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

Anonymous ACL submission

Abstract

Large language models (LLMs) have recently demonstrated remarkable advancements in embodying diverse personas, enhancing their effectiveness as conversational agents and virtual assistants. Consequently, LLMs have made significant strides in processing and integrating multimodal information. However, even though human personas can be expressed in both text and image, the extent to which the modality of a persona impacts the embodiment by the LLM remains largely unexplored. In this paper, we investigate how do different modalities influence the expressiveness of personas in multimodal LLMs. To this end, we create a novel modality-parallel dataset of 40 diverse personas varying in age, gender, occupation, and location. This consists of four modalities to equivalently represent a persona: image-only, text-only, a combination of image and small text, and typographical images, where text is visually stylized to convey persona-related attributes. We then create a systematic evaluation framework with 60 questions and corresponding metrics to assess how well LLMs embody each persona across its attributes and scenarios. Comprehensive experiments on 5 multimodal LLMs show that personas represented by detailed text show more linguistic habits, while typographical images often show more consistency with the persona. Our results reveal that LLMs often overlook persona-specific details conveyed through images, highlighting underlying limitations and paving the way for future research to bridge this gap.

1 Introduction

"An image is worth a thousand words", the adage describes how visuals have the unique capacity to encapsulate not only long texts but also more abstract concepts (Paivio, 2013). Large language models (LLMs) are increasingly being adopted as role-playing agents for specific tasks (Li et al., 2024b) and personalized conversational agents (Tseng et al., 2024). However, these personas are often represented using only elaborate textual descriptions (Wang et al., 2024a; Shen et al., 2023; Samuel et al., 2024). As these models become more capable of handling different modalities (Liu et al., 2024), it is imminent that they can express these personas by incorporating information from representative images (Ahn et al., 2023). Thus, it becomes crucial to conduct a systematic study to study the contribution of the persona modality in specifying these personas to the LLMs.

Figure 1: A comparison of visual and textual persona interactions for a chef from Paris. The left side presents an image persona, while the right side features a text persona derived from the image.

Multimodal LLMs have pushed the frontier of virtual assistants by enabling realistic image and voice-based interactions (Liu et al., 2024; Li et al., 2024a; GPT-4o, 2025). These advancements have enabled processing and generating content across multiple modalities, bridging the gap between text-based understanding and richer, more immersive experiences. However, significant gaps remain in

these models' ability to accurately capture visual information, leading to a subpar performance on visual understanding and reasoning tasks (Tong et al., 2024; Chen et al., 2024). Given the fast adoption of LLMs as all-purpose agents, it is important to understand the extent to which these models can accurately capture visual personas.

LLMs have shown remarkable capabilities in manifesting given roles/personas, as highlighted by their ability to succinctly answer specific questions by adapting their styles according to the prescribed personas (Tu et al., 2024; Tseng et al., 2024; Samuel et al., 2024). Furthermore, it has been shown that one can further improve the models' role-playing and personalization capabilities by incorporating both visual and textual information to create a multimodal persona (Sun et al., 2024; Ahn et al., 2023; Dai et al., 2025). However, no systematic comparison exists between personas represented in different modalities.

In this work, we present the first comprehensive study of the influence of the modality of a persona on its embodiment by multimodal LLMs. Figure 1 illustrates our analysis with an example and our contributions can be summarized as follows:

1. We conduct the **first systematic study** on how a persona's modality affects how it is expressed by multimodal LLMs.
2. We create a **novel text-image parallel dataset** of 40 diverse personas along with 60 probing questions about their attributes and scenarios.
3. **Comprehensive evaluation** using both LLM-based and linguistic metrics show that text-based personas are expressed better than the corresponding image representations.
4. We also conduct **stratified analysis** to show the stability of our results regardless of the type of personas, evaluators, and questions.

2 Related Work

Persona Evaluation Prior work has established several frameworks for evaluating language models' role-playing capabilities. Wang et al. (2024a) introduced RoleBench, an evaluation benchmark with QA pairs based on character profiles. Wang et al. (2024b) developed InCharacter, assessing role-playing fidelity through psychological scales in an interview format. Tu et al. (2024) created CharacterEval, a Chinese benchmark derived from novels and scripts with multi-interaction dialogues, while Shen et al. (2023) established RoleEval, a bilingual benchmark with multiple-choice questions testing persona knowledge and reasoning. Samuel et al. (2024) introduced PersonaGym, a dynamic evaluation framework for automated assessment of persona adherence across diverse interactions. Our work further extends the literature by performing the first systematic evaluation to understand the influence of the persona modality.

Multimodal Personas Recent work has explored integrating visual elements into LLM persona systems. Ahn et al. (2023) introduced MPCHAT, demonstrating that incorporating visual episodic memories alongside text improves dialogue consistency and persona grounding. Sun et al. (2024) investigated how visual personas influence LLMs' behavior in negotiation contexts, showing models can adapt their responses based on perceived visual personality traits. Dai et al. (2025) developed MMRole, a framework for training and evaluating multimodal role-playing agents. While these works establish the potential of visual personas and others extensively evaluate textual personas (Li et al., 2016; Xiao et al., 2024; Samuel et al., 2024), there has been no systematic comparison of how different modalities of persona representation affect model performance. Our work addresses this gap by directly evaluating text, visual, and hybrid approaches across a range of persona-based tasks.

Modality Alignment Language models demonstrate strong in-context learning capabilities in unimodal textual settings (Shanahan et al., 2023; Salewski et al., 2023). However, extending these capabilities to multimodal inputs remains challenging. When visual information is introduced, models often struggle to transfer knowledge effectively from text to vision (and vice versa), resulting in noticeably weaker performance with visual in-context demonstrations compared to textual ones (Zhao et al., 2024a; Jiang et al., 2024). Such cross-modal gaps manifest in several ways: for instance, catastrophic forgetting of text-based instruction following can occur when models are finetuned on images (Zhang 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.

3 Influence of Modality in Persona Embodiment of Multimodal LLMs

The problem of embodying a persona can be defined as the task of generating responses consistent with a specified character, identity, or role (Samuel et al., 2024). 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 the persona p in different modalities, denoted as $\mathcal{R}(p)$, on multimodal LLMs. In particular, we consider two common modalities, text and image, and evaluate the LLM's performance on these equivalent representations. Additionally, we also consider combining visual and textual features of the personas. We describe these 4 different representations $\mathcal{R}(p)$ in more detail below.

3.1 Persona Modality Representations

- **Text ($\mathcal{T}$):** Textual descriptions of a persona correspond to a sequence of sentences characterizing the persona in natural language.
- **Image ($\mathcal{I}$):** A persona can also be depicted visually using an image of the person in a representative environment that characterizes the persona visually.
- **Assisted Image ($\mathcal{I}_A$):** Since certain features may be obscured in the image, textual attributes of the persona can also be included explicitly as text.
- **Descriptive Image ($\mathcal{I}_D$):** In this case, we include the textual attributes in the image itself using typography instead of in the text.

4 Modality-Parallel Persona Dataset

4.1 Personas

We introduce a novel dataset of personas $\mathcal{P} = \{p_i\}$, such that each persona p can be represented equivalently in four modalities $\mathcal{I}, \mathcal{T}, \mathcal{I}_A, \mathcal{I}_D$. To ensure effective representation across both text and image modalities, we construct personas based on key demographic attributes that are easily visualizable (Todorov et al., 2015). Specifically, each persona is defined by a unique combination of age, gender, occupation, and location. A persona can thus be written as:

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

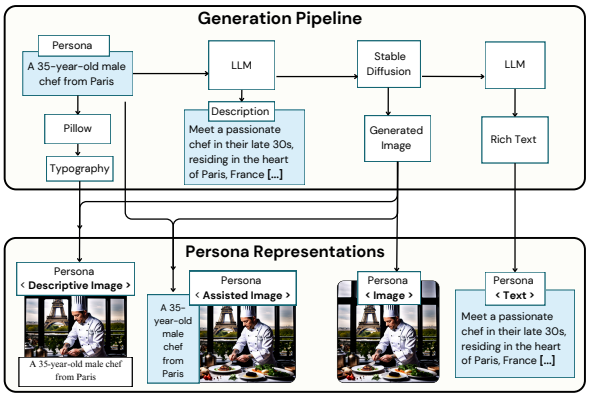


Figure 2: Our pipeline begins with curating a set of personas. Each persona receives a detailed text description, which is then fed into Stable Diffusion to generate $\mathcal{I}$. A separate model examines the image and generates an independent textual description, forming text persona $\mathcal{T}$. Pairing p with $\mathcal{I}$ produces an assisted image $\mathcal{I}_\mathcal{A}$, while combining a typographic representation of p with $\mathcal{I}$ produces a descriptive image $\mathcal{I}_\mathcal{D}$.

Table 1: **Persona Dataset Summary**

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

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.

To promote diversity, we systematically categorize these attributes into distinct groups and uniformly sample from each category. 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 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.

4.2 Equivalent Modality Representations

From above, we have a diverse set of textual persona descriptions as described by the four demographic attributes. Next, we construct a modality-parallel dataset, we require that each persona p can be equivalently depicted in 4 representations $\mathcal{R}(p)$: image $\mathcal{I}(p)$, text $\mathcal{T}(p)$, assisted image $\mathcal{I}_A(p)$, and descriptive image $\mathcal{I}_D(p)$. Figure 2 illustrates the step-by-step procedure to obtain these modality representations for a persona description $\mathcal{P}$.

1. We first convert the persona description made from the four attributes into a more detailed visual description using an LLM [2] with the prompt provided in Appendix A.1.
2. Next, we use a text-to-image generative model, particularly, Stable Diffusion XL [3] to generate a 768×768 px image conditioned on the more complete description of the persona found above. 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 the image $\mathcal{I}$.
3. Since the generated image can contain extra information due to underspecified textual prompts, we prompt the LLM one more time to generate a complete description of the persona as described in the image using a detailed prompt as given in Appendix . Thus, we obtain the text $\mathcal{T}$.

These steps enable us to convert a dataset of persona descriptions $\{p\} \rightarrow \{(\mathcal{T}(p), \mathcal{I}(p))\}$ such that $\mathcal{T}(p) \leftrightarrow \mathcal{I}(p)$ are equivalent to each other. One can now also obtain the assisted and descriptive image representations of the persona by pairing the image $\mathcal{I}$ with the text persona p for the assisted image $\mathcal{I}_A$, and by rendering [4] p as black text at the bottom of the image on a white background using Arial font at size 20 for $\mathcal{I}_D$. At each step, we ensure through manual verification that all the *important attributes of the persona are present* in each persona representation.

4.3 Question Generation

To evaluate how well a model embodies a given persona, we create a set of 60 questions that specifically probe for a given attribute either directly or in naturalistic scenarios. In particular, we create 10 questions per attribute for the two sets. Gender was excluded from our evaluation question set due to methodological constraints. While age and location can be objectively probed through factual knowledge, gender assessment would inevitably rely on stereotypes or normative expectations. Moreover, there may be a high possibility of refusal from the LLMs due to their safety training. Thus, we obtain two question sets Q^D and Q^S for L: location, O: occupation, and A: age.

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

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

4.3.1 Direct Testing

Questions were designed to probe specific knowledge across age, location, and occupation categories while enabling objective evaluation. For example, location questions assess knowledge of local customs and landmarks, while occupation questions may test domain expertise. For example, for age, we ask "*what life experiences do you consider most defining for your generation?*" while for location, we have "*what is the most visited tourist attraction in your area?*". We provide the complete list in Table 7 in Appendix.

4.3.2 Situational Testing

Scenarios accomplish similar knowledge evaluation but through naturalistic situations, requiring personas to implicitly demonstrate both knowledge and behavioral consistency. For example, for age scenarios, 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. A complete list is provided in Table 8 in Appendix.

4.4 Evaluation

For each persona $p \in \mathcal{P}$ and question $q \in \mathcal{Q}^D \cup \mathcal{Q}^S$, we find the response answer $a \leftarrow \mathcal{M}(\mathcal{R}(p), q)$ from a multimodal LLM $\mathcal{M}$, where $\mathcal{R}(p)$ denotes a modality representation of the persona $\mathcal{P}$. Thus, we obtain $(q, \mathcal{R}(p)) \rightarrow_{\mathcal{M}} (p, q, a)$. We now evaluate the quality of the response a based on the question asked q and the persona description p.

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

4.4.1 LLM-based evaluation

Following Samuel et al. (2024), we employ an LLM-based evaluator to judge the quality of the responses based on different metrics defined in the prompt. In particular, we prompt the LLM judge $\mathcal{J}$ with the question asked q, response a, and the persona description p on these metrics as described by the corresponding prompts.

For each evaluation criterion, $\mathcal{J}(p, q, a)$ outputs a score from a 5-point Likert scale based on the corresponding system prompt. For situational testing, we evaluate using `action justification`, `expected action`, `linguistic habits` while for direct testing, we use `persona consistency` and `linguistic habits`. Note that we combine the scores for linguistic habits across the two testing sets to find the average score.

4.4.2 Comparative Evaluation

We employ two comparative evaluation methods to assess relative performance across modalities, using evaluator $\mathcal{J}$ with the prompt:

```
You are given a persona description and multiple
responses to a prompt.
Persona Description: <p>
Prompt: <q>
Candidate Responses: <responses>
Choose the single response that best fits
the persona's style, values, and consistency.
Respond with 'Response X' where X is the number
of the chosen response.
```

Pairwise Comparison To compare responses across the text and image modalities, we first directly compare responses $a_{\mathcal{T}}$ and $a_{\mathcal{I}}$.

Swiss System Comparison To collectively evaluate all four modalities, we adopt the Swiss tournament system, which reduces the number of required comparisons compared to pairwise evaluation while maintaining ranking quality. Specifically, for $n = 4$, pairwise evaluation requires $\binom{4}{2} = 6$ comparisons, whereas the Swiss system reduces to 3 comparisons.

4.4.3 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:

- **Types:** $|\{r\}|$, unique token count.
- **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).
- **Measure of Textual Lexical Diversity (MTLD):** Following McCarthy and Jarvis (2010), calculates the mean length of text segments that maintain a type-token ratio (TTR) $> \tau = 0.72$.

5 Experimental Setup

Models. We evaluate the performance of 5 representative multimodal large language models: (1) GPT-4o [5], (2) GPT-4o mini [6], (3) Llama 3.2 11B [7], (4) Llama 3.2 90B [8], and (5) Pixtral 12B [9]. We select these models as they represent a diverse cross-section of three major model families and have varying parameter sizes.

Evaluators. We utilize two LLM evaluators, using GPT-4o and Gemini 2.0 Flash [10], with deterministic sampling with zero temperature. 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.

6 Results

6.1 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; numbers represent the mean between our two evaluator models. 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. Since these criteria are oriented more toward the actions taken by the persona instead of

[5] GPT-4o [6] GPT-4o-mini [7] Llama 3.2 11B [8] Llama 3.2 90B [9] [10] Gemini 2.0 Flash

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

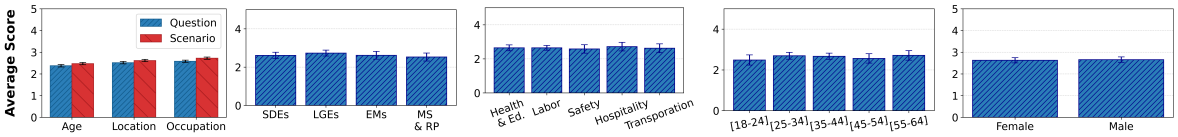
Table 2: **LLM-based evaluation [1-5] of responses under different persona modality representations.**

the generated language, we believe the models are trained to specifically attend to embedded text to generate directed responses based on the text embedded within the image. We also 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 6.3.

Next, we can also note that GPT-4o shows the highest average alignment scores across 4 criteria in each modality representation. This shows that GPT-4o is the most capable model of embodying these personas for the curated questions. 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 more direct pairwise comparison, we find that image-based personas are almost never chosen (selecting text up to 99%). This further strengthens our claims from above regarding the lack of capabilities of current multimodal models to embody visual personas.

Linguistic Evaluation Next, we compare the linguistic diversity of the responses generated through different persona modalities. Table 4 shows the mean and standard deviation in the three metrics



(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)**

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

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, which also show significantly more variation (at least ~ 2 more). We note that the Llama models are highly selective and show extremely high linguistic diversity in text modality than other modalities, as Llama-3.2-90B generates up to 150 more types when prompted with a textual persona as compared to an image and up to 2 times root token-type ratio.

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 to judge how well the response is aligned with the persona for the given question. Table 5 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, showing the highest alignment for text followed by the descriptive image modality, while assisted image and image perform similarly. Human scores show a strong Pearson correlation with LLM

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

evaluations – 0.99 for linguistic habits and 0.78 for action justification. This indicates that humans are particularly attentive to persona-consistent language during evaluation. We defer other survey details to Appendix C.

6.2 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 LLM judge regardless of the persona modality for different categories. Except for a slight preference for "scenario" over direct "questions", we do not observe any major effect of these factors, confirming that the LLM-assigned scores are not confounded on factors such as question and persona attributes. This further emphasizes the role played by persona modality in Table 2.

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

6.3 Investigating assisted image vs. image

Finally, we investigate the counterintuitive finding that the text-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 for 15 questions.

Figure 4 shows 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

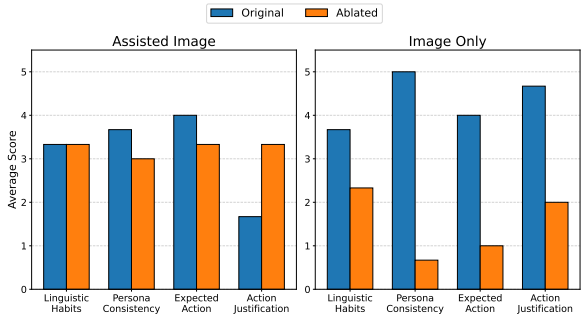


Figure 4: Effect of replacing persona images with random noise. The Image condition (right) shows dramatic performance decreases, while the Assisted Image condition (left) is less affected.

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.

7 Discussion & Conclusion

In this work, we study the influence of the modality of a persona in multimodal LLMs. We first create a novel modality-parallel dataset with equivalent representations across 4 different modality representations: image, text, assisted image, and descriptive image. Using a set of manually curated hard questions about the persona, we find 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. 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 also believe our modality-parallel dataset lays the foundation for future advancements in visual persona understanding in multimodal LLMs. Future works can also comprehensively analyze different configurations of combining image and text modalities (*e.g.*, different fonts, sizes, placements, etc. in descriptive image) to enhance the embodiment of multimodal personas (Dai et al., 2025) and test the model's robustness. 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 worlds" within an image or to understand any of them at all.

Limitations

A limitation of our work is that we only deal with 40 personas. However, due to a lack of any persona dataset with equivalent representations in different modalities, we see this as our contribution and leave it for future works to expand the scale of the study. Furthermore, we specifically increase the diversity of these personas across 4 well-grounded categories, focusing on the quality of our dataset. As the field of persona alignment in LLMs is still quite nascent, we believe quality becomes more important than quantity. 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.

Ethical Considerations

We intend 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

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.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024. 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*.

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.

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.

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.

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.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024. Wings: Learning multimodal llms without text-only forgetting. *Preprint*, arXiv:2406.03496.

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.

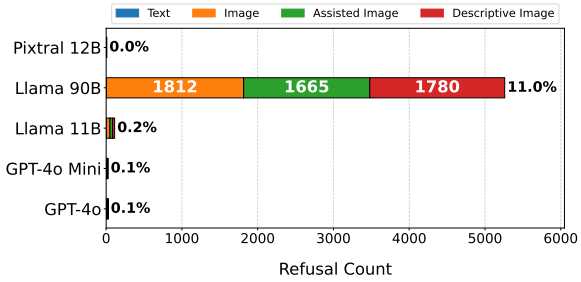


Figure 5: The rate and number of refusals in response to persona prompts. Llama 3.2 90B shows a strong aversion to multimodal persona prompts, while other models rarely refuse.

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 5). 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 Human survey design

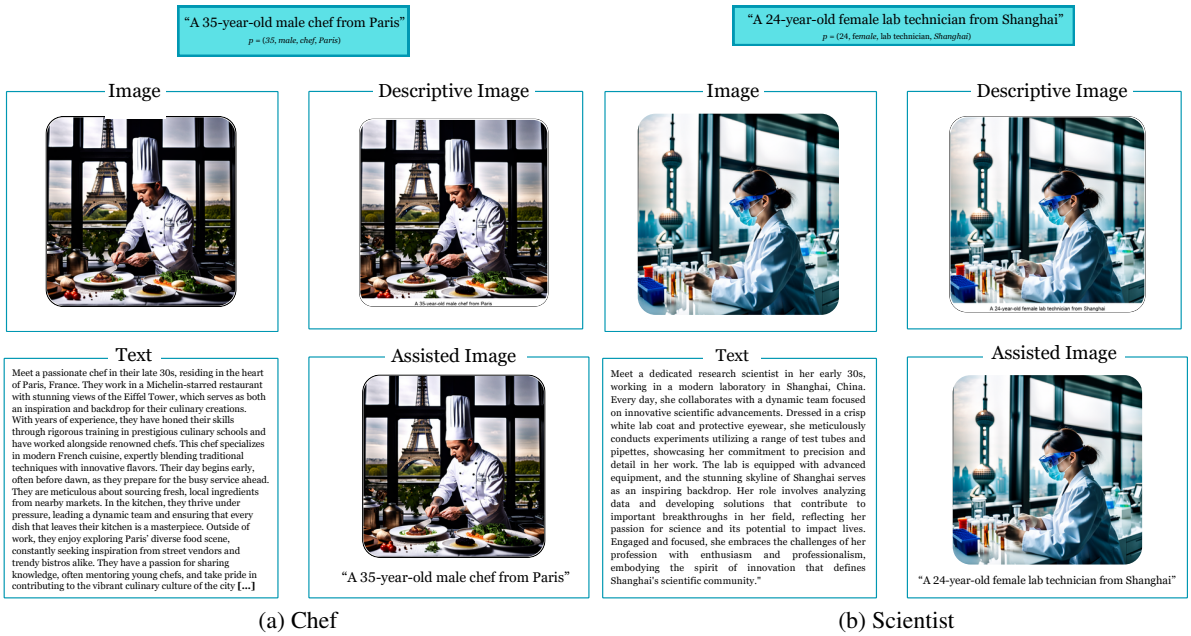
Figure 7 demonstrates our survey design that we conduct on 8 independent annotators to evaluate the quality of LLM evaluators. In particular, we first show the instructions to evaluate the responses

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

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

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

for a prompt and a persona, followed by 10 such questions.



(a) Chef (b) Scientist

Figure 6: **Persona examples:** Representation of a chef and lab technician persona across four modalities: Text, Image-only, Assisted Image, and Descriptive Image.

Table 7: Direct testing question list

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 8: Scenarios for situational testing

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 ...

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.

GPT-4o				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
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				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
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				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
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				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
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				
Modality	**Linguistic Habits**	**Persona Consistency**	**Expected Action**	**Action Justification**
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.

You are being asked to be a volunteer in a research study. The purpose of this study is to **evaluate the alignment of given personas in multimodal large language models**. The **survey** will take approximately **10 minutes** to complete.

In the following 10 pages, you will be given a persona, a prompt, and a response. You have to evaluate ***how closely the response aligns with the expected response of the persona to the given prompt***. An ideal response should:

- Adhere to the persona's characteristically unique syntax, tone, and lingo, ensuring that these elements are consistently utilized throughout the persona's dialogue.
- Be faithful and specific to the given persona and avoid generic characteristics
- Reflect the persona's internal reasoning and motivations for the action, aligning with its established characteristics and context.
- Take actions that are logically expected of the persona in the setting of the question.

→

(a) Instruction

Persona
A 50-year-old male farmer from Sao Paulo

Prompt
During a job interview, you're asked about your generation's unique approach to work-life balance and career expectations. You describe ...

Response
Being a 50-year-old farmer from São Paulo, I've come to deeply value a strong sense of community and the rhythm of nature. My generation has a unique approach to work-life balance in that we recognize the importance of hard work but also the necessity of taking time to recharge and connect with loved ones.

Unlike the urban hustle, rural life has taught us to respect the seasons and understand that there's a time for every activity. We don't necessarily strive for the same corporate ladder growth that younger generations might; instead, we aim for stability and sustainability. We place a high value on relationships, family, and a sense of purpose that comes from seeing the fruits of our labor every day.

When it comes to career expectations, we seek fulfillment in the quality of our work and its impact on our community. For me, that means knowing that what I grow not only feeds my family but also supports my neighbors and local economy. It's less about rapid advancements and more about continuous, steady progress and contributing to something greater than ourselves.

- ○ Not aligned with the persona at all
- ○ Partially aligned with the persona
- ○ Mostly aligned with the persona
- ○ Fully aligned with the persona

(b) Question

Figure 7: Human survey design