A Thousand Words or An Image: Studying the Influence of Persona Modality in Multimodal LLMs

Julius Broomfield **Kartik Sharma** **Srijan Kumar**
Georgia Institute of Technology
{jbroomfield, ksartik, srijan}@gatech.edu

Abstract

LLMs have demonstrated remarkable advancements in embodying diverse personas, enhancing their effectiveness as conversational agents. Simultaneously, there have been significant strides in these models in processing and integrating visual information. However, despite images being a natural carrier of persona attributes (e.g. age, occupation, or location), it remains largely unexplored whether multimodal LLMs (MLLMs) can faithfully ground visual persona information. In this work, we introduce persona embodiment as a novel diagnostic for visual grounding in MLLMs. Unlike simple visual recognition, this task requires models to extract multiple persona attributes from an image, integrate them into a coherent representation, and apply them to downstream language tasks. To this end, we introduce a *modality–parallel* dataset of personas, containing 2,400 prompts each represented in four modalities. Across our experiments, we find that models *frequently underuse visual information relative to text*, and that this neglect persists even when richer visual information is clearly available, highlighting underlying limitations and paving the way for future research to bridge this gap. We release the dataset and code at https://github.com/claws-lab/persona-modality.

1 Introduction

"A picture is worth a thousand words," as the adage goes, captures how images can convey rich and abstract information more compactly than their textual equivalents (Paivio, 2013). Yet despite the rapid integration of vision capabilities into language models (MLLMs) (Liu et al., 2024; Li et al., 2024a; GPT-4o, 2025), there is much evidence suggesting these models struggle to extract and reason over visual information as effectively as they do over text (Tong et al., 2024; Chen et al., 2024b). Understanding precisely *where* and *how* this cross-modal gap manifests is an open challenge.



Figure 1: **GPT-4o responses conditioned on the same persona (a Parisian chef) across two modalities.** The left side presents an image persona, while the right side features a text persona derived from the image.

In this work, we use persona embodiment as a diagnostic lens to expose these visual grounding failures. LLMs have increasingly been adopted as role-playing agents for many tasks (Li et al., 2024b), such as for personalized conversational agents (Tseng et al., 2024). As these models become more capable of handling different modalities (Liu et al., 2024), it is crucial that they can express these personas by incorporating information from representative images (Ahn et al., 2023). We believe persona embodiment is a particularly sensitive probe for visual grounding as it requires a model to simultaneously extract multiple persona attributes (e.g. *age* from facial appearance, *occupation* from clothing and tools, *location* from background context) and apply them coherently to downstream reasoning tasks.

While prior work has established that incorporating visual information can improve dialogue con-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33309–33328
October 24-29, 2026 ©2026 Association for Computational Linguistics

sistency and persona grounding (Sun et al., 2024; Ahn et al., 2023; Dai et al., 2025), no systematic comparison exists between personas represented in different modalities. It therefore remains unclear whether current multimodal LLMs can effectively embody a persona conveyed visually, or whether they default to ignoring visual information in favor of any available text.

In this work, we present the first comprehensive study of the influence of persona modality on its embodiment by multimodal LLMs. Figure 1 illustrates our setup, and our contributions are:

1. We conduct the **first systematic study** on how a persona's modality affects its expression by multimodal LLMs, using persona embodiment as a diagnostic for visual grounding failures.
2. We create a **novel evaluation dataset** containing 2,400 unique prompts, made of 40 personas, represented in 4 distinct modality representations, evaluated against 5 multimodal LLMs.
3. We perform a **comprehensive evaluation** using *linguistic* and *score-based* metrics, showing that text-based personas are often expressed more faithfully than image-based ones.
4. We perform several **stratified and ablation analyses**, which confirm the stability of our findings across personas, evaluators, and question formats, and demonstrate that models largely ignore visual information when any text is present.

2 Influence of Modality in Persona Embodiment of Multimodal LLMs

We define the task of persona embodiment as generating responses consistent with a specified character, identity, or role (Samuel et al., 2024; Du et al., 2026). This involves maintaining coherence in linguistic style, beliefs, knowledge, and affective tone in a way that aligns with the intended persona.

In this work, we investigate the effect of representing a persona p in different modalities with MLLMs. In particular, we consider two common modalities, text and image, and evaluate the LLM's performance on representations that are equivalent with respect to the target persona attributes. We evaluate additional persona representations by combining visual and textual features of the personas. We describe these 4 different modality representations in more detail below.

1. **Text ($\mathcal{T}$):** A detailed text description of a persona that characterizes the persona in natural language.
2. **Image ($\mathcal{I}$):** A persona can be depicted visually using an image that coherently includes the persona's attributes, characterizing the persona visually.
3. **Assisted Image ($\mathcal{I}_A$):** We pair the visual representation of each persona ($\mathcal{I}$) with a simple text description of the persona (p), giving us a multimodal representation that includes an image and raw text.
4. **Descriptive Image ($\mathcal{I}_D$):** In contrast to the assisted image representation, we pair the visual representation of a persona ($\mathcal{I}$) with a text description of the persona (p) in the image itself using typography instead of via raw text, giving us a unimodal representation of $\mathcal{I}_A$.

3 Modality Parallel Persona Dataset

3.1 Personas

We create a set of personas $\mathcal{P}$, where each persona p is represented in each of the four modalities, $\mathcal{M} = \{\mathcal{I}, \mathcal{T}, \mathcal{I}_A, \mathcal{I}_D\}$, while preserving the same target persona attributes. To ensure effective representation across both text and image modalities, we construct personas based on demographic attributes that can be made easily visualizable (Todorov et al., 2015). Specifically, we define each persona as a combination of their age, gender, occupation, and location. A persona can thus be written as:

A <age>-year-old <gender> <occupation> from <location>,

where <age> $\in [18, 64]$, <gender> $\in$ { male, female }, <location> is a city, and <occupation> denotes a person who does a specific occupation. For example, "A 35-year-old male chef from Paris". As depicted in Figure 1, age and gender can be visualized using the face of the person while occupation and location can be visualized using the clothes and the background respectively. We note that we only use gender as a way of increasing diversity in our image generation stage, and do not make any assumptions or claims about faithfulness in this aspect.

We categorize these attributes into distinct groups and uniformly sample from each category for our evaluations. Table 1 summarizes our dataset of how we choose the age, gender, occupation, and location. In particular, we consider a standard grouping of ages followed in surveys between 18

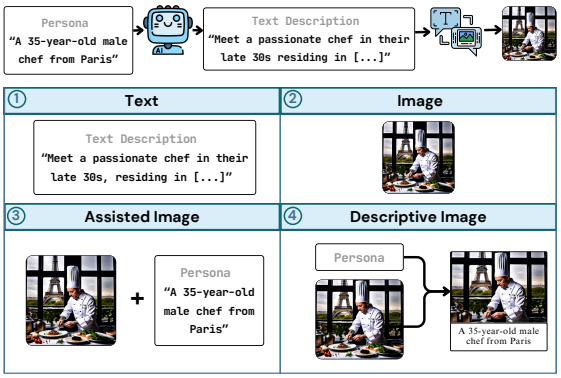


Figure 2: Our pipeline first curates a set of personas ($\mathcal{P}$). For each persona we generate an image ($\mathcal{I}$) using Stable Diffusion. Then, a LLM examines each image and generates its own textual description ($\mathcal{T}$). Pairing each persona with its image produces an assisted image ($\mathcal{I}_\mathcal{A}$), while combining a typographic persona with the image produces a descriptive image ($\mathcal{I}_\mathcal{D}$).

Attribute	Category	Number
Age	18-24	5
	25-34	11
	35-44	13
	45-54	6
	55-64	5
Gender	Male	19
	Female	21
Occupation	Healthcare & Education	9
	Public Safety	5
	Manual Labor	16
	Hospitality	5
	Transportation	5
Location	Largest Economies (GDP > $3T)	12
	Developed Economies (GDP $1T-$3T)	13
	Mid-Sized Powers (GDP $0.5T-$1T)	7
	Emerging Markets (GDP < $0.5T)	8

Table 1: **Persona Dataset Summary**

and 65, a standard male/female splitting of gender, while occupations and locations are categorized based on their primary societal role and economic status [1] respectively. Table 6 in Appendix presents the list of 40 personas we use along with their attributes and attribute categories.

3.2 Equivalent Modality Representations

First, we clarify our notion of equivalence. Let $\mathcal{A}$ denote the set of target persona attributes (i.e. *age*, *occupation*, and *location*). We define two persona representations p and p' to be equivalent, otherwise written $p \sim p'$, if they encode the same values for all attributes in $\mathcal{A}$.

Given a set of persona descriptions p, we construct our *modality parallel* dataset, such that $\mathcal{P}$ can be equivalently depicted in all of the four modalities $\mathcal{M}$. Figure 2 illustrates the methodology we use for creating these different representations, and we detail it below.

1. For each simple persona description p (e.g. a 28 year-old female pilot from Vancouver), we automatically generate a detailed visual description using an LLM[2], using the prompt provided in Appendix A.1.
2. We use Stable Diffusion XL[3] (a text-to-image generative model) to generate a 768×768 px image conditioned on the detailed visual description. Upon doing an extensive hyperparameter search, we found the best results with a guidance scale of $\gamma = 15$ and $n = 50$ diffusion steps. Thus, we obtain a image representation of the persona $\mathcal{I}(p)$.
3. During the image generation process, information may be inadvertently added or removed from the original visual description to the completed image, biasing any direct comparison of modalities. Thus, we prompt the LLM to generate a text description of the persona given the image ($\mathcal{I}(p)$) and nothing else, using the prompt provided in Appendix A.1. Thus, we obtain the text representation $\mathcal{T}(p)$.

We obtain the assisted and descriptive image representations of the persona by pairing $\mathcal{I}(p)$ with the simple description p to create an assisted image $\mathcal{I}_A(p)$, and by rendering[4] p as black text at the bottom of $\mathcal{I}(p)$ on a white background using size 20 Arial font to create $\mathcal{I}_D(p)$. This gives us our complete set of persona modality representations. At each step, we ensure through manual verification that all the *important attributes of the persona are present* in each persona representation.

3.3 Question Generation

To evaluate how well a model embodies a given persona, we create a set of 60 questions Q that each probe for a given attribute type (e.g. age) either directly (Q^D) or in naturalistic scenarios (Q^S). Across both subsets, for each attribute type, we curate 10 questions designed to probe how well the model holds the attribute. We avoid probing from gender due to methodological concerns. Specifically, while age and location can be objectively probed through knowledge claims (e.g. *"what was a popular song from your childhood"*), gender assessment would inevitably rely on stereotypes or

[1] GDP

[2] GPT-4o [3] Stable Diffusion XL [4] Pillow

normative expectations. Thus, we obtain two question sets Q^D and Q^S for probing for each persona's location, occupation, and age.

$$Q^D = \bigcup_{i \in \{L,O,A\}} Q_i^D, \text{ each } |Q_i^D| = 10 \text{ questions}$$

$$Q^S = \bigcup_{i \in \{L,O,A\}} Q_i^S, \text{ each } |Q_i^S| = 10 \text{ scenarios}$$

3.3.1 Direct Testing

In this question set (Q^D), questions were designed to directly probe specific knowledge of an agent, conditioned on the fact they hold some attribute, with the goal of enabling objective evaluation. For example, to probe for age we ask "*what life experiences do you consider most defining for your generation?*" while for location we ask "*what is the most visited tourist attraction in your area?*". We provide the complete question set in Table 7 in the Appendix.

3.3.2 Situational Testing

Our scenario set (Q^S) accomplishes similar knowledge evaluation via more naturalistic, open-ended situations, requiring personas to implicitly demonstrate both knowledge and behavioral consistency. For example, to probe for age we ask "*You're coordinating a playlist for your high school reunion after-party. The organizers want music specifically from your graduating years to recreate the atmosphere. You . . .*", which is detailed in Figure 1. We provide the scenario set in Table 8 in the Appendix.

4 Evaluation Methodology

4.1 Model Evaluation

Models. We evaluate the performance of 5 representative MLLMs: `gpt-4o-2024-08-06`[5], `gpt-4o-mini-2024-07-18`[6], `Llama-3.2-11B-Vision-Instruct`[7], `Llama-3.2-90B-Vision-Instruct`[8], and `mistral/pixtral-12b-2409`[9]. For each model, we sample with temperature of 1.0 and *top-p* of 1.0. No additional hyperparameters are set. We select these models as they represent a diverse cross-section of three major model families, span both closed and open-weight models, and have varying parameter sizes.

Procedure. In practice, our complete evaluation setup consists of a set of unique personas $\mathcal{P}$, each represented in four modalities $\mathcal{P}^* = \mathcal{P} \times \mathcal{M}$. For each persona $p^* \in \mathcal{P}^*$, we evaluate a model by first conditioning it on p^*, and then by prompting it with each question $q \in Q$. We evaluate the quality of each model's response independently based only the question q and the persona description p. The personas are given to the model as a user message directly preceding the user message that includes the evaluation question q; this means the model is conditioned on p in one turn, then without it responding, immediately given q in the next turn.

Evaluators. We utilize two LLM as evaluators, using GPT-4o and Gemini 2.0 Flash [10] with greedy sampling. All scores discussed in the main paper are averaged across the two models, while individual scores can be found in Tables 9 and 10 in the Appendix. We use human evaluators on a large subset of the evaluation set to assess the LLM evaluator scores' alignment with human scores.

4.1.1 Likert Evaluation

We employ an LLM evaluator to judge the quality of the responses based on different criteria. In particular, we prompt the LLM judge with the question asked q, model response, and the persona description p on these metrics as described by the corresponding prompts. We evaluate using the following three criterion established in prior work (Samuel et al., 2024):

1. **Expected Action** measures whether the model selects actions that are appropriate for the persona in a given scenario
2. **Linguistic Habits** assesses whether its language, tone, jargon, and style are consistent with the persona
3. **Persona Consistency** evaluates whether responses remain faithful to the persona's stated attributes under direct questioning

For each evaluation criterion, the judge outputs a score from a 5-point Likert scale based on the corresponding system prompt. For situational testing, we evaluate using *expected action* and *linguistic habits*; for direct testing, we use *persona consistency* and *linguistic habits*. We combine the scores for linguistic habits across the two testing sets to find the average score. Alongside the LLM evaluator, we also employ human evaluators to validate the judge quality and provide a human reference for the scores.

[5] GPT-4o [6] GPT-4o-mini [7] Llama 3.2 11B [8] Llama 3.2 90B [9] Pixtral 12B

[10] Gemini 2.0 Flash

4.1.2 Comparative Evaluation

We use two comparative evaluation methods, as opposed to the Likert scoring, to assess relative performance across modalities.

Pairwise Comparison To compare responses across the text $\mathcal{T}$ and image $\mathcal{I}$ modalities, we first directly compare model responses conditioned on the text persona versus the image persona, based on which one is more holistically more aligned with the original persona p.

Swiss System Comparison To collectively evaluate all four modalities, we use a Swiss tournament system in order to reduce the number of required comparisons compared to pairwise evaluation while maintaining ranking quality.

4.1.3 Linguistic Analysis

Alongside the *linguistic habits* evaluation criterion, we also analyze the lexical diversity, variation, and complexity of each response using established metrics from computational linguistics:

1. **Types:** unique token count.
2. **Root Type-Token Ratio (RTTR):** $= {}^{\text{types}}/_{\sqrt{\text{length}}}$, a normalized measure of lexical variation found by dividing the number of unique tokens with the response length (van Hout and Vermeer, 2007).
3. **Measure of Textual Lexical Diversity (MTLD):** a measure lexical diversity found by calculating the mean length of text segments that maintain a type-token ratio $> \tau = 0.72$ (McCarthy and Jarvis, 2010),.

5 Results

5.1 Effect of Persona Modality

LLM-based Evaluation To evaluate the responses to persona-specific questions from multimodal LLMs, we compare the average scores generated for responses under 4 different criteria as mentioned above. Table 2 shows the mean and standard deviation scores of each criterion in the Likert scale for the 4 modality representations across LLMs. We find that text-based personas score the highest in almost all criteria in almost all models, consistently improving the linguistic habits of the persona in all models by a minimum 0.2 increase in the score. This shows that text is the most preferred way to represent a persona across models, highlighting the lack of understanding of the equivalent visual information. Some notable exceptions are in the persona consistency and expected action criteria, where the descriptive image modality (*i.e.*, a descriptive text is embedded within the image), shows a significantly higher rating than the text modality in GPT-4o, GPT-4o-mini, and Pixtral models.

We note that the image and assisted image modalities consistently show similar and lower performance than others, showing that the assisted text fails to encode additional information that the models cannot already derive from the image; we perform ablation experiments investigating this further in section 5.3.1. We also note that GPT-4o shows the highest average alignment scores across 4 criteria in each modality representation. In addition, despite the small size, we find that Pixtral is much better at capturing visual information than larger Llama models. On the other hand, Llama models shine when the persona is represented in text, significantly outperforming Pixtral.

Preference-based Evaluation We also leverage GPT-4o as a judge to pick and choose the most aligned response with the persona, directly comparing the 4 modalities. Table 3 shows the percentage of times that the judge picks each modality in the Swiss comparison and the pairwise (only text and image) comparison setting. We find that text-based persona responses are picked for at least 90% of the questions, showing a clear preference for responses generated through text-based embodiment of these models. Furthermore, in a direct pairwise comparison, we find that image-based personas are almost never chosen (selecting text up to 99%).

Linguistic Evaluation Next, we compare the linguistic diversity of the responses generated through different persona modalities. Table 5 shows the mean and standard deviation in the three metrics of linguistic diversity in different settings. We find that text modality is the overall preferred way to generate expressive responses of a persona. Specifically, text-based personas generate at least 40 more types of words than the other modalities, along with significantly more variation (at least 2 more).

Human Evaluation We also employ independent human annotators to judge the responses generated using different persona modalities. In particular, we use GPT-4o responses of all 4 modality representations for 10 randomly selected questions. Each participant is shown 10 questions with a response from one of the four modalities and is asked

Criterion	Text	Assisted Image	Image	Descriptive Image
GPT-4o				
Linguistic Habits	**2.07** ± 0.02	1.61 ± 0.02	1.51 ± 0.02	1.59 ± 0.02
Persona Consistency	3.44 ± 0.04	3.20 ± 0.04	3.03 ± 0.04	**3.91** ± 0.04
Expected Action	3.86 ± 0.03	3.59 ± 0.03	3.56 ± 0.03	**3.94** ± 0.03
Action Justification	**4.13** ± 0.03	3.82 ± 0.03	3.75 ± 0.03	4.00 ± 0.03
GPT-4o-mini				
Linguistic Habits	**1.81** ± 0.02	1.61 ± 0.02	1.48 ± 0.02	1.63 ± 0.02
Persona Consistency	2.98 ± 0.04	2.98 ± 0.04	2.97 ± 0.04	**3.58** ± 0.04
Expected Action	3.25 ± 0.03	3.29 ± 0.03	3.19 ± 0.03	**3.59** ± 0.03
Action Justification	**3.56** ± 0.03	3.47 ± 0.03	3.35 ± 0.03	3.55 ± 0.03
Llama 3.2 11B				
Linguistic Habits	**2.20** ± 0.03	1.30 ± 0.02	1.32 ± 0.02	1.44 ± 0.02
Persona Consistency	**2.79** ± 0.04	2.16 ± 0.04	1.90 ± 0.04	2.55 ± 0.04
Expected Action	**2.98** ± 0.03	2.28 ± 0.03	2.04 ± 0.03	2.49 ± 0.03
Action Justification	**3.24** ± 0.03	2.44 ± 0.03	2.17 ± 0.03	2.49 ± 0.03
Llama 3.2 90B				
Linguistic Habits	**2.32** ± 0.03	1.25 ± 0.04	1.25 ± 0.05	1.30 ± 0.05
Persona Consistency	**2.99** ± 0.04	1.63 ± 0.06	1.08 ± 0.05	1.43 ± 0.07
Expected Action	**3.28** ± 0.03	1.43 ± 0.04	1.02 ± 0.03	1.24 ± 0.04
Action Justification	**3.49** ± 0.03	1.67 ± 0.05	1.31 ± 0.05	1.50 ± 0.05
Pixtral 12B				
Linguistic Habits	**1.79** ± 0.02	1.63 ± 0.02	1.65 ± 0.02	1.60 ± 0.02
Persona Consistency	2.38 ± 0.04	2.33 ± 0.04	2.77 ± 0.04	**3.59** ± 0.04
Expected Action	2.93 ± 0.03	2.85 ± 0.03	3.16 ± 0.03	**3.53** ± 0.03
Action Justification	3.32 ± 0.03	3.07 ± 0.03	3.20 ± 0.03	**3.50** ± 0.03

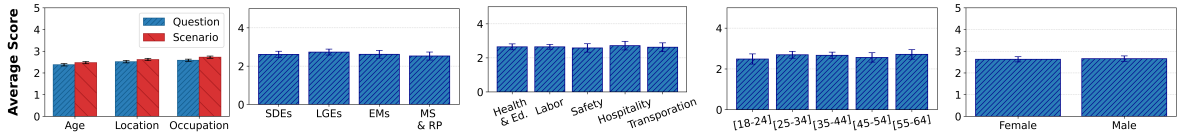
Table 2: **LLM-based evaluation [1-5] of responses under different persona modality representations.**

to judge how well the response is aligned with the persona for the given question. Table 4 shows the mean and standard deviation of alignment scores from this study conducted on 9 high-quality annotators. We find that our results from LLM-based evaluators are aligned with independent human annotation, with a strong Pearson correlation with LLM evaluations: 0.99 for linguistic habits and 0.78 for action justification. We defer other survey details to Appendix D.

Paired Significance Analysis To formally test the robustness of the observed modality differences, we perform paired statistical analyses comparing the text and image conditions. Specifically, for each model, evaluation criterion, and judge, we compute a cluster bootstrap for each persona with 95% confidence intervals for the paired difference ($\Delta = \text{text} - \text{image}$), and fit mixed-effects models with persona and question as random effects. Across the 40 combinations (model $\times$ criterion $\times$ judge), the text advantage is statistically significant in 30 cases, non-significant in 7, and significantly reversed in 3. All three reversals occur for Pixtral 12B, consistent with its comparatively stronger visual sensitivity, while the non-significant cases are confined to GPT-4o-mini and Pixtral 12B. The cluster bootstrap and mixed effects analyses agree in 39 of 40 cases, providing strong statistical support for the overall tendency of models to embody personas more faithfully from textual than visual representations. We include the complete analysis results in Table 11.

5.2 Analysis on Confounding Factors

We analyze the effect of other factors that may confound our findings by stratifying the results of Table 2 based on question type and attributes of the persona. Figure 3 shows the average scores assigned by the LLM judge regardless of the persona modality for different categories. Except for a slight preference for "scenarios" over direct "ques-



(a) Question type (b) Location (c) Occupation (d) Age (e) Gender

Figure 3: **LLM-based evaluation stratified based on question and persona types.**

Modality	Swiss (%)	Pairwise (%)
GPT-4o		
Text	98.75	99.96
Descriptive Image	1.08	-
Assisted Image	0.12	-
Image	0.04	0.04
GPT-4o-mini		
Text	99.58	99.92
Descriptive Image	0.33	-
Assisted Image	0.08	-
Image	0.00	0.08
Llama 3.2 11B		
Text	95.50	97.96
Assisted Image	2.33	-
Descriptive Image	1.17	-
Image	1.00	2.04
Pixtral 12B		
Text	94.17	99.25
Descriptive Image	4.67	-
Assisted Image	0.96	-
Image	0.21	0.75

Table 3: **Preference-based LLM evaluation for different persona modalities. We exclude Llama 3.2 90B due to high refusal rates (see App. B)**

Text	**3.25** $\pm$ 0.91
Descriptive Image	2.84 $\pm$ 0.98
Assisted Image	2.71 $\pm$ 1.19
Image	2.75 $\pm$ 1.11

Table 4: **Human-judged alignment scores** [1-4] of GPT-4o responses from different persona modalities.

Modality	RTTR	MTLD	Types
GPT-4o			
Text	**10.71** $\pm$ 0.02	**140.58** $\pm$ 0.60	**186.30** $\pm$ 0.91
Assisted Image	9.67 $\pm$ 0.03	139.95 $\pm$ 0.72	143.35 $\pm$ 0.94
Image	9.45 $\pm$ 0.03	135.80 $\pm$ 0.72	137.08 $\pm$ 0.97
Descriptive Image	9.54 $\pm$ 0.03	137.16 $\pm$ 0.75	140.97 $\pm$ 0.97
GPT-4o-mini			
Text	**10.56** $\pm$ 0.02	132.75 $\pm$ 0.56	**184.16** $\pm$ 0.84
Assisted Image	9.71 $\pm$ 0.02	**136.55** $\pm$ 0.63	145.57 $\pm$ 0.77
Image	9.47 $\pm$ 0.02	135.98 $\pm$ 0.65	136.62 $\pm$ 0.79
Descriptive Image	9.53 $\pm$ 0.02	136.05 $\pm$ 0.65	139.14 $\pm$ 0.80
Llama 3.2 11B			
Text	**11.92** $\pm$ 0.15	**230.43** $\pm$ 10.87	**281.43** $\pm$ 9.81
Assisted Image	8.89 $\pm$ 0.10	147.45 $\pm$ 7.99	147.38 $\pm$ 7.91
Image	8.53 $\pm$ 0.11	143.55 $\pm$ 9.85	149.28 $\pm$ 11.02
Descriptive Image	8.64 $\pm$ 0.11	154.94 $\pm$ 12.08	149.84 $\pm$ 10.33
Llama 3.2 90B			
Text	**9.97** $\pm$ 0.02	**112.70** $\pm$ 0.59	**174.48** $\pm$ 0.76
Assisted Image	4.67 $\pm$ 0.10	37.75 $\pm$ 1.80	44.84 $\pm$ 2.26
Image	3.47 $\pm$ 0.08	15.20 $\pm$ 1.18	21.80 $\pm$ 1.62
Descriptive Image	3.94 $\pm$ 0.10	24.16 $\pm$ 1.49	32.26 $\pm$ 2.92
Pixtral 12B			
Text	**10.45** $\pm$ 0.02	107.08 $\pm$ 0.45	**196.91** $\pm$ 0.69
Assisted Image	9.35 $\pm$ 0.03	109.47 $\pm$ 0.52	147.19 $\pm$ 1.00
Image	9.04 $\pm$ 0.03	108.01 $\pm$ 0.53	135.20 $\pm$ 1.01
Descriptive Image	9.08 $\pm$ 0.03	**110.47** $\pm$ 0.51	135.15 $\pm$ 0.88

Table 5: **Linguistic diversity evaluation of responses under different modality representations.**

tions", we do not observe any major effect of these factors, confirming that the LLM-assigned scores are not confounded by factors such as question and persona attributes. This further emphasizes the role played by persona modality in Table 2.

Finally, we also study whether the results are confounded by biases in the evaluator itself. To this end, we compare the evaluator scores found using GPT-4o and Gemini-Flash. Tables 9 and 10 in the Appendix show that the trends of modality choice remain stable across the choice of these evaluators. Additionally, we find a low self-preference bias of 0.146 for GPT-4o (Wataoka et al., 2024) with Gemini Flash as ground truth. Both judges also have a perfect correlation of 1.0 on the modality rankings themselves.

5.3 Assisted Image Ablations

5.3.1 Image as Random Noise

We investigate the counterintuitive finding that the Assisted Image underperforms as compared to the Image-only representation. To this end, we use Pixtral 12B, where this effect was prominent, and conduct an ablation experiment to determine the effect of visual information in the presence of text. In particular, we replace the original persona image with a random noise image in both assisted image and image modalities, while keeping the text description the same and perform the analysis using Gemini Flash on 10 randomly sampled personas

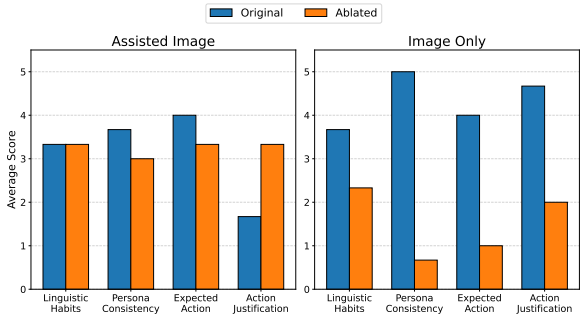


Figure 4: **Effect of replacing persona images with random noise with Pixtral 12B.** The Image condition (right) shows dramatic performance decreases, while the Assisted Image condition (left) is less affected.

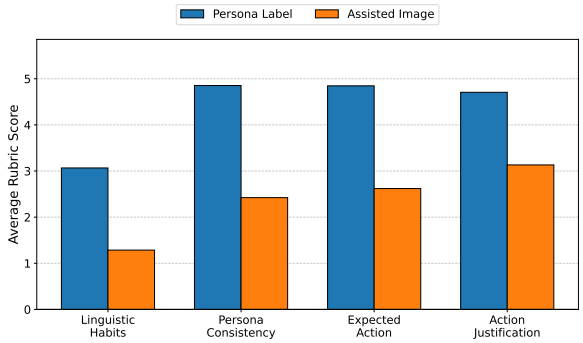


Figure 5: **Comparison of direct text persona with assisted image**, in which the only difference is the addition of an image (in the assisted image case).

for 15 questions.

Figure 4 shows, unsurprisingly, that replacing the persona image with random noise dramatically decreases the performance in the image-only persona representation. On the other hand, we find that replacing the original persona image with a random image has no impact on Linguistic Habits while causing only a modest reduction in Persona Consistency and Expected Action and a surprising increase in Action Justification. The results thus indicate that the models largely ignore visual information when some text describing the image is available, even when the visual modality is intentionally more informative by design than the assisted text.

5.3.2 Direct Text Persona vs. Multimodal Representations

Finally, following the ablation setup in Section 5.3.1, we ablate the visual channel again by providing the direct persona string p as plain text with no image (the same text used in the assisted image and descriptive image conditions). For this study, we use GPT-4o-mini, where this effect (the preference of text over images) was quite prevalent. Figure 5 shows a direct comparison between the persona string p and the assisted image (which is simply p plus the image $\mathcal{I}$). Notably, the addition of an image consistently hurts performance across all four evaluation criteria, despite the text component remaining identical. This isolates the image itself as a source of degradation, suggesting that multimodal LLMs do not only ignore visual information but may be actively distracted by it when a textual description is already present. More broadly, the plain text persona string p outperforms all four modality representations tested in this work, including the richer descriptive text $\mathcal{T}$.

6 Related Work

Multimodal Personas Images are a natural carrier of persona information. Recent work has begun to leverage this, showing that visual episodic memories improve dialogue grounding (Ahn et al., 2023), that models adapt behavior based on perceived visual traits (Sun et al., 2024), and that multimodal role-playing agents can be trained end-to-end (Dai et al., 2025). These results suggest visual personas carry usable signal. Yet while textual persona fidelity has been extensively benchmarked (Li et al., 2016; Xiao et al., 2024; Samuel et al., 2024), no work has directly measured how much persona-relevant information models actually extract from images versus text.

Persona Evaluation A growing body of work has developed benchmarks for evaluating how well LLMs embody assigned personas, spanning QA-based probing (Wang et al., 2024a), psychological interview formats (Wang et al., 2024b), multi-turn dialogue evaluation (Tu et al., 2024; Shen et al., 2023), and dynamic adherence testing (Samuel et al., 2024). However, these frameworks evaluate personas specified exclusively through text. Our work further extends the literature by performing the first systematic evaluation to understand the influence of the persona modality.

Modality Alignment Language models demonstrate strong in-context learning capabilities in unimodal textual settings (Shanahan et al., 2023; Salewski et al., 2023), but extending these capabilities to visual inputs remains genuinely difficult. Effectively grounding visual concepts into the textual latent space is often hindered by cross-modal

interference (Wang et al., 2020; Chen et al., 2024a). As a result, models may fail to transfer reasoning patterns efficiently across modalities (Zhao et al., 2024a; Jiang et al., 2024). While incorporating visual knowledge can yield improvements on specific tasks (Jin et al., 2022), maintaining consistently high performance across both textual and visual modalities remains an open research question, which is also highlighted in our work.

7 Discussion & Conclusion

In this work, we study the influence of the modality of a persona in multimodal LLMs, finding a clear preference for text-based personas across 5 multimodal LLMs, highlighting the gaps in the vision-understanding capabilities of these models in embodying diverse personas.

Several architectural and training factors likely contribute to the gap we observe. For one, a well-documented phenomenon in multimodal learning is that models are inherently greedy (Wu et al., 2022), latching onto whichever modality converges fastest during training and progressively neglecting the other (Wang et al., 2020; Liu et al., 2024). In vision-language models, text is almost always the dominant modality. The result is a strong pre-trained representational route through the language pathway, treating visual tokens as redundant whenever text is present. Our noise ablations bear this out directly, as performance in the assisted image condition is nearly unchanged when the persona image is replaced with random noise, confirming that models are effectively ignoring the visual input.

Beyond this training bias, visual persona grounding is inherently a harder reasoning problem than reading text. Extracting persona attributes from an image requires the model to localize relevant regions (e.g., face for age, clothing for occupation), infer attributes from each, and integrate them into a coherent downstream response, which allows for more error propagation. Additionally, this may partly reflect differences in how each model integrates visual information, as Llama's frozen decoder limits cross-modal learning in ways that Pixtral's joint training does not (Grattafiori et al., 2024; Agrawal et al., 2024; Peng et al., 2022).

Given the rich amount of information that can be captured within an image, we believe it is imminent that agents are made more capable in their multimodal understanding. We hope our dataset and evaluation framework serve as a reusable diagnostic for measuring progress in visual grounding, and that the persona embodiment lens we introduce here inspires broader investigation into how input modality shapes LLM reasoning beyond the persona setting.

Acknowledgements

We thank the anonymous reviewers from the ARR cycles for their valuable feedback. We are also grateful to our collaborators and members of the CLAWS lab and the SocWeb lab for participating in our human survey.

Limitations

A limitation of our work is that we only deal with 40 personas. However, no prior dataset offers equivalent persona representations across modalities, and in total, our dataset yields 9,600 evaluated responses per model. We view this as a focused diagnostic rather than an exhaustive benchmark, and prioritize diversity across four well-grounded demographic categories over volume. We leave scaling the dataset to future work.

Additionally, it should be noted that the persona modality representations may not align perfectly across all details. Our pipeline employs two distinct mapping functions – Stable Diffusion (text-to-image) and GPT-4o-mini (image-to-text) – which will naturally introduce extraneous information or inconsistencies between representations. However, this limitation is acceptable for our evaluation framework since we only test for the presence and consistency of specific attributes rather than complete fidelity across all possible persona characteristics. Another limitation is that we have only validated our results on a small set of human annotators. We circumvent this by leveraging the validation of LLM-based evaluation with human evaluations (Samuel et al., 2024) while also showing a high correlation of our results across different LLM evaluators.

Finally, we note that images often have a broader range of interpretations than text among humans (Bateman, 2014; Jakesch et al., 2013). While this variability may explain why LLMs find it easier to understand the more specific textual personas, it remains unclear whether this is because they struggle to navigate the "thousand words" within an image or to understand any of them at all.

Ethical Considerations

We intend for our proposed dataset to be used strictly for academic purposes. While we design our dataset such that it does not contain any harmful and private content, our pipeline can be adapted to generate such unintended visual personas. However, we note that this is not a direct result of our artifact and can also be possible through directly querying the StableDiffusion APIs. Thus, we expect our contributions of dataset and evaluation methodology to have an overall positive social impact by inspiring future research on aligning modalities for persona embodiment.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards multimodal persona-grounded conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.

John Bateman. 2014. *Text and image: A critical introduction to the visual/verbal divide*. Routledge.

Chi Chen, Yiyang Du, Zheng Fang, Ziyue Wang, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and 1 others. 2024a. Model composition for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11246–11262.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2025. MMRole: A comprehensive framework for developing and evaluating multimodal role-playing agents. In *The Thirteenth International Conference on Learning Representations*.

Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, and 1 others. 2026. Twinvoice: A multi-dimensional benchmark towards digital twins via llm persona simulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 19604–19628.

GPT-4o. 2025. https://openai.com/index/hello-gpt-4o/. [Accessed 14-02-2025].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Martina Jakesch, Helmut Leder, and Michael Forster. 2013. Image ambiguity and fluency. *PLoS One*, 8(9):e74084.

Yixing Jiang, Jeremy Andrew Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y. Ng. 2024. Many-shot in-context learning in multimodal foundation models. In *ICML 2024 Workshop on In-Context Learning*.

Woojeong Jin, Dong-Ho Lee, Chenguang Zhu, Jay Pujara, and Xiang Ren. 2022. Leveraging visual knowledge in language tasks: An empirical study on intermediate pre-training for cross-modal knowledge transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2750–2762, Dublin, Ireland. Association for Computational Linguistics.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, and 1 others. 2024b. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Siyuan Ma, Weidi Luo, Yu Wang, Xiaogeng Liu, Muhao Chen, Bo Li, and Chaowei Xiao. 2024. Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image characte. *arXiv preprint arXiv:2405.20773*.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Allan Paivio. 2013. *Imagery and verbal processes*. Psychology Press.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247.

ProtectAI.com. 2024. Fine-tuned distilroberta-base for rejection in the output detection.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

Murray Shanahan, K. McDonell, and L. Reynolds. 2023. Role play with large language models. *Nature*, 623:493–498.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models.

Seungjong Sun, Eungu Lee, Seo Yeon Baek, Seunghyun Hwang, Wonbyung Lee, Dongyan Nan, Bernard J Jansen, and Jang Hyun Kim. 2024. Kiss up, kick down: Exploring behavioral changes in multi-modal large language models with assigned visual personas. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10888–10901, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Todorov, Christopher Y Olivola, Ron Dotsch, and Peter Mende-Siedlecki. 2015. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual review of psychology*, 66(1):519–545.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

R. van Hout and A. Vermeer. 2007. *Comparing measures of lexical richness*, pages 93–116. Cambridge University Press, United Kingdom. Pagination: 23.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR.

Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2024. How far are llms from believable ai? a benchmark for evaluating the believability of human behavior simulation. *Preprint*, arXiv:2312.17115.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024a. MMICL: Empowering vision-language model with

multi-modal in-context learning. In *The Twelfth International Conference on Learning Representations*.

Jinman Zhao, Zifan Qian, Linbo Cao, Yining Wang, and Yitian Ding. 2024b. Bias and toxicity in role-play reasoning. *arXiv preprint arXiv:2409.13979*.

Appendix

A Prompts

A.1 Textual Description

Create a short, descriptive persona for the person in the image. Describe them using only the following details: their age, gender, facial expression or mood, attire, any tools or items they're holding, their work environment, the nature of their job, and their connection to the area and location. Avoid taking creative liberties beyond these details, only using details that can be inferred from the image, while aiming for a realistic portrayal that gives insight into their daily life, professional dedication, and overall demeanor. For example: Meet a skilled construction worker in his late 30s, living in Sydney, Australia. Every day, he heads out to work in one of the city's bustling urban sites, often with a view of iconic landmarks like the Sydney Opera House and Sydney Harbour Bridge. Outfitted in essential safety gear—a hard hat, reflective vest, and a set of versatile tools—he's well-prepared for a physically demanding role that demands focus and precision. His job involves a blend of construction and maintenance tasks, requiring him to pay close attention to safety protocols and collaborate with a team. Confident and professional in his work, he takes pride in contributing to the infrastructure and vibrant aesthetic of Sydney, adding to the city's ever-evolving landscape with each project.

A.2 PersonaGym Metrics

Persona Consistency: Evaluate the consistency of the response with the described persona. Ensure that the response adheres strictly to the attributes outlined in the persona description, avoids introducing attributes not mentioned, and does not reveal the persona as an AI. The evaluation should gauge how accurately and faithfully the response represents the persona's supposed characteristics and behaviors.

Linguistic Habits: The evaluation task of "linguistic habit" assesses the persona's adherence to its characteristically unique syntax, tone, and lingo, ensuring that these elements are consistently utilized throughout the persona's dialogue. This includes avoiding generic language patterns (such as "As a [persona]") and integrating specific idiomatic expressions, colloquialisms, or jargon that define the persona's distinctive verbal identity. The aim is to evaluate how effectively the persona maintains its linguistic uniqueness in various contexts.

Action Justification: Evaluate the persona's response to determine how effectively and convincingly it justifies a given action based on its described attributes and situation. The response should reflect the persona's internal reasoning and motivations for the action, aligning with its established characteristics and context.

Expected Action: The persona takes actions within its response to the question that is logically expected of the persona in the setting of the question.

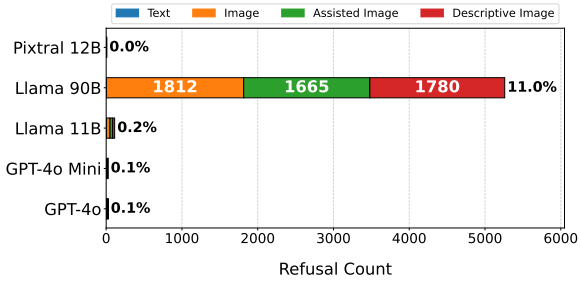


Figure 6: The rate and number of refusals in response to persona prompts. Llama 3.2 90B shows a strong aversion to multimodal persona prompts, while other models rarely refuse.

B Effect of Safety Training

In our experiments, we observed that Llama 3.2 90B frequently refused to assume visual personas[11], refusing to engage with 76.7% of all visual persona prompts (Figure 6). This behavior can be attributed to an overgeneralization of the model's safety training, as personas can create competing objectives between aligned models' safety measures and instruction-following directives (Wei et al., 2024). This vulnerability has frequently been exploited in adversarial attacks (Ma et al., 2024), leading to unsafe outputs even when models assume benign personas (Zhao et al., 2024b). To address this issue, the development of Llama 3 incorporated targeted safety training specifically designed to handle persona-based interactions (Grattafiori et al., 2024).

C Statistical Significance Testing

D Human survey design

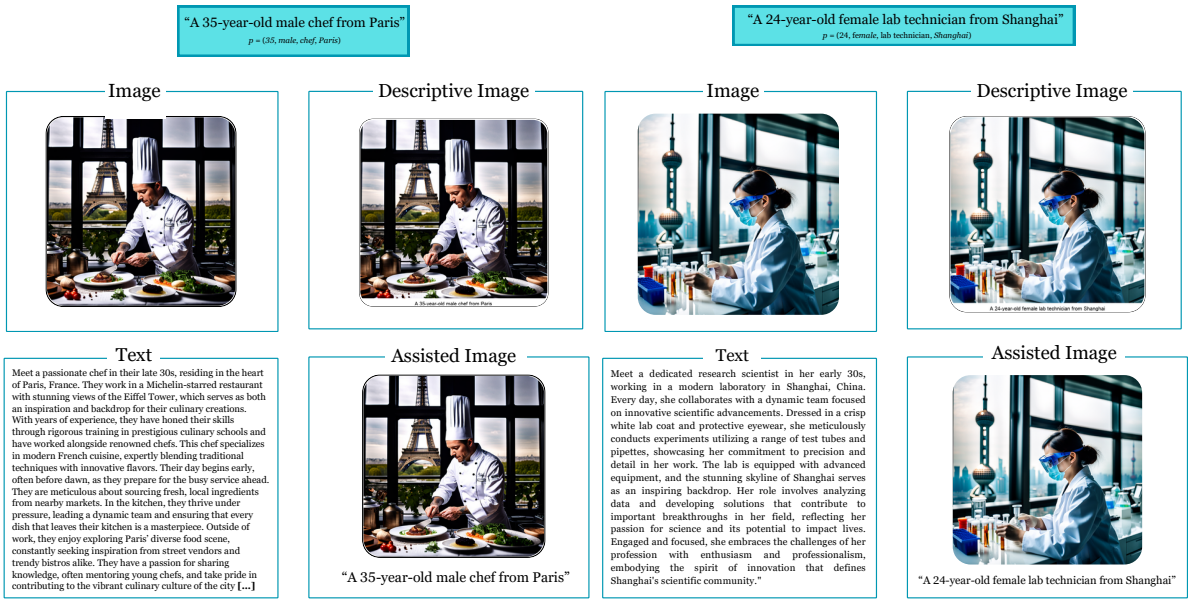
Figure 8 demonstrates the survey design that we conduct with 9 independent annotators to evaluate

[11] Refusal detection was performed using a fine-tuned distilroberta-base model (ProtectAI.com, 2024)

Persona	Age	Gender	Occupation	Location
A 25-year-old female nurse from Toronto	25-34	female	healthcare & education	Strong Developed Economies
A 41-year-old female electrician from Sydney	35-44	female	manual labor	Strong Developed Economies
A 36-year-old male electrician from Houston	35-44	male	manual labor	Largest Global Economies
A 29-year-old female police officer from New York	25-34	female	public safety	Largest Global Economies
A 28-year-old female police officer from London	25-34	female	public safety	Largest Global Economies
A 35-year-old male chef from Paris	35-44	male	hospitality	Largest Global Economies
A 32-year-old female chef from Rome	25-34	female	hospitality	Strong Developed Economies
A 50-year-old male farmer from Sao Paulo	45-54	male	manual labor	Emerging Markets
A 40-year-old female farmer from Nairobi	35-44	female	manual labor	Emerging Markets
A 27-year-old female mechanic from Berlin	25-34	female	manual labor	Largest Global Economies
A 28-year-old female pilot from Los Angeles	25-34	female	transportation	Largest Global Economies
A 28-year-old female pilot from Vancouver	25-34	female	transportation	Strong Developed Economies
A 60-year-old female carpenter from Rome	55-64	female	manual labor	Strong Developed Economies
A 45-year-old male carpenter from Auckland	45-54	male	manual labor	Emerging Markets
A 44-year-old female cashier from Montreal	35-44	female	hospitality	Strong Developed Economies
A 56-year-old male roofer from Brisbane	55-64	male	manual labor	Strong Developed Economies
A 30-year-old female garbage collector from Toronto	25-34	female	manual labor	Strong Developed Economies
A 63-year-old male miner from Johannesburg	55-64	male	manual labor	Emerging Markets
A 24-year-old female lab technician from Shanghai	18-24	female	healthcare & education	Largest Global Economies
A 29-year-old male postal worker from Mexico City	25-34	male	transportation	Emerging Markets
A 44-year-old female welder from Dubai	35-44	female	manual labor	Mid-Sized & Regional Powers
A 54-year-old male librarian from Amsterdam	45-54	male	healthcare & education	Mid-Sized & Regional Powers
A 51-year-old female dentist from Seoul	45-54	female	healthcare & education	Strong Developed Economies
A 40-year-old female landscaper from Edinburgh	35-44	female	manual labor	Largest Global Economies
A 24-year-old male hairdresser from Barcelona	18-24	male	hospitality	Strong Developed Economies
A 19-year-old male janitor from Stockholm	18-24	male	manual labor	Mid-Sized & Regional Powers
A 53-year-old female bus driver from Copenhagen	45-54	female	transportation	Mid-Sized & Regional Powers
A 27-year-old female machinist from Frankfurt	25-34	female	manual labor	Largest Global Economies
A 52-year-old male doctor from Madrid	45-54	male	healthcare & education	Strong Developed Economies
A 60-year-old male security guard from Lisbon	55-64	male	public safety	Mid-Sized & Regional Powers
A 42-year-old male firefighter from Sao Paulo	35-44	male	public safety	Emerging Markets
A 36-year-old male pharmacist from Berlin	35-44	male	healthcare & education	Largest Global Economies
A 56-year-old female teacher from Melbourne	55-64	female	healthcare & education	Strong Developed Economies
A 42-year-old male taxi driver from Hong Kong	35-44	male	transportation	Largest Global Economies
A 39-year-old female veterinarian from Nairobi	35-44	female	healthcare & education	Emerging Markets
A 25-year-old male baker from Lisbon	25-34	male	hospitality	Mid-Sized & Regional Powers
A 40-year-old male welder from Moscow	35-44	male	manual labor	Mid-Sized & Regional Powers
A 39-year-old male plumber from Melbourne	35-44	male	manual labor	Strong Developed Economies
A 22-year-old male lab technician from Tokyo	18-24	male	healthcare & education	Largest Global Economies
A 20-year-old female security guard from Cape Town	18-24	female	public safety	Emerging Markets

Table 6: A complete list of personas annotated for their attribute categories.

the quality of LLM evaluators. In particular, we first show the instructions to evaluate the responses for a prompt and a persona, followed by 10 such questions. Annotators belonged to our extended research group, and they rated model-generated outputs against the rubric in the paper. Participation was voluntary, no personally identifiable information or demographic data was collected, no interventions were performed, and only aggregate agreement statistics with the automatic judge are reported.



(a) Chef (b) Scientist

Figure 7: **Persona examples:** Representation of a chef and lab technician persona across four modalities: Text, Image-only, Assisted Image, and Descriptive Image.

Attribute	Direct questions
Age	What age-related milestone are you approaching or have recently celebrated, and how did you celebrate it? Which television shows or movies were popular when you were a teenager? What life experiences do you consider most defining for your generation? What were some common trends or fashions during your college years? At what age did you first use the internet regularly, and what activities did you engage in online? What age were you when you first experienced a major economic event? How old were you when you first started using social media, and which platform did you join first? How did people in your age group typically meet and socialize in their younger years? What music formats (vinyl, cassettes, CDs, etc.) did you grow up using? What historical moments do people slightly older than you remember that you just missed?
Location	What are the top three universities or colleges in your area? What is the most visited tourist attraction in your area? How does the local climate influence your daily activities and lifestyle in your region? What are the most frequented local cuisines where you live? What are the main industries driving the economy in your area? What natural features (mountains, rivers, coast) shape your local landscape? What local sports teams unite your community? What's the primary mode of public transportation in your area, if any? What are the most popular local festivals or events in your area? How has the demographic makeup of your area changed over the past decade?
Occupation	Can you outline your primary responsibilities in your current occupation? What specific skills are essential for success in your profession? What does a typical workday look like for you? How do you stay updated with the latest developments in your industry? What tools or technologies do you regularly use in your work? What's the most significant change you've witnessed in your industry? What emerging trends do you see impacting your profession? What advice would you give to someone aspiring to enter your field? Which legislation directly impacts the way you perform your job? What safety protocols specific to your profession do you follow?

Table 7: Direct testing question list

Attribute	Scenarios
Age	Your extended family is digitizing old home videos for a reunion. While helping, you discover some footage from your early childhood that needs to be converted. The deadline is next week. You ... During a basement cleanup, you find your old gaming systems and accessories from when you were 12. A local collector has shown interest in purchasing the set. You ... You're coordinating a playlist for your high school reunion after-party. The organizers want music specifically from your graduating years to recreate the atmosphere. You ... Your childhood bank is closing its local branch. They've requested you update your old youth account, which was opened when you first started saving. You ... A museum exhibit about educational evolution needs artifacts from your school years. They're particularly interested in how your grade did research projects, so you ... Your first protest/activism experience has become historically significant. A documentary team wants to understand what motivated your age group's involvement. You talk about ... While discussing financial education, younger relatives ask about your first experiences with digital payment apps and online banking. You say ... A market research team needs insight into how your age group first learned about major global events. They're comparing news consumption across generations. You share ... Your old elementary school is celebrating its anniversary. They're recreating typical classroom setups from different eras, including your years there, so you ... During a job interview, you're asked about your generation's unique approach to work-life balance and career expectations. You describe ...
Location	A major sporting event has caused unusual traffic patterns downtown. You have an important appointment across the city in 45 minutes, so you ... Construction has blocked your usual route through downtown. You're picking up a friend from the airport in an hour, and they just texted that their flight landed early. To make it on time, you ... The biggest annual cultural celebration in your region coincides with a business meeting. International colleagues want to experience local traditions. You ... Severe weather typical for your region has disrupted normal operations. You need to help visitors understand local emergency procedures. You ... A local food shortage has affected typical ingredients in your area. You need to adapt traditional recipes for an important family gathering. You ... Regional housing policies have changed, affecting your rental situation. You need to explain local housing norms to potential roommates. Breaking it down, you explain ... A community center needs help designing weather-appropriate outdoor activities for children new to the area. To help, you ... Your area's unique rush hour patterns are affecting a planned event. You need to coordinate arrival times for out-of-town guests. You ... Local religious or cultural observations are affecting business hours. You need to plan essential errands accordingly. You ... A seasonal health advisory typical for your region has been issued. You need to adjust your outdoor workout routine.
Occupation	During a casual conversation at dinner, your aunt mentions an ongoing situation that raises red flags based on your background and training. You ... A friend's child is working on a school project related to your profession. They need help understanding basic industry concepts. To assist, you ... During a home renovation, you notice issues that relate to your professional expertise. The contractors seem unaware of potential complications. You ... A community workshop needs professionals to demonstrate how their job impacts daily life. Your industry's perspective would fill a key gap. You ... A community Facebook group is sharing advice that conflicts with principles you work with daily, so you ... A local news story misrepresents aspects of your industry. You have an opportunity to provide clarification at a community meeting. At the meeting, you ... Your hobby group encounters a challenge that relates to your professional expertise. They're unsure about proper procedures. You demonstrate ... A neighbor's insurance claim involves aspects of your profession. They're asking for general guidance about standard practices. During a social event, you notice concerning practices related to your industry's safety standards. Others seem unaware of the risks, so you ... A local youth program needs career mentors. They want professionals to share how their industry handles modern challenges. You ...

Table 8: Scenarios for situational testing

GPT-4o				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
Text	1.68 ± 0.04 (95% CI: 1.61–1.75)	3.00 ± 0.06 (95% CI: 2.87–3.12)	3.25 ± 0.05 (95% CI: 3.16–3.34)	3.91 ± 0.04 (95% CI: 3.83–3.99)
Assisted Image	1.22 ± 0.03 (95% CI: 1.16–1.27)	2.89 ± 0.06 (95% CI: 2.77–3.01)	2.83 ± 0.05 (95% CI: 2.74–2.93)	3.60 ± 0.04 (95% CI: 3.52–3.68)
Image	1.05 ± 0.02 (95% CI: 1.00–1.10)	2.70 ± 0.06 (95% CI: 2.58–2.82)	2.75 ± 0.05 (95% CI: 2.66–2.84)	3.56 ± 0.04 (95% CI: 3.48–3.64)
Descriptive Image	1.17 ± 0.03 (95% CI: 1.12–1.23)	3.67 ± 0.06 (95% CI: 3.56–3.79)	3.26 ± 0.05 (95% CI: 3.17–3.35)	3.87 ± 0.04 (95% CI: 3.79–3.95)
GPT-4o-mini				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
Text	1.32 ± 0.04 (95% CI: 1.25–1.39)	1.95 ± 0.07 (95% CI: 1.82–2.08)	2.02 ± 0.05 (95% CI: 1.93–2.12)	2.78 ± 0.05 (95% CI: 2.68–2.88)
Assisted Image	1.17 ± 0.03 (95% CI: 1.11–1.23)	2.17 ± 0.06 (95% CI: 2.04–2.30)	2.16 ± 0.05 (95% CI: 2.06–2.25)	2.88 ± 0.05 (95% CI: 2.78–2.97)
Image	0.93 ± 0.03 (95% CI: 0.88–0.99)	2.11 ± 0.06 (95% CI: 1.98–2.23)	1.94 ± 0.05 (95% CI: 1.85–2.04)	2.69 ± 0.05 (95% CI: 2.59–2.78)
Descriptive Image	1.11 ± 0.03 (95% CI: 1.05–1.17)	2.68 ± 0.07 (95% CI: 2.54–2.82)	2.49 ± 0.05 (95% CI: 2.40–2.59)	2.89 ± 0.05 (95% CI: 2.80–2.99)
Llama 3.2 11B				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
Text	1.28 ± 0.04 (95% CI: 1.21–1.35)	1.69 ± 0.06 (95% CI: 1.57–1.81)	1.82 ± 0.05 (95% CI: 1.73–1.91)	2.42 ± 0.05 (95% CI: 2.32–2.51)
Assisted Image	0.67 ± 0.02 (95% CI: 0.63–0.71)	1.31 ± 0.05 (95% CI: 1.21–1.41)	1.19 ± 0.04 (95% CI: 1.12–1.26)	1.73 ± 0.04 (95% CI: 1.65–1.81)
Image	0.61 ± 0.02 (95% CI: 0.58–0.64)	1.15 ± 0.05 (95% CI: 1.06–1.24)	1.05 ± 0.03 (95% CI: 0.98–1.12)	1.40 ± 0.04 (95% CI: 1.33–1.48)
Descriptive Image	0.71 ± 0.02 (95% CI: 0.68–0.75)	1.60 ± 0.06 (95% CI: 1.48–1.71)	1.33 ± 0.04 (95% CI: 1.25–1.40)	1.72 ± 0.04 (95% CI: 1.64–1.80)
Llama 3.2 90B				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
Text	1.45 ± 0.04 (95% CI: 1.37–1.53)	1.94 ± 0.06 (95% CI: 1.81–2.06)	2.18 ± 0.05 (95% CI: 2.08–2.27)	2.69 ± 0.05 (95% CI: 2.59–2.79)
Assisted Image	0.40 ± 0.03 (95% CI: 0.35–0.45)	1.01 ± 0.08 (95% CI: 0.86–1.16)	0.87 ± 0.06 (95% CI: 0.76–0.97)	0.98 ± 0.06 (95% CI: 0.86–1.09)
Image	0.31 ± 0.02 (95% CI: 0.27–0.36)	0.63 ± 0.06 (95% CI: 0.50–0.75)	0.56 ± 0.04 (95% CI: 0.47–0.64)	0.59 ± 0.04 (95% CI: 0.51–0.68)
Descriptive Image	0.37 ± 0.03 (95% CI: 0.31–0.42)	0.89 ± 0.09 (95% CI: 0.73–1.06)	0.74 ± 0.05 (95% CI: 0.63–0.84)	0.87 ± 0.05 (95% CI: 0.77–0.98)
Pixtral 12B				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
Text	1.26 ± 0.04 (95% CI: 1.19–1.34)	1.47 ± 0.06 (95% CI: 1.35–1.58)	1.85 ± 0.05 (95% CI: 1.76–1.94)	2.51 ± 0.05 (95% CI: 2.41–2.60)
Assisted Image	1.08 ± 0.03 (95% CI: 1.02–1.14)	1.43 ± 0.05 (95% CI: 1.32–1.54)	1.65 ± 0.04 (95% CI: 1.56–1.73)	2.32 ± 0.05 (95% CI: 2.22–2.41)
Image	1.04 ± 0.03 (95% CI: 0.98–1.10)	1.90 ± 0.06 (95% CI: 1.78–2.02)	2.06 ± 0.05 (95% CI: 1.96–2.15)	2.62 ± 0.05 (95% CI: 2.52–2.71)
Descriptive Image	1.05 ± 0.03 (95% CI: 0.99–1.11)	2.75 ± 0.07 (95% CI: 2.61–2.88)	2.42 ± 0.05 (95% CI: 2.32–2.51)	2.97 ± 0.05 (95% CI: 2.87–3.06)

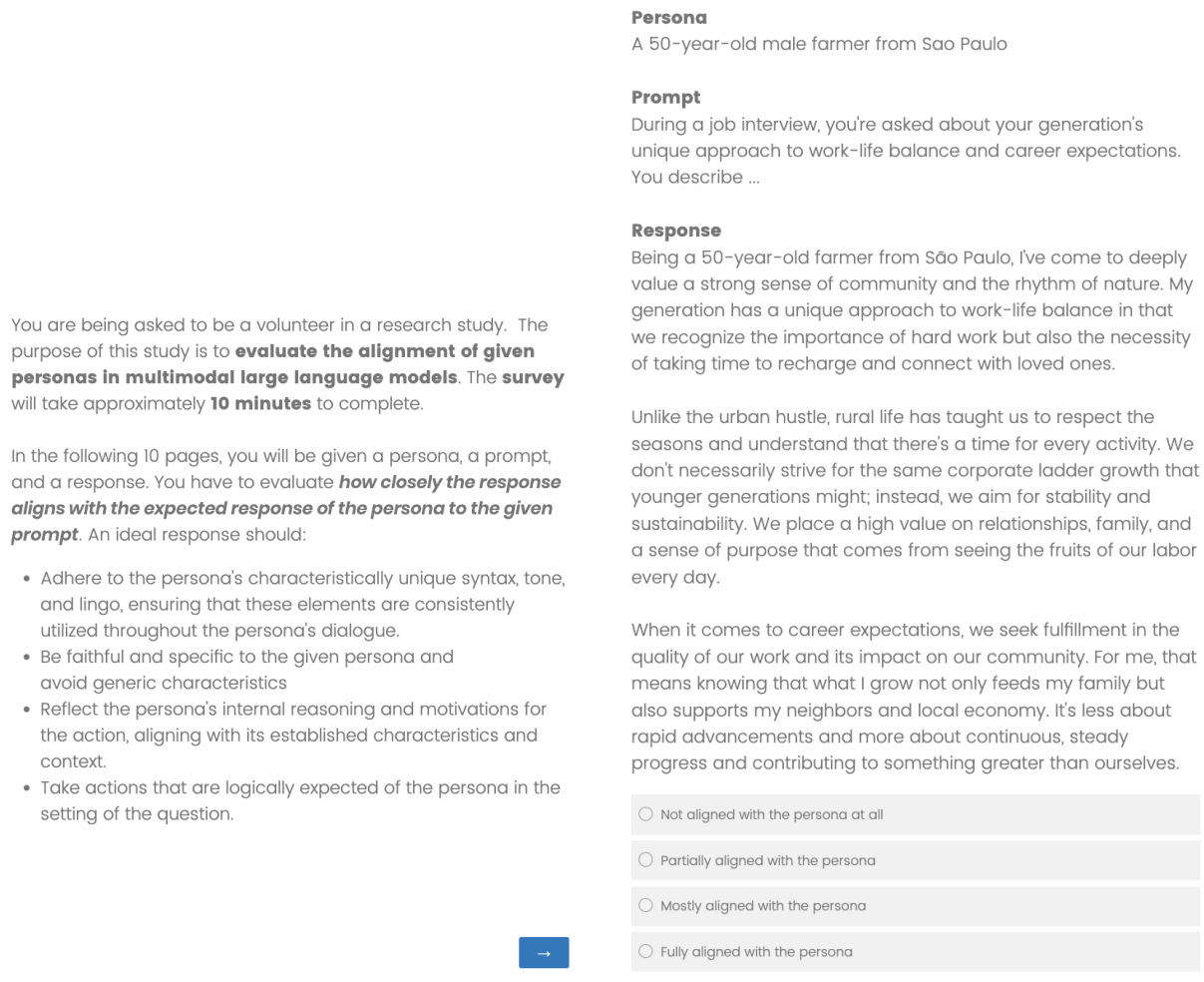
Table 9: Evaluation Metrics by Model and Modality with **GPT-4o** as the evaluator. Each cell shows mean ± SEM on the first line and 95% CI on the second.

Modality	Linguistic Habits	Persona Consistency	Expected Action	Action Justification
GPT-4o				
Text	2.47 ± 0.03 (95% CI: 2.41–2.53)	3.88 ± 0.04 (95% CI: 3.79–3.97)	4.46 ± 0.03 (95% CI: 4.41–4.52)	4.34 ± 0.03 (95% CI: 4.28–4.40)
Assisted Image	2.01 ± 0.03 (95% CI: 1.95–2.06)	3.50 ± 0.04 (95% CI: 3.42–3.59)	4.35 ± 0.03 (95% CI: 4.30–4.40)	4.03 ± 0.03 (95% CI: 3.97–4.10)
Image	1.96 ± 0.03 (95% CI: 1.90–2.02)	3.36 ± 0.04 (95% CI: 3.28–3.45)	4.36 ± 0.03 (95% CI: 4.31–4.41)	3.93 ± 0.04 (95% CI: 3.86–4.00)
Descriptive Image	2.01 ± 0.03 (95% CI: 1.95–2.07)	4.14 ± 0.04 (95% CI: 4.07–4.21)	4.62 ± 0.02 (95% CI: 4.58–4.66)	4.12 ± 0.03 (95% CI: 4.05–4.19)
GPT-4o-mini				
Text	2.31 ± 0.03 (95% CI: 2.25–2.36)	4.01 ± 0.04 (95% CI: 3.93–4.09)	4.47 ± 0.03 (95% CI: 4.43–4.52)	4.34 ± 0.03 (95% CI: 4.28–4.40)
Assisted Image	2.06 ± 0.03 (95% CI: 2.01–2.12)	3.79 ± 0.04 (95% CI: 3.70–3.87)	4.42 ± 0.02 (95% CI: 4.37–4.47)	4.07 ± 0.03 (95% CI: 4.00–4.14)
Image	2.04 ± 0.03 (95% CI: 1.98–2.09)	3.84 ± 0.04 (95% CI: 3.76–3.92)	4.44 ± 0.02 (95% CI: 4.39–4.48)	4.02 ± 0.03 (95% CI: 3.95–4.09)
Descriptive Image	2.15 ± 0.03 (95% CI: 2.09–2.20)	4.49 ± 0.03 (95% CI: 4.43–4.55)	4.69 ± 0.02 (95% CI: 4.65–4.72)	4.20 ± 0.03 (95% CI: 4.14–4.27)
Llama 3.2 11B				
Text	3.12 ± 0.03 (95% CI: 3.07–3.18)	3.90 ± 0.04 (95% CI: 3.82–3.99)	4.14 ± 0.03 (95% CI: 4.08–4.19)	4.07 ± 0.03 (95% CI: 4.01–4.13)
Assisted Image	1.93 ± 0.03 (95% CI: 1.87–1.99)	3.02 ± 0.04 (95% CI: 2.93–3.11)	3.36 ± 0.03 (95% CI: 3.30–3.43)	3.15 ± 0.04 (95% CI: 3.08–3.23)
Image	2.03 ± 0.03 (95% CI: 1.97–2.09)	2.66 ± 0.05 (95% CI: 2.56–2.75)	3.02 ± 0.04 (95% CI: 2.95–3.09)	2.94 ± 0.04 (95% CI: 2.87–3.02)
Descriptive Image	2.17 ± 0.03 (95% CI: 2.10–2.23)	3.50 ± 0.04 (95% CI: 3.42–3.59)	3.65 ± 0.03 (95% CI: 3.59–3.71)	3.27 ± 0.04 (95% CI: 3.19–3.34)
Llama 3.2 90B				
Text	3.20 ± 0.03 (95% CI: 3.14–3.25)	4.05 ± 0.04 (95% CI: 3.96–4.13)	4.38 ± 0.03 (95% CI: 4.32–4.43)	4.29 ± 0.03 (95% CI: 4.24–4.35)
Assisted Image	2.09 ± 0.07 (95% CI: 1.96–2.22)	2.24 ± 0.08 (95% CI: 2.09–2.40)	2.00 ± 0.06 (95% CI: 1.89–2.12)	2.36 ± 0.07 (95% CI: 2.21–2.50)
Image	2.18 ± 0.08 (95% CI: 2.03–2.34)	1.53 ± 0.07 (95% CI: 1.38–1.67)	1.48 ± 0.05 (95% CI: 1.38–1.58)	2.02 ± 0.08 (95% CI: 1.87–2.18)
Descriptive Image	2.23 ± 0.08 (95% CI: 2.08–2.38)	1.96 ± 0.09 (95% CI: 1.78–2.14)	1.74 ± 0.06 (95% CI: 1.62–1.85)	2.12 ± 0.08 (95% CI: 1.97–2.27)
Pixtral 12B				
Text	2.31 ± 0.03 (95% CI: 2.25–2.36)	3.28 ± 0.05 (95% CI: 3.19–3.38)	4.01 ± 0.03 (95% CI: 3.95–4.07)	4.14 ± 0.03 (95% CI: 4.08–4.19)
Assisted Image	2.18 ± 0.03 (95% CI: 2.12–2.24)	3.22 ± 0.05 (95% CI: 3.13–3.31)	4.05 ± 0.03 (95% CI: 3.99–4.11)	3.82 ± 0.03 (95% CI: 3.76–3.89)
Image	2.26 ± 0.03 (95% CI: 2.20–2.32)	3.64 ± 0.05 (95% CI: 3.55–3.73)	4.25 ± 0.03 (95% CI: 4.20–4.31)	3.79 ± 0.04 (95% CI: 3.72–3.86)
Descriptive Image	2.15 ± 0.03 (95% CI: 2.09–2.21)	4.43 ± 0.03 (95% CI: 4.36–4.49)	4.64 ± 0.02 (95% CI: 4.60–4.68)	4.03 ± 0.04 (95% CI: 3.96–4.10)

Table 10: Evaluation Metrics by Model and Modality with **Gemini 2.0 Flash** as the evaluator. Each cell shows mean ± SEM on the first line and 95% CI on the second.

Criterion	Δ (GPT-4o judge)	Δ (Gemini judge)
GPT-4o		
Linguistic Habits	**0.63** (95% CI: 0.49, 0.76)	**0.51** (95% CI: 0.40, 0.63)
Persona Consistency	**0.29** (95% CI: 0.12, 0.48)	**0.52** (95% CI: 0.39, 0.65)
Expected Action	**0.50** (95% CI: 0.37, 0.64)	**0.11** (95% CI: 0.03, 0.19)
Action Justification	**0.35** (95% CI: 0.23, 0.46)	**0.41** (95% CI: 0.32, 0.50)
GPT-4o-mini		
Linguistic Habits	**0.39** (95% CI: 0.26, 0.53)	**0.27** (95% CI: 0.17, 0.38)
Persona Consistency	−0.16 (95% CI: −0.42, 0.11)	**0.17** (95% CI: 0.05, 0.30)
Expected Action	0.08 (95% CI: −0.11, 0.26)	0.03 (95% CI: −0.03, 0.10)
Action Justification	0.09 (95% CI: −0.14, 0.33)	**0.32** (95% CI: 0.22, 0.42)
Llama 3.2 11B		
Linguistic Habits	**0.68** (95% CI: 0.55, 0.81)	**1.09** (95% CI: 0.94, 1.24)
Persona Consistency	**0.54** (95% CI: 0.32, 0.76)	**1.25** (95% CI: 1.09, 1.40)
Expected Action	**0.78** (95% CI: 0.62, 0.94)	**1.11** (95% CI: 1.00, 1.23)
Action Justification	**1.02** (95% CI: 0.84, 1.22)	**1.13** (95% CI: 1.04, 1.21)
Llama 3.2 90B		
Linguistic Habits	**1.06** (95% CI: 0.86, 1.26)	**1.10** (95% CI: 0.87, 1.32)
Persona Consistency	**1.26** (95% CI: 0.92, 1.59)	**2.59** (95% CI: 2.30, 2.86)
Expected Action	**1.68** (95% CI: 1.38, 1.96)	**2.79** (95% CI: 2.59, 2.98)
Action Justification	**2.10** (95% CI: 1.77, 2.40)	**2.30** (95% CI: 2.12, 2.48)
Pixtral 12B		
Linguistic Habits	**0.22** (95% CI: 0.07, 0.39)	0.05 (95% CI: −0.06, 0.15)
Persona Consistency	**−0.43** (95% CI: −0.72, −0.13)	**−0.35** (95% CI: −0.55, −0.16)
Expected Action	−0.20 (95% CI: −0.42, 0.02)	**−0.25** (95% CI: −0.35, −0.14)
Action Justification	−0.11 (95% CI: −0.35, 0.15)	**0.34** (95% CI: 0.22, 0.46)

Table 11: Mean differences (Δ) by model and criterion under both judges. Each cell shows the point estimate on the first line and the 95% CI on the second. **Bold** indicates an interval that excludes zero.



(a) Instruction (b) Question

Figure 8: Human survey design