad dimensional ordering (Appendix D), while a paired justification-minus-caption contrast within the matched design confirms larger differences for all four attributes (Appendix E).

Additional comparisons with no-persona generations show similarly strong descriptive convergence but greater variability in justificatory language; full results are reported in Appendix F.

4.3 Persona Effects on Topic Structure

We further explore interpretive framing through the thematic content of justifications. We focus on justifications because captions yielded semantically uniform and unstable BERTopic results, with insufficient variation to produce differentiated topics. The ten most prevalent justification topics include evaluative themes related to natural beauty, development, accident response, tranquility, decay, and neglect. Both models are analyzed using the same BERTopic pipeline. Topic nicknames are manu-

Persona dimension	Qwen3-VL	Gemma4
Economic status	0.311	0.718
Political orientation	0.281	0.557
Personality	0.095	0.579
Gender	0.026	0.172

Table 4: Between-group Jensen–Shannon divergence of justification-topic distributions, averaged over group pairs and images. Higher values indicate greater thematic divergence within a model. Absolute values are not directly comparable across models because their BERTopic models were fitted independently.

ally assigned by the authors after inspecting the highest-weight c-TF-IDF terms and BERTopic’s representative documents for each topic.

To quantify how strongly persona dimensions structure thematic content, Table 4 reports the between-group JSD of justification-topic distributions. Within both models, economic status produces the largest divergence and gender the smallest. Political orientation is the second-largest dimension for Qwen3-VL, whereas personality and political orientation have similar magnitudes in Gemma4. All dimensions reach significance under the within-image permutation test ($p_{\text{emp}} \leq 0.001$), but substantive interpretation emphasizes JSD magnitude rather than significance alone.

Figure 3 shows differences in a profile-level view. For Qwen3-VL, most visible contrasts occur across economic status and political orientation profiles, particularly for topics related to neglect, development, order, and natural beauty. Gender and personality profiles appear comparatively uniform, consistent with their smaller JSD values.

Some Qwen3-VL topics, including *Historical & preserved* and *City & nature*, remain relatively stable across profiles. Others, such as *Typical urban scene* and *Natural landscape & beauty*, vary more visibly. This suggests that persona conditioning changes which evaluative aspects of the same scene become thematically salient.

Gemma4 exhibits economic status variation, although the direction is topic-specific rather than uniformly negative or positive. In its heatmap, accident- and neglect-related topics appear more often in some low-income profiles, and peaceful, natural, and architectural themes in selected high-income profiles. These are profile-level patterns over the ten displayed topics rather than a tested directional effect of economic status.

As shown in Figure G1, Appendix G, both models show consistent alignment between justification topics and sentiment, with accident, conflict, ne-

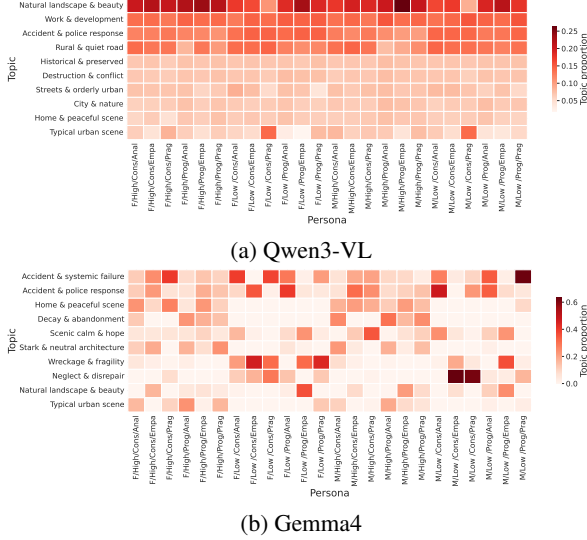


Figure 3: Column-normalized proportions of the ten most prevalent assigned justification topics across persona profiles. BERTopic was fitted independently for Qwen3-VL and Gemma4; topic labels, proportions, and color intensities should therefore be interpreted within each panel rather than compared directly across models.

glect, and decay topics strongly associated with negative sentiment, and nature, beauty, and peaceful scenes with positive sentiment. In contrast, everyday urban infrastructure and typical city scenes are predominantly neutral or mixed, reflecting the ambiguity of ordinary urban environments.

These patterns require cautious interpretation. BERTopic assigned topic -1 to 15% of Qwen3-VL justifications and 57% of Gemma4 justifications. Because the analysis conditions on assigned topics, and the displayed topics cover different proportions of the two corpora, absolute JSD values and heatmap contrasts should not be compared directly across models.

The sharper Gemma4 contrasts therefore do not by themselves indicate a stronger persona effect. The more reliable cross-model result is the within-model ordering of dimensions: economic status produces the largest thematic divergence and gender the smallest.

Overall, the topic analysis complements the similarity results: persona conditioning changes the thematic emphasis of justifications, supporting the distinction between shared descriptive grounding and variable interpretive framing.

5 Conclusion

This work examined how persona prompting shapes language generated by Qwen3-VL and

Gemma4 for urban perception. We introduced a functional framework distinguishing descriptive grounding, an intermediate semantic layer, and interpretive framing.

Across roughly 60,000 persona-conditioned annotations per model, captions remained highly similar across profiles, whereas justifications showed substantially greater variation, particularly for economic status, political orientation, and personality. Economic status produced the largest differences and gender the smallest, an ordering broadly preserved by the matched-factorial analysis. Perception tags likewise showed greater within-group than cross-group similarity across attributes.

The prominence of economic status is consistent with recent urban-perception research reporting socioeconomic differences in visual perception and its relationship with sentiment (Shen et al., 2026). This parallel does not establish that persona-conditioned outputs reproduce human income-group differences, but highlights model sensitivity to economic status prompting.

Gemma4 generally exhibited larger persona-associated differences than Qwen3-VL, although their magnitudes may also reflect differences in serving stack and precision. Nevertheless, the shared dimensional ordering and strong cross-model agreement indicate a consistent qualitative pattern whose magnitude varies across the evaluated configurations.

Overall, structured labels and descriptive outputs may understate persona-conditioned variation, which emerges most clearly in open-ended explanations. Persona-conditioned MLLMs should therefore be treated as generators of prompted interpretive perspectives rather than simulations of human demographic groups. Future work should validate these patterns against demographically matched human annotations and across additional models, languages, and visual domains.

Limitations and Ethics Considerations

Our factorial design represents personas through a controlled set of demographic and personality attributes. Although this enables attribute-level comparisons, the profiles are prompting conditions rather than comprehensive representations of individual identities. We therefore interpret the observed differences as persona-conditioned model behavior, not as estimates of how corresponding human groups perceive urban scenes. Establishing

such correspondence would require demographically matched human annotations.

Similarly, the 50 instances generated for each persona profile should not be interpreted as 50 independent or autonomous agents. They are repeated model inferences under the same persona-conditioned configuration and are used to characterize variability across generations. Differences among them may therefore reflect decoding stochasticity rather than systematic instance-level variation. Future work could introduce persistent identities or state when modeling distinct simulated individuals.

Using the same 50 images for both models enables a controlled comparison but limits the generality of the findings, which are specific to the evaluated models, prompts, personas, and image set. Future work should assess their robustness across additional configurations, models, and datasets.

Several design choices qualify our comparisons. First, the prompt asks for captions to be written independently of the persona, whereas justifications are produced in the persona’s own voice (Appendix H). The observed gap between descriptive and interpretive outputs may therefore partly reflect this task design, although captions still exhibit small but systematic attribute-associated differences. Second, justifications may repeat or paraphrase persona cues from the prompt, so some within-group similarity may reflect shared lexical content rather than interpretive framing alone. Future work should test this possibility by masking persona-related terms before recomputing Δ_d and by controlling for output length. Third, the models differ in architecture, serving stack, and weight precision; cross-model differences should therefore be interpreted as differences between the evaluated configurations rather than architecture alone. Finally, the topic analysis is exploratory: BERTopic leaves 57% of Gemma4 justifications unassigned. A shared topic model and directional analyses of topic and sentiment distributions by attribute level remain important extensions.

Synthetic personas may reproduce biased or stereotypical associations. To support scrutiny of these risks, we release the prompts, configurations, and analysis code for replication and auditing.

Acknowledgments

This study was supported in part by the National Council for Scientific and Technological

Development - CNPq (processes 441444/2023-7, 403646/2026-0, and 444724/2024-9) and INCT TILD-IAR (proc. 408490/2024-1). This research was also supported in part by a CIFAR AI Catalyst Grant, “Towards Socially Grounded AI Safety: Integrating Causal and Institutional Reasoning in Language Models” (PI: Matt Ratto; Co-PI Zhijing Jin, University of Toronto).

References

- Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate multiple humans and replicate human subject studies](#). In *Proc. of ICML*, Honolulu, Hawaii, USA. JMLR.org.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337–351.
- Christopher A. Bail. 2024. [Can generative ai improve social science?](#) *PNAS*, 121(21):e2314021121.
- Jose Balsa-Barreiro, Shahid Rabbani, Djellel Difallah, Monica Menendez, and Manuel Cebrian. 2026. [A large-scale llm-generated dataset for exploring social interactions and urban experiences across 21 global cities](#). *Discover Data*, 4(1):5.
- Christopher Barrie and Roberto Cerina. 2026. [Synthetic personas distort the structure of human belief systems](#). *OSF: osf.io/preprints/socarxiv/n7fq8_v1*.
- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. [Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting](#). In *Proc. of EACL*, pages 2589–2615, St. Julian’s, Malta.
- Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. 2025. [Machine bias. how do generative language models answer opinion polls?](#) *Sociological Methods & Research*, 54(3):1156–1196.
- Neemias B. da Silva, John Harrison, Rodrigo Minetto, Myriam R. Delgado, Bogdan T. Nassu, and Thiago H. Silva. 2025. [Multimodal llms see sentiment](#). *ArXiv: https://arxiv.org/abs/2508.16873*.
- Neemias B da Silva, Rodrigo Minetto, Daniel Silver, and Thiago H Silva. 2026. Stable behavior, limited variation: Persona validity in LLM agents for urban sentiment perception. In *Proc. of IEEE DCOSS-IoT-UrbCom*, Reykjavik, Iceland.
- Apostolos Filippas, John J. Horton, and Benjamin S. Manning. 2024. [Large language models as simulated economic agents: What can we learn from homo silicus?](#) In *Proc. of EC*, page 614–615, New Haven, CT, USA. ACM.

- Jun He, Yi Lin, Zilong Huang, Jiacong Yin, Junyan Ye, Yuchuan Zhou, Weijia Li, and Xiang Zhang. 2026. [Urbanfeel: A comprehensive benchmark for temporal and perceptual understanding of city scenes through human perspective](#). In *Proc. of ICLR*, Rio de Janeiro, Brazil.
- Tiancheng Hu and Nigel Collier. 2024. [Quantifying the persona effect in LLM simulations](#). In *Proc. of ACL*, pages 10289–10307, Bangkok, Thailand. ACL.
- Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Liang, Luis Alonso, and Kent Larson. 2025. [Simulating society requires simulating thought](#). *Preprint*, arXiv:2506.06958.
- Cesar Rafael Lopes, Rodrigo Minetto, Myriam Regattieri Delgado, and Thiago H Silva. 2023. [Perceptsent - exploring subjectivity in a novel dataset for visual sentiment analysis](#). *IEEE Transactions on Affective Computing*, 14(3):1817–1831.
- Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. [The prompt makes the person\(a\): A systematic evaluation of sociodemographic persona prompting for large language models](#). In *Proc. of EMNLP*, pages 23212–23237, Suzhou, China. ACL.
- Ananya Malik, Nazanin Sabri, Melissa M. Karnaze, and Mai ElSherief. 2025. [Are LLMs empathetic to all? investigating the influence of multi-demographic personas on a model’s empathy](#). In *Proc. of EMNLP*, pages 24938–24959, Suzhou, China. ACL.
- Wyverson Bonasoli de Oliveira, Leyza Baldo Dorini, Rodrigo Minetto, and Thiago H. Silva. 2020. [Outdoorsent: Sentiment analysis of urban outdoor images by using semantic and deep features](#). *ACM Trans. Inf. Syst.*, 38(3).
- Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. [Generative agent simulations of 1,000 people](#). *arXiv preprint arXiv:2411.10109*.
- Ketong Shen, Jian Liu, and Xintao Liu. 2026. [Social inequality in sight: Exploring urban visual perception and sentiment across income levels](#). *Transactions in GIS*, 30(2).
- Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. [Large language models that replace human participants can harmfully misportray and flatten identity groups](#). *Nat Mach Intell*, 7(3):400–411.
- Pengda Wang, Huiqi Zou, Zihan Yan, Feng Guo, Tianjun Sun, Ziang Xiao, and Bo Zhang. 2024. [Not yet: Large language models cannot replace human respondents for psychometric research](#). *OSF: osf.io/preprints/osf/rwy9b_v1*.
- Songtai Wu, Wenbing Wang, Chengzhi Zhang, Qisheng Zeng, Haiying Wang, Jinyao Lin, and Shaoying Li. 2026. [Exploring multimodal large language models’ potential in simulating human perception of urban cycling environments: A street-view perspective](#). *Information Geography*, 2(1):100037.
- Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. 2025. [Flame: Learning to navigate with multimodal llm in urban environments](#). In *Proc. of AAAI*, volume 39, page 9005–9013, Philadelphia, USA.

A Inter-Profile Similarity

This section reports the image-conditioned 24×24 similarity matrices for Qwen3-VL and Gemma4, for captions (Figure A1) and justifications (Figure A2), both under cosine scale, and for perceptions (Figure A3) under Jaccard scale.

In each matrix, cell (p_x, p_y) is the mean (cosine or Jaccard) similarity between profiles p_x and p_y , computed on the same image and averaged over shared images. Note that the similarity of the perceptions is the *Jaccard* index (intersection over union of the set of controlled-vocabulary tags), so the values lie on a different scale from the cosine heatmaps for the caption and justification cases.

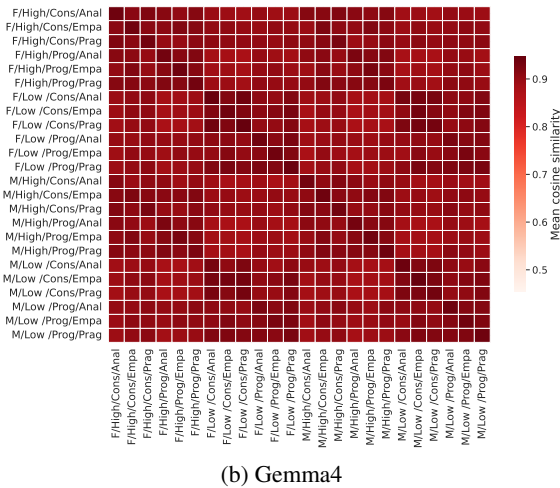
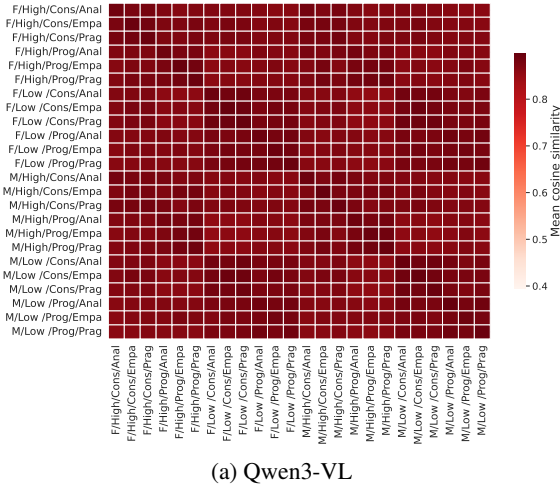


Figure A1: Inter-profile *caption* (descriptive grounding) similarity. Notice that, for both models (Qwen3-VL(A1a) and Gemma4(A1b)), persona conditioning has only a marginal effect on captions ($DSI = 2.3\%$ and 3.7% , respectively), yielding highly correlated profile-level similarity patterns. The two models agree strongly on this structure (Pearson $r = 0.89$; Table 3).

as captured by captions (Figure A1). For both models, caption similarity spans a narrow range across all cells, including the diagonal (0.85 – 0.90 for Qwen3-VL and 0.87 – 0.95 for Gemma4), with no discernible block structure. For the similarity matrix depicted in Figure A1(a) (Qwen3-VL model), $DSI = \frac{\mu_D - \mu_O}{\mu_D} = \frac{0.8928 - 0.8719}{0.8928} = 0.0234 \approx 2.3\%$ and for the similarity matrix depicted in Figure A1(b) (Gemma4 model), $DSI = 0.0368 \approx 3.7\%$. This indicates that agents sharing the same profile produce captions that are only marginally more similar than those from different profiles, confirming that descriptive content is largely invariant to persona conditioning in both models. The two caption matrices are strongly correlated (Pearson $r = 0.89$; Table 3).

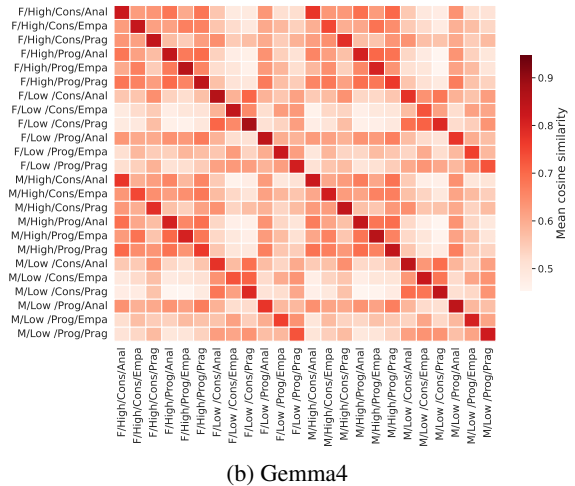
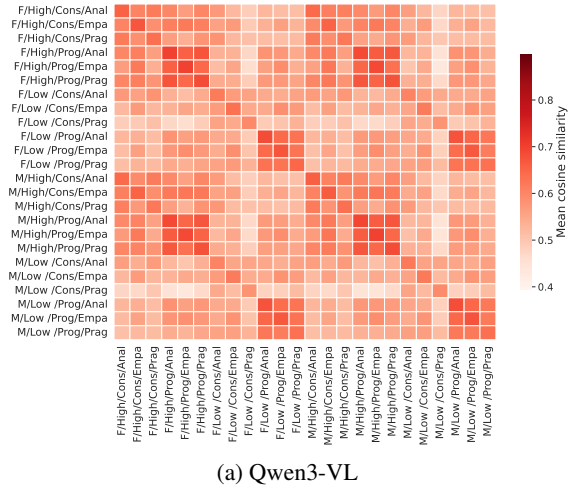


Figure A2: Inter-profile *justification* (interpretive framing) similarity. Both models show a wide range and high-valued diagonal (within-profile) cells: $DSI = 14.9\%$ for Qwen3-VL and 30.0% for Gemma4 which exhibits a more pronounced structure, yet the two patterns remain strongly correlated (Pearson $r = 0.80$).

We begin by examining *descriptive grounding*,

Figure A2 shows that justification similarity

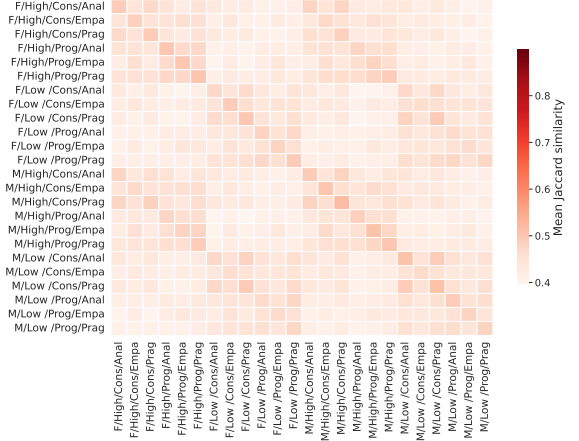
spans a wider range (Qwen3-VL: 0.44–0.70; Gemma4: 0.45–0.88) and diagonal cells are consistently among the highest values in their respective rows and columns for both models, confirming that agents within the same profile generate more similar justifications than cross-profile pairs (Figure A2). Gemma4 shows an even more pronounced structure (a higher diagonal and a wider off-diagonal spread), but the profiles that resemble each other are shared across models (Pearson $r = 0.8$). Thus, persona effects concentrate in interpretive framing rather than descriptive content, consistently across models (highest DSI values for justifications).

Perception tag Jaccard similarity (Figure A3) is reported on a different scale and is not directly comparable to cosine values, as perception tags are controlled-vocabulary sets rather than free text (Section 3.2). Within each model, profiles that diverge most in justification also tend to diverge in perception-tag selection (Qwen3-VL: Pearson $r = 0.67$; Gemma4: $r = 0.82$), suggesting that differences in interpretive framing are partially reflected in structured semantic selection. The perception-tag matrices are also strongly correlated across models ($r = 0.88$; Table 3). However, the dimensional analysis shows that the magnitudes of these attribute-associated differences vary substantially across output types and models (Section 4.2).

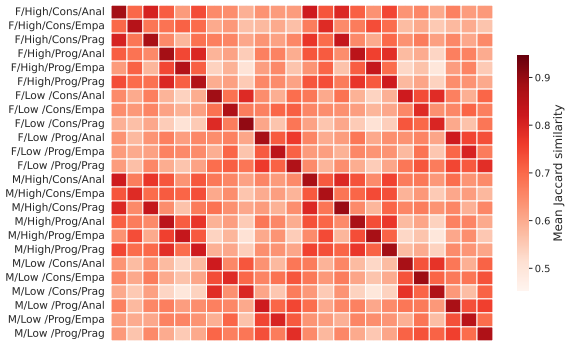
B Decoding-Sensitivity Ablation at $T = 0$

To test whether the strongest persona-associated pattern identified in the main analysis depends on stochastic decoding, we conduct a controlled ablation at $T = 0$. We focus on economic status, which produces the largest justification difference in both models in the main analysis. We hold the remaining attributes fixed as Female, Conservative, and Analytical and vary only the economic-status label (Low vs. High). For each model, we generate outputs under both conditions for all $|I| = 50$ images. In this ablation, both models run through Hugging Face transformers with bf16 weights (Qwen3-VL as Qwen/Qwen3-VL-8B-Instruct), so for Qwen3-VL the ablation also differs from the main corpus in serving stack and precision.

Under the evaluated deterministic configuration, repeated generations within each economic-status condition produced byte-identical outputs because the instance identifier was not included in the



(a) Qwen3-VL



(b) Gemma4

Figure A3: Inter-profile *perception-tag* similarity. DSI = 11.7% for Qwen3-VL and 25.1% for Gemma4, which exhibits higher DSI and similarity values, yet the patterns remain strongly correlated (Pearson $r = 0.88$).

prompt. Within-condition similarity was therefore 1.000 and is not informative for this analysis. Instead, we compute, for each image, the cross-condition similarity between the High- and Low-income outputs. We use cosine similarity for captions and justifications and Jaccard similarity for perception tags, following Section 3.2.

Changing the economic-status label leaves captions unchanged in 24 of 50 images (48%) for Qwen3-VL, whereas no Gemma4 caption pairs are byte-identical. Justifications differ for every image in both models (see Figure B1). Nevertheless, caption similarity remains high across economic status conditions (Qwen3-VL: 0.960; Gemma4: 0.823), whereas justification similarity is substantially lower (0.612 and 0.554, respectively). The corresponding mean caption-minus-justification similarity differences are 0.348 for Qwen3-VL and 0.269 for Gemma4. Paired

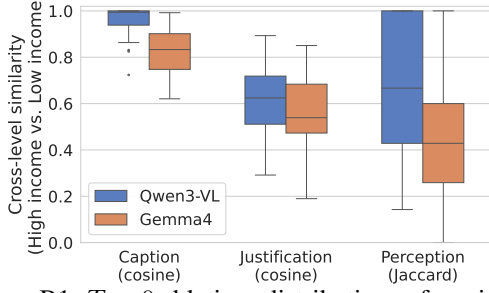


Figure B1: $T = 0$ ablation: distributions of per-image cross-condition similarity between outputs generated with the High- and Low-income persona labels for Qwen3-VL and Gemma4, with all other persona attributes held fixed. Caption and justification values use cosine similarity, whereas perception tags use Jaccard similarity.

Wilcoxon signed-rank tests over the 50 images confirm that cross-condition similarity is lower for justifications than for captions in both models (Qwen3-VL: $p = 1.8 \times 10^{-15}$; Gemma4: $p = 2.2 \times 10^{-12}$).

Perception-tag Jaccard similarity is 0.673 for Qwen3-VL and 0.423 for Gemma4. Because perception tags use Jaccard similarity rather than the cosine similarity used for captions and justifications, these values are interpreted separately.

Thus, the economic-status caption–justification contrast persists under deterministic decoding in both models. Changing the economic-status label affects interpretive framing more strongly than descriptive grounding, even when stochastic sampling is removed, supporting the robustness of the main economic-status result to decoding temperature.

C Full Paired Attribute-Level Statistics

Table C1 reports the full paired image-level results for both models. Panel A reports the mean within-minus-cross difference across the images, Δ_d , for each persona dimension and output type. Panel B reports the corresponding justification-minus-caption contrast. The terms in Panel B are directly comparable because captions and justifications are represented using the same embedding model and evaluated using cosine similarity.

For each comparison, n_+ denotes the number of images with a positive paired difference. Let W_+ and W_- denote the sums of the positive and negative signed ranks, respectively. We report $W = \min(W_+, W_-)$ and the matched-pairs rank-biserial correlation $r_{tb} = (W_+ - W_-)/(W_+ + W_-)$. Mean differences and their 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals summarize effect magnitude, while p_{BH} reports the two-

sided Wilcoxon signed-rank result after Benjamini–Hochberg correction.

All signed-rank results in Table C1 remain significant after Benjamini–Hochberg correction within their respective testing families. Because statistical significance is nearly uniform across comparisons, substantive interpretation focuses on the mean differences and their BCa confidence intervals. In particular, the Qwen3-VL gender contrast in Panel B has a small mean of +0.0004 and a confidence interval that includes zero, despite a significant signed-rank result. This difference arises because the BCa interval summarizes the arithmetic mean, whereas the signed-rank test evaluates the ranked location of the paired differences.

Exact two-sided sign tests provide an additional sensitivity analysis and also remain significant after correction. The largest adjusted sign-test p -values are 4.5×10^{-10} in Panel A and 0.00260 in Panel B.

A complete case sensitivity analysis yields the same substantive conclusions. For Qwen3-VL, restricting the analysis to the 33 images with all 1,200 persona-conditioned generations leaves the mean gender justification-minus-caption contrast unchanged at +0.0004, although its adjusted signed-rank result becomes $p_{BH} = 0.0801$. The other three Qwen3-VL contrasts remain substantively unchanged. For Gemma4, all four contrasts remain substantively unchanged when the analysis is restricted to its 49 complete images.

Reading example. For Qwen3-VL economic status, $\Delta_d = +0.0617$ indicates that justifications generated under profiles sharing the same economic status label are, on average, 0.0617 cosine points more similar than justifications generated under profiles with different economic status labels. The 95% BCa interval $[+0.0538, +0.0710]$ excludes zero, and all 50 image-level differences are positive.

D Matched-Factorial Contrast

The analysis in Appendix C compares pairs that share or differ on a target persona dimension while averaging over the remaining three dimensions. Consequently, profiles in a cross-group pair may also differ on one or more additional attributes. As a complementary controlled analysis, we repeat the comparison using matched profile pairs that agree on all other attributes and differ only on the target dimension.

		Qwen3-VL						Gemma4					
Type	Dimension	Estimate [95% BCa CI]	n_+	W	r_{rb}	p_{BH}		Estimate [95% BCa CI]	n_+	W	r_{rb}	p_{BH}	
Panel A. Within-group minus cross-group similarity													
Caption	Economic status	+0.0138 [+0.0102, +0.0224]	50	0	1.000	2.13×10^{-15}		+0.0211 [+0.0159, +0.0282]	50	0	1.000	2.37×10^{-15}	
	Political orientation	+0.0072 [+0.0035, +0.0217]	50	0	1.000	2.13×10^{-15}		+0.0089 [+0.0059, +0.0144]	49	1	0.998	4.26×10^{-15}	
	Personality	+0.0032 [+0.0025, +0.0043]	50	0	1.000	2.13×10^{-15}		+0.0070 [+0.0056, +0.0086]	50	0	1.000	2.37×10^{-15}	
	Gender	+0.0017 [+0.0012, +0.0028]	50	0	1.000	2.13×10^{-15}		+0.0024 [+0.0018, +0.0034]	46	10	0.984	7.64×10^{-14}	
Justification	Economic status	+0.0617 [+0.0538, +0.0710]	50	0	1.000	2.13×10^{-15}		+0.1015 [+0.0922, +0.1127]	50	0	1.000	2.37×10^{-15}	
	Political orientation	+0.0444 [+0.0390, +0.0520]	50	0	1.000	2.13×10^{-15}		+0.0561 [+0.0500, +0.0639]	50	0	1.000	2.37×10^{-15}	
	Personality	+0.0225 [+0.0200, +0.0257]	50	0	1.000	2.13×10^{-15}		+0.0632 [+0.0583, +0.0699]	50	0	1.000	2.37×10^{-15}	
	Gender	+0.0021 [+0.0018, +0.0026]	50	0	1.000	2.13×10^{-15}		+0.0073 [+0.0064, +0.0084]	50	0	1.000	2.37×10^{-15}	
Perception	Economic status	+0.0358 [+0.0273, +0.0479]	49	1	0.998	3.88×10^{-15}		+0.1068 [+0.0853, +0.1393]	50	0	1.000	2.37×10^{-15}	
	Political orientation	+0.0171 [+0.0137, +0.0214]	50	0	1.000	2.13×10^{-15}		+0.0650 [+0.0481, +0.0920]	50	0	1.000	2.37×10^{-15}	
	Personality	+0.0122 [+0.0097, +0.0156]	50	0	1.000	2.13×10^{-15}		+0.0469 [+0.0379, +0.0572]	50	0	1.000	2.37×10^{-15}	
	Gender	+0.0030 [+0.0023, +0.0040]	48	5	0.992	1.78×10^{-14}		+0.0099 [+0.0074, +0.0133]	49	6	0.991	2.71×10^{-14}	
Panel B. Justification minus caption within-cross difference													
Justification – caption	Economic status	+0.0478 [+0.0389, +0.0575]	49	29	0.955	6.18×10^{-12}		+0.0804 [+0.0707, +0.0913]	50	0	1.000	2.37×10^{-15}	
	Political orientation	+0.0372 [+0.0307, +0.0438]	49	40	0.937	2.06×10^{-11}		+0.0472 [+0.0416, +0.0537]	50	0	1.000	2.37×10^{-15}	
	Personality	+0.0193 [+0.0167, +0.0225]	50	0	1.000	7.11×10^{-15}		+0.0562 [+0.0512, +0.0632]	50	0	1.000	2.37×10^{-15}	
	Gender	+0.0004 [†] [−0.0003, +0.0009]	36	328	0.485	0.00236		+0.0049 [+0.0036, +0.0062]	45	90	0.859	3.71×10^{-9}	

Table C1: Paired image-level analysis for both MLLMs. Panel A reports the within-minus-cross difference Δ_d , and Panel B reports the justification-minus-caption contrast $\Delta_d^{\text{just}} - \Delta_d^{\text{cap}}$. n_+ is the number of images with a positive paired difference, W is the smaller signed-rank sum, and r_{rb} is the matched-pairs rank-biserial correlation. The displayed p -values are from two-sided Wilcoxon signed-rank tests with Benjamini–Hochberg correction applied separately within each model and panel: 12 comparisons in Panel A and four in Panel B. Confidence intervals are 95% BCa bootstrap intervals for the mean image-level difference. Because the confidence intervals and signed-rank tests target different summary quantities, they need not always yield identical conclusions. [†] marks a confidence interval that includes zero.

The four persona dimensions define 24 complete profiles. For a target dimension d , let $\mathcal{M}_d$ denote the set of profile pairs (p, q) that agree on every other attribute and differ only on d . This produces 12 matched pairs for each binary dimension. For personality, the eight combinations of the remaining binary attributes each contribute three pairwise comparisons among Pragmatic, Empathetic, and Analytical, producing 24 matched pairs.

For image i and matched pair $(p, q) \in \mathcal{M}_d$, we define

$$\delta_{i,pq,d} = \frac{1}{2} \left(s_{i,p}^{\odot} + s_{i,q}^{\odot} \right) - s_{i,pq}^{\otimes},$$

where $s_{i,p}^{\odot}$ is the mean pairwise similarity among outputs generated under profile p for image i , and $s_{i,pq}^{\otimes}$ is the mean similarity between outputs generated under profiles p and q for that image. We then average over the matched profile pairs:

$$\Delta_{i,d}^{\text{match}} = \frac{1}{|\mathcal{M}_d|} \sum_{(p,q) \in \mathcal{M}_d} \delta_{i,pq,d}.$$

This procedure yields one matched-factorial contrast for each of the 50 images. We apply the same image-level inferential procedure used in Appendix C: a two-sided Wilcoxon signed-rank test, a 95% BCa bootstrap confidence interval obtained by resampling images, an exact sign-test sensitivity analysis, and Benjamini–Hochberg correction

across the 12 comparisons within each model. Similarities are computed using all available valid generations. Every image–profile cell contains at least two valid outputs, so all within-profile and between-profile similarities are defined.

Table D1 reports the matched-factorial estimates. The economic-status dimension produces the largest matched contrast and gender the smallest across both models and all three output types. The relative ordering of the four dimensions is preserved in five of the six model–output type combinations. For Gemma4 justifications, political orientation slightly exceeds personality (+0.2037 versus +0.1976), reversing their second- and third-ranked positions in the marginal analysis.

The marginal and matched analyses provide complementary views of persona-associated variation. Appendix C summarizes within-group and cross-group similarity across the full set of profiles, averaging over the remaining persona dimensions. The matched-factorial analysis instead isolates each target dimension by comparing profiles that are identical on the other three attributes. Because the two analyses use different profile-pair sets and different within-group reference quantities, their numerical magnitudes should not be compared directly.

The matched estimator also incorporates within-profile reproducibility through $s_{i,p}^{\odot}$, which compares outputs generated under the same complete

Type	Dimension	Pairs	Mean Δ^{match} [95% BCa CI]	
			Qwen3-VL	Gemma4
Caption	Economic status	12	+0.0229 [+0.0180, +0.0319]	+0.0380 [+0.0305, +0.0473]
	Political orientation	12	+0.0156 [+0.0110, +0.0306]	+0.0239 [+0.0179, +0.0324]
	Personality	24	+0.0098 [+0.0079, +0.0126]	+0.0199 [+0.0165, +0.0239]
	Gender	12	+0.0074 [+0.0058, +0.0099]	+0.0112 [+0.0088, +0.0147]
Justification	Economic status	12	+0.0976 [+0.0876, +0.1084]	+0.2684 [+0.2501, +0.2861]
	Political orientation	12	+0.0805 [+0.0732, +0.0903]	+0.2037 [+0.1877, +0.2202]
	Personality	24	+0.0456 [+0.0414, +0.0507]	+0.1976 [+0.1858, +0.2098]
	Gender	12	+0.0106 [+0.0095, +0.0117]	+0.0662 [+0.0609, +0.0726]
Perception	Economic status	12	+0.0619 [+0.0518, +0.0758]	+0.2349 [+0.1950, +0.2778]
	Political orientation	12	+0.0413 [+0.0363, +0.0479]	+0.1749 [+0.1425, +0.2134]
	Personality	24	+0.0325 [+0.0282, +0.0375]	+0.1411 [+0.1169, +0.1659]
	Gender	12	+0.0171 [+0.0147, +0.0197]	+0.0619 [+0.0496, +0.0773]

Table D1: Matched-factorial contrasts for both MLLMs. Each row averages over profile pairs that agree on the other three persona attributes and differ only on the stated dimension. For each matched pair, the contrast compares the average within-profile similarity of the two profiles with their between-profile similarity. *Pairs* denotes the number of matched profile pairs. All 50 image-level contrasts are positive in every comparison, yielding $n_+ = 50$, $W = 0$, $r_{\text{tb}} = 1.000$, and $p_{\text{BH}} = 1.78 \times 10^{-15}$ throughout; these repeated statistics are omitted from the table. Exact sign tests also remain significant after Benjamini–Hochberg correction. Caption and justification results use cosine similarity, whereas perception-tag results use Jaccard similarity and are interpreted separately.

persona prompt. Because all image-level matched contrasts are positive, their magnitudes and the relative ordering of dimensions provide the most informative comparisons. The economic-status dimension remains the largest, and gender remains the smallest across models and output types.

The marginal analysis remains the primary presentation because it summarizes attribute-associated similarity across the full corpus and corresponds directly to the distributions in Figure 2. The matched-factorial analysis provides complementary controlled evidence that the main dimensional pattern persists when the other three persona attributes are held fixed.

E Matched Justification–Caption Contrast

Table D1 reports the matched caption and justification differences separately. Because both output types are represented using the same embedding model and evaluated with cosine similarity, their matched differences are directly comparable. For each image i and persona dimension d , we define

$$\Gamma_{i,d} = \Delta_{i,d}^{\text{just,match}} - \Delta_{i,d}^{\text{cap,match}}.$$

Positive values indicate that the matched attribute-associated difference is larger for justifications than for captions.

We analyze the 50 image-level contrasts using a two-sided Wilcoxon signed-rank test and report a 95% bias-corrected and accelerated (BCa)

bootstrap confidence interval for the mean contrast, obtained by resampling images. Benjamini–Hochberg correction is applied across the four persona dimensions within each model. An exact two-sided sign test provides an additional sensitivity analysis. Perception tags are excluded because their Jaccard similarities are not commensurate with the cosine similarities used for captions and justifications.

Table E1 reports the resulting contrasts. All mean values are positive, ranging from +0.0032 to +0.0747 for Qwen3-VL and from +0.0550 to +0.2303 for Gemma4. Economic status produces the largest justification-minus-caption contrast and gender the smallest in both models. The confidence intervals exclude zero for all eight model–dimension comparisons, and all Wilcoxon results remain significant after correction.

The Qwen3-VL gender result provides a useful sensitivity case. In the marginal analysis, the mean justification-minus-caption contrast is +0.0004, with a confidence interval that includes zero (Table C1). In the matched design, the mean contrast remains small but is positive at +0.0032, with a 95% BCa interval of [+0.0012, +0.0048]. Thus, the broader pattern of greater attribute-associated variation in justifications than in captions persists when the remaining persona attributes are held fixed. Across both models, the matched analysis also preserves the main ordering, with economic status producing the largest contrast and gender the smallest.

Dimension	Qwen3-VL						Gemma4					
	Mean Γ [95% BCa CI]	n_+	W	r_{rb}	p_{BH}		Mean Γ [95% BCa CI]	n_+	W	r_{rb}	p_{BH}	
Economic status	+0.0747 [+0.0638, +0.0862]	49	5	0.992	2.37×10^{-14}		+0.2303 [+0.2119, +0.2489]	50	0	1.000	1.78×10^{-15}	
Political orientation	+0.0649 [+0.0567, +0.0739]	49	4	0.994	2.37×10^{-14}		+0.1799 [+0.1638, +0.1962]	50	0	1.000	1.78×10^{-15}	
Personality	+0.0358 [+0.0315, +0.0413]	50	0	1.000	7.11×10^{-15}		+0.1777 [+0.1659, +0.1900]	50	0	1.000	1.78×10^{-15}	
Gender	+0.0032 [+0.0012, +0.0048]	38	270	0.576	2.52×10^{-4}		+0.0550 [+0.0491, +0.0617]	50	0	1.000	1.78×10^{-15}	

Table E1: Matched-factorial justification-minus-caption contrasts, $\Gamma_{i,d} = \Delta_{i,d}^{\text{just,match}} - \Delta_{i,d}^{\text{cap,match}}$. Positive values indicate that the matched attribute-associated difference is larger for justifications than for captions. Each row uses 50 image-level contrasts. Confidence intervals are 95% bias-corrected and accelerated bootstrap intervals for the mean, obtained by resampling images. The displayed p_{BH} values are from two-sided Wilcoxon signed-rank tests with Benjamini–Hochberg correction across the four dimensions within each model. n_+ is the number of positive image-level contrasts, W is the smaller signed-rank sum, and r_{rb} is the matched-pairs rank-biserial correlation. Exact two-sided sign tests also remain significant after correction.

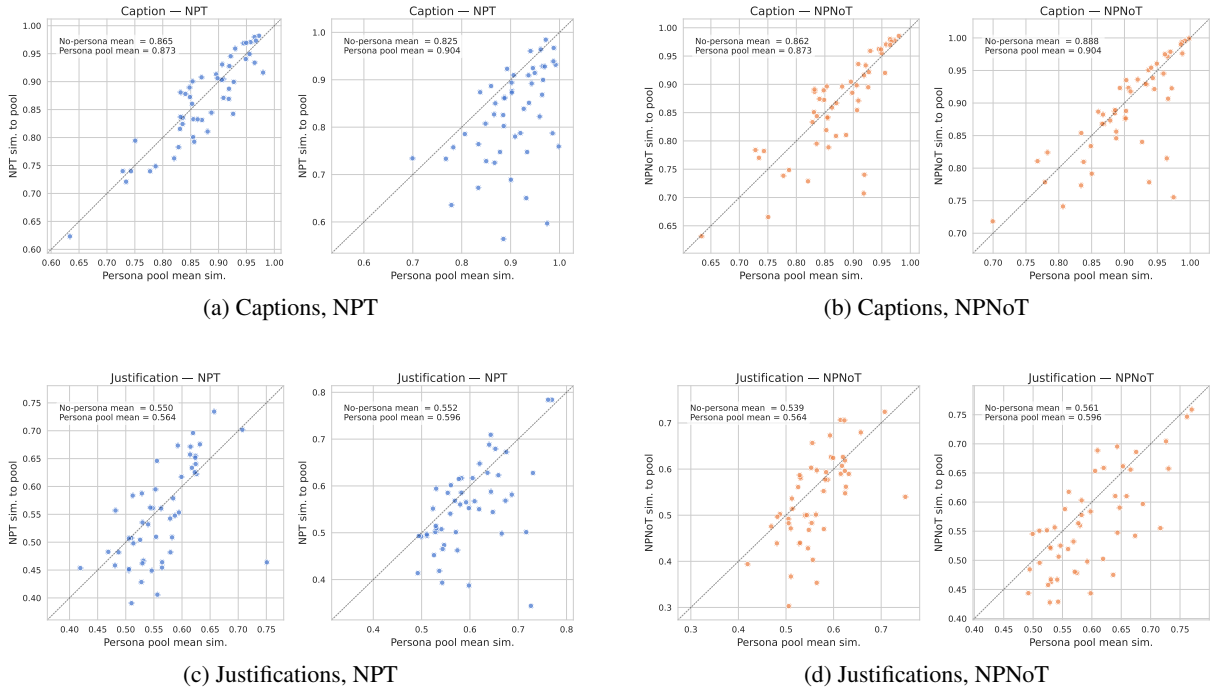


Figure F1: Per-image cosine similarity of no-persona outputs to the persona-conditioned pool: (a) captions with extended reasoning (NPT), (b) captions without extended reasoning (NPNoT), (c) justifications with extended reasoning (NPT), and (d) justifications without extended reasoning (NPNoT). Within each panel, Qwen3-VL is shown on the left and Gemma4 on the right. Each point is one image; the dashed line marks $y = x$.

F Persona vs. No-Persona Effects

To contextualize the role of persona conditioning, we compare two no-persona variants (NPT and NPNoT), for each model, against the model’s persona-conditioned pool using the same embedding-based cosine similarity described in Section 3.2, computed separately for each image.

For each image $i \in I$, we consider two quantities: (i) the average cosine similarity among persona-conditioned outputs (persona-persona), and the cosine similarity between the no-persona output and the mean representation of the persona-conditioned outputs. Figure F1 plots these for

both models. Each point indicates how closely a no-persona response aligns with the persona pool relative to the internal agreement among persona-conditioned agents, allowing us to assess whether persona conditioning produces systematic shifts in generated language beyond the model’s internal persona-conditioned variability.

Note that this comparison is structurally asymmetric: the persona-conditioned pool comprises $\approx 1,200$ annotations per image, whereas each no-persona setting contributes only a handful. The no-persona values should therefore be interpreted as indicative rather than statistically equivalent base-lines.

For captions, both no-persona variants yield similarity levels close to the persona within-pool average (Qwen3-VL —NPT: 0.865; NPNNoT: 0.862; persona within-pool: 0.873; Gemma4 — NPT: 0.825, NPNNoT: 0.888, within-pool: 0.904). This is consistent with the strong convergence in descriptive grounding observed in Section 4.1, indicating that descriptions remain stable even without persona conditioning.

In contrast, justification similarities are lower overall (Qwen3-VL — NPT: 0.550; NPNNoT: 0.539; within-pool: 0.564; Gemma4 — NPT: 0.552, NPNNoT: 0.561, within-pool: 0.596), reflecting the greater semantic variability of interpretive framing relative to captions. In both models, the no-persona outputs remain broadly aligned with, yet slightly below, the persona within-pool average, consistent with the broader pattern observed in Section 4.2, where persona conditioning introduces structured variation primarily in justificatory language.

Beyond these mean differences, Figure F1 shows that this pattern holds consistently across images for both models: caption similarities cluster tightly near the persona within-pool level, whereas justification similarities are both lower and more variable, reinforcing that persona effects primarily operate in interpretive language.

These comparisons are consistent with the main finding that descriptive grounding is more stable than interpretive framing. Because the persona and no-persona conditions contain substantially different numbers of generations, however, they should be interpreted as contextual evidence rather than as a definitive test separating persona-conditioned effects from stochastic variation.

G Topics Analysis

Figure G1 shows that, in both models, justification topics align with sentiment in interpretable ways: accident, conflict, and decay topics skew negative; nature/beauty and warmth topics skew positive; and infrastructure or neutral/stark topics sit in between, reflecting ambiguity in how agents interpret ordinary urban development scenes.

H Prompt Templates

Figure H1 shows the prompt templates used for all annotations in both models. In the persona condition, the profile is specified through a direct second-person role instruction in the system prompt, with each placeholder filled with the corresponding at-

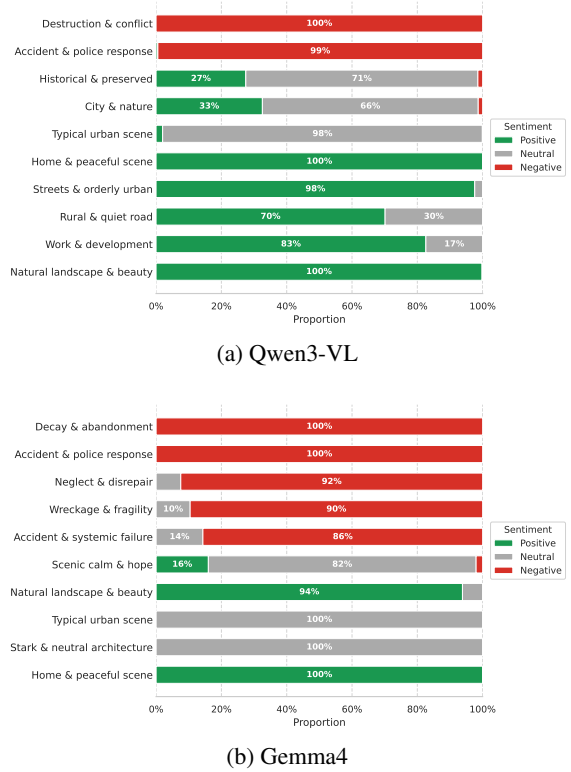


Figure G1: Sentiment distribution across the top 10 justification topics generated by persona-conditioned agents for Qwen3-VL and Gemma4.

tribute level (e.g., Low income or High income for economic status). The same annotation task and output instructions are used in the persona and no-persona conditions; only the persona-specific block and associated rules are replaced in the no-persona condition. The prompt explicitly requests a caption written independently of the persona and a justification in the persona’s own voice, a design choice discussed in the Limitations section.

[System prompt, persona condition]

You are a person with the following demographic profile:

- Gender: {gender}
- Economic status: {economic_status}
- Political leaning: {political_spectrum}
- Personality archetype: {personality}

CRITICAL RULES:

1. You are this person. You are NOT an AI.
2. Evaluate the image exactly as this person would, based on your background.
3. Do not break character. Do not apologize for your views.

TASK – IMAGE SENTIMENT ANNOTATION:

Look at the image and respond with a JSON object and nothing else.

The JSON must have exactly these four keys:

- "sentiment" : one of ['Positive', 'SlightlyPositive', 'Neutral', 'SlightlyNegative', 'Negative']
- "perceptions" : a JSON array of 1-5 labels chosen EXCLUSIVELY from the PERCEPTION VOCABULARY listed below – do NOT invent new labels
- "caption" : a 1-2 sentence objective description of what is shown in the image, written independently of your persona
- "justification": a single sentence in YOUR voice explaining your sentiment score

PERCEPTION VOCABULARY (593 labels – choose 1-5 that best match the image):

["\"Closed\" environment", "\"Home\" feeling, something reminiscent of childhood", "\"Water mirror\"", "Abandoned", ...]

Example response:

```
{
  "sentiment": "Negative",
  "perceptions": ["Violence", "Debris/Destruction"],
  "caption": "A damaged street with rubble and broken glass scattered across the pavement.",
  "justification": "Looks like the same mess they always leave after those marches downtown."
}
```

Return ONLY the JSON object. No markdown, no code fences, no additional text.

[System prompt, no-persona condition: the persona block and CRITICAL RULES are replaced by]

You are an impartial image annotator. Evaluate each image objectively, without any personal or demographic lens.

[User turn, both conditions: this text, followed by the image]

What is the sentiment of this image?

Figure H1: Prompt templates. Placeholders in braces are filled with the persona’s attribute levels. The perception vocabulary is truncated here (593 labels in full); line breaks in the vocabulary, example response, and no-persona prompt were added for layout.