DREAM: A Multicultural Multimodal Dataset Linking Dialogues and Realistic Image Sequences

Juan Mallo de la Calle, Marcos Estecha-Garitagoitia, Ricardo Córdoba, Luis Fernando D'Haro

Speech Technology and Machine Learning Group, Universidad Politécnica de Madrid (UPM)
ETSI de Telecomunicación, Information Processing and Telecommunication Center, Madrid, Spain
juan.mallo@alumnos.upm.es, {mestecha, ricardo.cordoba, luisfernando}@upm.es

Abstract

An ongoing challenge in multimodal language research is creating and interpreting dialogues that preserve visual and cultural consistency across turns. We introduce **DREAM** (Dialogue to REAlistic Multicultural Image Sequences), a multicultural multimodal resource that ties dialogues grounded in explicit persona profiles to photorealistic, storyboard-like image sequences. Each of the 1,000 dialogues includes two rich persona profiles (structured traits plus descriptive language), two matching photorealistic portraits, and a collection of scene-level images depicting key dialogue moments. The pipeline integrates profile augmentation, culturally-sensitive prompt engineering, and turn selection to craft cohesive visual narratives, promoting character consistency across images. This is accomplished through a controlled generation process employing large language and image models. Beyond dialogue grounding, DREAM supports *appearance-based demographic perception* and *culture-aware rendering*: models can be evaluated on their ability to (i) perceive age, gender presentation, and broad ethnicity appearance clusters from profile portraits, and (ii) maintain these characteristics in dialogue scenes. We provide a unified JSON format integrating profiles, dialogue text, and visual turns, facilitating research on visually anchored dialogue understanding, consistency, and generation. A dual evaluation protocol combines human judgments (realism, coherence, consistency, and demographic perception) with automated portrait analysis via GPT-5. Ethical considerations, limitations, and recommended applications are discussed.

Keywords: multimodal dataset, multicultural, persona profiling, visual grounding, appearance-based demographic perception, bias and robustness, evaluation, synthetic data

1. Introduction

The demand is increasing for resources that ensure both *visual* and *cultural* continuity in multimodal dialogue systems. Although existing corpora often focus on textual exchanges, they frequently overlook the integration of dialogues with images that (i) maintain character identity over time and (ii) systematically represent diverse demographic cues (e.g., age groups, gender presentation, appearance-based ethnicity clusters, and context such as nationality or residence). This omission limits progress toward visually grounded dialogue understanding, consistency-aware content generation, and systematic analysis of demographic representation patterns.

Introducing DREAM, a multicultural multimodal resource that connects dialogues rooted in persona profiles to photorealistic, storyboard-like image sequences. Each of the 1,000 dialogues comes with two detailed personas (including structured traits and narratives), two matching profile portraits, and a series of scene-level images aligned with visually important dialogue turns. Our process includes persona expansion, culturally-informed prompt engineering with safeguards, and a selection method for visual turns that capture key actions, settings, and emotions while maintaining continuity.

Beyond visual coherence, DREAM facilitates *appearance-aware* training and evaluation: models can infer broad, perceptual demographic cues from portraits and preserve them in scene generation. This supports both research (representation, robustness, bias audits) and opt-in applications (e.g., adaptive avatars), explicitly avoiding identity recognition. We describe these cues as approximate visual signals and emphasize minimal-data and privacy-preserving deployment.

Diverging from earlier studies on persona-based dialogues and vision-language datasets, DREAM introduces a *unified JSON schema* that simultaneously encompasses (i) both structured and narrative persona information, (ii) dialogue text with distinct turn indices, and (iii) profile images along with visual turns. This framework facilitates research focused on (a) the impact of persona grounding on visual coherence, (b) approaches for evaluating character consistency across sequences, and (c) the dynamics between demographic diversity and model robustness and bias.

We also propose a dual evaluation protocol: (1) human evaluations of realism, coherence between dialogue and images, character consistency across dialogue turns, and demographic perception in profile portraits; and (2) automated demographic estimation from profile images using an advanced

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9266–9281
11–16 May 2026.

vision–language model (GPT-5). Together, these evaluations help audit visual quality and consistency, while offering insights for fairness and bias-focused studies.

Contributions.

- A **multicultural multimodal resource** consisting of 1,000 persona-grounded dialogues, each matched with sequences of photorealistic images and two consistent profile portraits per dialogue.
- A **unified JSON schema** that integrates persona traits and narratives with the interleaved dialogue text, including turn indices and inserted image IDs; this format is *tooling-friendly* and allows for direct loading for training, inference, and UI rendering.
- A **controlled, reproducible generation pipeline** that combines persona augmentation, culturally-aware prompt creation, and visual selection at the turn level to maintain identity consistency across scenes; this pipeline is *model-agnostic* (LLM and text-to-image (T2I) components can be interchanged) and *batchable* for large-scale executions.
- A **dual evaluation protocol** encompassing both human evaluations (realism, coherence, consistency, demographic perception) and automated demographic analysis of profile portraits.
- A **diversity-first curation** spanning multiple age groups, gender identities, and appearance-based ethnicity clusters, enabling research on bias, representation, and robustness.

2. Related Work

2.1. Persona-grounded Dialogue Corpora

Zhang et al. (2018) presented *PersonaChat*, a dataset for conversational chit-chat created through crowd-sourcing. In this corpus, each participant's dialogue is influenced by a short textual profile, typically consisting of 4–5 sentences, intended to encourage engaging and consistent conversations. This dataset was pivotal in making persona conditioning popular in open-domain dialogues and is accessible through the ParlAI framework Miller et al. (2017). We utilize a maintained mirror of this dataset on Hugging Face Awsaf49 (2018), specifically using the revised training split (`train_both_revised.txt`). Despite its impact, *PersonaChat* profiles are short textual descriptions without structured demographic information or visual content, which limits controlled multicultural representation (e.g., balanced age groups, gender identity categories, appearance-based ethnicity clusters, or nationality/residence context) and consistency of visual identity across images.

2.2. Comparative and Profile-driven Dialogues

ComperDial specializes in dialogues grounded in commonsense personas, highlighting comparative reasoning and behavior influenced by agent profiles (Wakaki et al., 2024). The dataset is available publicly on Hugging Face (Sony AI, 2024). Unlike *PersonaChat*, *ComperDial* presents scenarios where speakers actively use profile attributes to contrast preferences or options over multiple exchanges. However, as a text-centric resource, *ComperDial* excludes portrait images or scene visuals and does not aim for explicitly curated demographic coverage across dimensions such as age, gender identity, appearance-based ethnicity, or nationality/residence. This limits its applicability for systematic explorations of multicultural representation or multimodal identity consistency.

2.3. Multimodal Dialogue and Visual Grounding

A wider body of work integrates dialogue with visual media such as images or videos, aligning dialogue exchanges with individual images, captions, or grounded visual contexts through vision–language learning paradigms (Das et al., 2017; Shuster et al., 2018; de Vries et al., 2017; Anderson et al., 2018). Common challenges in identity-focused studies include: (i) the absence of explicit persona representations linked to each speaker, (ii) inadequate or nonexistent handling of *sequence-level* visual grounding throughout multiple dialogue turns, and (iii) lack of mechanisms to ensure visual consistency of characters across different scenes. Consequently, while these resources successfully anchor dialogues in visual contexts, they are less suited for evaluating long-range character continuity or storyboard-style conversations with persistent identities.

2.4. Cultural Representation, Fairness, and Synthetic Demographics

Previous discussions within the community have highlighted the critical role of representation and fairness in datasets involving demographic elements. A frequent challenge is distinguishing between *self-identified attributes* and *perception based* on appearance: many evaluation tasks require annotators (or models) to assess visible signals (e.g., age range, gender presentation, and

broad appearance clusters) instead of identity itself. Synthetic pipelines, where personas and images are generated rather than collected from real individuals, enable appearance-based perception studies without relying on real personal data or photographs. Recent works have also explored LLM-driven prompt design and synthetic multimodal dataset construction (Brade et al., 2023; Huang and Huang, 2024). However, such pipelines must still avoid stereotypical representations and ensure age-appropriate content. Resources that explicitly document these design choices support auditing, robustness analysis, and bias evaluation in multimodal dialogue systems.

3. Dataset Construction

Figure 1 illustrates the pipeline. Initially, we choose dialogue sources, followed by expanding personas into detailed structured-narrative bundles and creating a unique portrait for each speaker. Subsequently, we extract *visual turns* from the dialogue text and produce scene images, ensuring identity is maintained through portrait references. Lastly, all elements are compiled into a cohesive JSON format, ready to be evaluated.

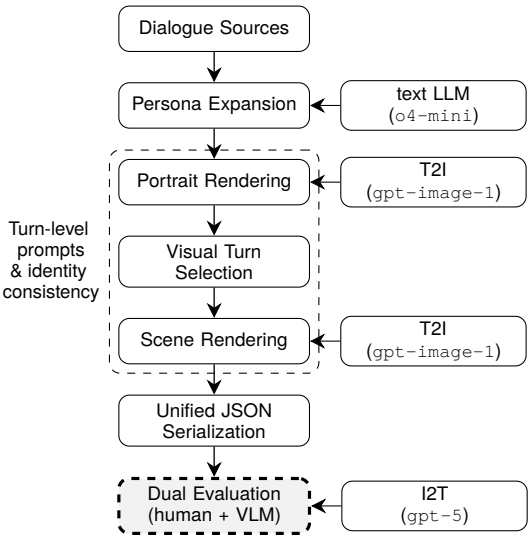


Figure 1: DREAM pipeline: dialogue selection, persona expansion, portrait rendering, visual turn selection, scene rendering, unified JSON, and evaluation.

3.1. Dialogue Sources

We utilize two public persona-grounded resources: *PersonaChat* (Zhang et al., 2018) and *CamperDial* (Wakaki et al., 2024). In DREAM, our dataset consists of **1,000 dialogues**: **900** are selected from `PersonaChat` (file `train_both_revised.txt`, with a total availability of **8,939**) and **100** from `CamperDial` (the entire public set available at the time of collection). The sources are standardized (Unicode and punctuation) and combined into a unified schema featuring clear A/B speaker labels and turn indices, without any alterations to the content.[1]

3.2. Persona Profiles (Profile Expansion)

Each dialogue has two speakers, A and B. For both individuals, we create:

- `profile_struct`: `age`, `gender_identity`, `sexual_orientation`, `ethnicity` (10-category appearance-based grouping), `nationality`, `residence_country`.
- `profile_narrative`: five succinct lists detailing: *personal_data*, *visual_appearance*, *environment*, *personality_attitudes*, *other_details*.
- `profile_prompt`: a self-sufficient, shoulders-up photorealistic prompt featuring stable visual characteristics and a subtly blurred environmental suggestion.

Evidence-first, otherwise controlled fallbacks. If a structured attribute is explicitly supported by the speaker's persona sentences or dialogue (e.g., "I am 24", "I grew up in X"), we set it accordingly. Otherwise, we assign it from predefined fallback distributions to avoid unconstrained demographic hallucination and to ensure broad coverage across the ten appearance clusters, age groups, and gender identities. Refer to Ethics & Limitations (Sec. 8) for the reasoning behind this taxonomy and the precautions in place.

Nationality vs. residence. We distinguish `nationality` (self-identified legal/citizenship cues) from `residence_country` (current living context). Mentions of work, studies, leagues, or daily-life context are treated as residence evidence only; nationality is changed only when explicitly self-declared. This enables controlled diaspora-like cases (`nationality` $\neq$ `residence_country`) without letting weak cues arbitrarily shift nationality.

Narrative augmentation. Each persona is expanded into five narrative sections (`personal_data`, `visual_appearance`, `environment`, `personality_attitudes`, `other_details`), each with at least five concrete,

[1] All personas and images in DREAM are synthetic; the dialogue text is either synthetic or publicly released for research. See Ethics & Limitations.

renderable sentences. Compared to `profile_struct`, this layer provides scene-useful hooks (rooms, props, routines) while keeping identity attributes stable and analyzable.

Contextual attire rules (limited scope). To improve visual plausibility in contexts with legally enforced public-dress requirements, we apply a small set of explicit prompt rules (e.g., hijab/abaya or headscarf) only when `residence_country` indicates such constraints (e.g., Iran, Afghanistan, Saudi Arabia). This mechanism is intentionally narrow in DREAM and is designed to be extensible to additional context-dependent realism rules in future versions.

Validation and safeguards. Rule-based checks remove contradictions (e.g., incompatible age/roles) and keep cross-national residence plausible by introducing it explicitly via `residence_country`. Portrait prompts prioritize stable identity cues (face, hair, persistent accessories), while scene prompts control dynamic elements (pose, temporary outfit, props, lighting, location), enabling storyboard-like variation without identity drift. We also enforce near-uniform coverage across the 10 appearance clusters to prevent collapse into a small set of dominant groups and to enable per-group analysis. See Sec. 8 for taxonomy rationale, bias sources, and attire limitations.

3.3. Profile Portrait Rendering

We generate one 1024×1024 portrait per speaker with **gpt-image-1**, using the `profile_prompt`. Portraits fix identity and background cues later reused to preserve consistency in dialogue scenes. Four representative persona portraits are shown in Figure 2, which serve as identity anchors for subsequent scenes.

3.4. Visual Turn Selection

From dialogue turns (`text_i`), we identify visually significant moments, combine consecutive lines to form a single moment, classify them under `scene_type` ∈ {`shared`, `memory`, `imagined`, `cutaway`, `montage`}, and denote `speaker_focus` ∈ {`A`, `B`, `both`}. An early `shared` establishing scene is required (an initial anchor shot that places both speakers in the ongoing setting, or split-screen if remote). Events that have already happened are labeled as `memory`, whereas unreal or speculative content is labeled as `imagined`. `Cutaway` encompasses inserts like objects or locations, and `montage` denotes a compact depiction of progress or time passage. Visual examples are reported in Appendix B.

Figure 2: Examples of persona portraits (1024×1024) used to anchor identity. Synthetic images; selection shown for illustration only.

3.5. Scene Rendering and Identity Control

For each visual turn, the generation prompt concatenates: (1) the portrait prompt(s) indicated by `speaker_focus` (used *only* to preserve identity), and (2) the scene prompt. All images are 1024×1024 with neutral/soft lighting unless the scene implies otherwise. A complete dialogue with scenes is reported in Appendix E.

3.6. Models and Batch Automation

We utilize: **o4-mini** to enhance profile text and extract visual turns; **gpt-image-1** for creating photorealistic portraits and scenes; and **GPT-5** as a vision—language model to automatically perceive demographics based on appearance in portraits (Section 5). Processing is conducted in dialogue-level batches to maintain consistency and efficiency.

4. Resource Overview

4.1. Size and Turn/Scene Statistics

We release **1,000 dialogues**, with **6,950** dialogue images and **2,000** profile portraits in total. Table 1 summarizes the number of turns and rendered scene images per dialogue.

Metric	Mean	Med.	Min	Max
Turns / dialogue	14.6	14	12	24
Images / dialogue (scene)	6.95	7	4	13

Table 1: Per-dialogue statistics.

4.2. Demographic Distributions of Personas

Tables below report the target demographic distributions used for controlled coverage. We use ten coarse, appearance-based ethnicity clusters as an *operational label space* for perception and consistency analysis (not self-identification categories); clusters are designed to be broad enough for annotators to apply consistently and to support per-group robustness auditing.

Group	Count	%
South Asian	183	9.15
East Asian	209	10.45
Southeast Asian	168	8.40
Sub-Saharan African	226	11.30
North African & Middle Eastern	230	11.50
European (Northern & Eastern)	192	9.60
European (Southern / Mediterranean)	206	10.30
North American	213	10.65
Central & South American	191	9.55
Oceanian / Pacific Islander	182	9.10

Table 2: Ethnicity clusters (appearance-based).

Group	Count	%
<18	134	6.70
18–29	628	31.40
30–39	451	22.55
40–49	316	15.80
50–59	247	12.35
60+	224	11.20

Table 3: Age group distribution (mean age 35.41).

Group	Count	%
Male	917	45.85
Female	977	48.85
Non-binary	46	2.30
Transgender male	27	1.35
Transgender female	33	1.65

Table 4: Gender identity distribution.

Other traits. The categories of sexual orientation (7), nationalities (91), and residence countries (87) are documented in Ethics & Limitations (Section 8), including safeguards for minors.

4.3. Standardized JSON Format

Each dialogue JSON file contains (i) two personas (`profile_struct`, `profile_narrative`, `profile_prompt`) and (ii) the dialogue sequence, where `image_id` entries are inserted at the appropriate positions between turns. Portrait prompts do not appear inside scene prompts; they are combined only at rendering time based on `speaker_focus`. A complete illustrative JSON example is provided in the Appendix A.1

4.4. Visual-turn Specs (Separate Files)

Prior to rendering, we generate a per-dialogue specification file listing the `visual_turns`. Each entry specifies the `image_id`, the covered `dialogue_indices` (e.g., `"text_7"` or `["text_7","text_8"]`), the `speaker_focus` (`A / B / both`), the `scene_type` (`shared`, `memory`, `imagined`, `cutaway`, `montage`), and a cinematic scene `prompt`. A full example is provided in the Appendix A.2. This two-step approach (specs → final JSON) keeps prompts reusable and facilitates reproducible rendering.

5. Evaluation

We evaluate DREAM through (i) a **human study** conducted via a web interface (implemented with Gradio) and (ii) an **automated portrait analysis** employing a GPT-family vision–language model ("GPT-5"). Details on eligibility, consent, and privacy protections are provided in Ethics & Limitations (Section 8).

5.1. Human Evaluation Protocol

Interface and instructions. Annotators used an English Gradio-based interface (Figure 3) and had to be at least 18 years old to participate. During registration, annotators provided a *non-identifying* nickname and coarse demographics for aggregate analysis: self-reported *gender identity* (Man, Woman, Other, Prefer not to say), *age group* (18–29, 30–39, 40–49, 50+), and *nationality*. We did not collect contact information or any free-text fields that could contain personally identifying information (PII). On each dialogue page, annotators viewed two profile portraits and a series of dialogue images interleaved with text. Instructions emphasized that all labels are *appearance-based perception categories* intended for evaluating synthetic images and should not be interpreted as identity assertions.

Tasks. (1) Profile portraits (A & B). Annotators estimated *appearance-based* attributes such as (i) ethnicity (allowing up to 3 selections from the 10 clusters used in DREAM), (ii) age group, and (iii) gender (Man, Woman, or Other/Can't tell).
(2) Dialogue images. For each generated image, annotators rated the (i) realism, (ii) coherence with the local dialogue turn(s), and (iii) character consistency throughout the sequence on a scale of 0 to 10 (with 0.5 increments); optional free-text comments were encouraged.

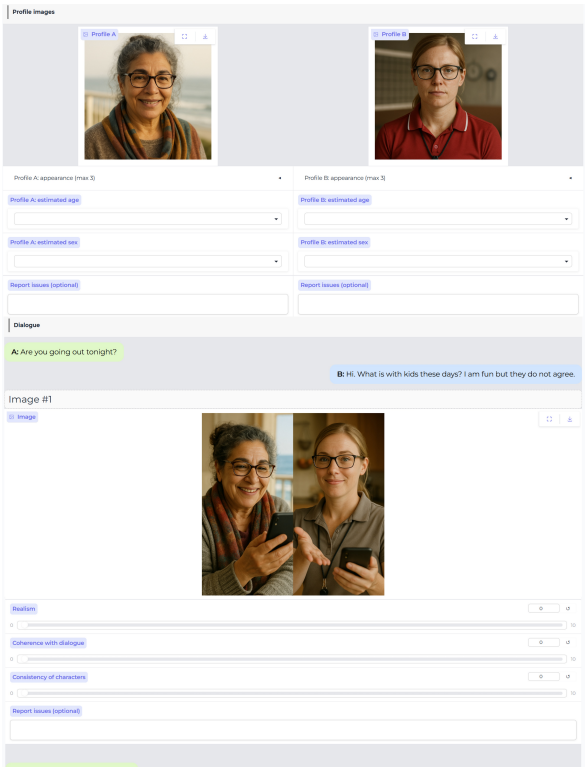


Figure 3: Human evaluation interface: profile portraits and demographic perception panel, interleaved dialogue image view, and rating sliders with optional issue reporting.

Participants and coverage. We received evaluations from **77** unique volunteers representing **9** different nationalities. Our evaluation covered a random **25%** subset of the corpus, consisting of **250** unique dialogues. Among these, **100** dialogues were independently assessed by three annotators, with two of these dialogues also reviewed by a fourth annotator. This resulted in a total of **452** dialogue-level evaluation tasks and **3,076** individual image ratings. Table 5 summarizes these numbers.

Quantity	Count
Unique annotators	77
Annotator nationalities	9
Unique dialogues evaluated	250
Dialogues with triple ratings	100
Total dialogue-level jobs	452
Total image ratings (all categories)	3,076

Table 5: Human evaluation coverage.

Scoring and label mapping. **Ethnicity.** When selecting up to three clusters, fractional credit is given as $1/1$, $1/2$, or $1/3$ depending on whether the correct cluster is among 1, 2, or 3 chosen clusters, respectively.[2] **Age.** We assess this against seven categories: 4—12, 13-–17, 18-–29, 30-–39, 40-–49, 50—59, and 60+. **Gender.** Our dataset contains gender *identity* categories (including transgender and non-binary). However, the human task is explicitly *appearance-based* (what is visually perceivable), not self-identification. For scoring consistency across portrait perception, we therefore map *transgender male* $\rightarrow$ Man and *transgender female* $\rightarrow$ Woman, and we treat *non-binary* as correct only when annotators select Other/Can't tell, reflecting that non-binary identity may not correspond to a reliable visual marker.

Results: dialogue images. Across **3,076** image ratings, the mean scores are: **realism** = 7.5, **coherence** = 8.3, **character consistency** = 7.5 (0–10 scale; 0.5 increments), as summarized in Table 6.

Metric	Realism	Coherence	Consistency
Mean	7.5	8.3	7.5

Table 6: Dialogue-image ratings (all images, all jobs).

Results: profile portraits. We present findings based on two human conditions: *All votes*, which involves every individual portrait assessment, and a *Majority (100 dialogues)* subset where 200 portraits were evaluated by three raters each and the consensus was determined by majority vote. The accuracy is assessed against the ground-truth profile fields (Sec. 3.2). Table 7 summarizes results.

Setting	Ethnicity	Age group	Gender
H (all votes)	70%	79%	97%
H (majority, 100)	89%	89%	97%

Table 7: Profile-portrait perception accuracy (human).

Qualitative findings. Feedback from free-text responses highlights common failure patterns in vision models. In total, annotators provided **576** comments. Of these, **224** were left unclassified due to insufficient semantic specificity, which is expected given the open-ended nature of human feedback. The remaining comments were grouped into eight recurrent categories.

Among identifiable issues, **anatomical errors** (especially hands and fingers) are the most frequent (**98** cases), followed by **text/iconography errors** (**68**) and **wardrobe or prop continuity**

[2] In the automated scenario (Sec. 5.2), the model is required to output a single ethnicity label, which we treat as a stricter condition.

problems (**66**). **Identity drift** appears in **47** cases and, while less frequent than low-level artifacts, is particularly disruptive because it directly breaks persona grounding and narrative continuity. Additional categories include **object semantics/world-knowledge errors** (**32**), **gaze or interaction misalignment** (**23**), and less frequent **scene-content mismatches** and **object physics issues** (**9** each).

Representative examples for each category are reported in Appendix D. Overall, these patterns align with known limitations of current T2I/VLM systems, particularly in *identity persistence across scenes*, *fine-grained anatomy*, and *interaction grounding*.

5.2. Automated Portrait Perception (GPT-5)

Protocol. We evaluate **2,000** profile portraits (A+B) using GPT-5. The prompt requests a strict JSON output with fields `ethnicity` (one of the 10 clusters), `gender` ({male,female,other}), and `age` (integer). The predicted `age` is mapped to the same seven bins as in the human study. Accuracy is computed with the same mapping rules for gender as above. Unlike the human protocol, ethnicity here is *single-label*.

Results and comparison. GPT-5 achieves **54%** (ethnicity), **63%** (age group), and **95%** (gender) accuracy (Table 8). Compared to humans (Table 7), single-label ethnicity is substantially harder for the model; gender is near-human; age grouping is below human but within a reasonable band for coarse perception.

Method	Ethnicity	Age group	Gender
GPT-5	54%	63%	95%

Table 8: Profile-portrait perception accuracy (automated).

Per-ethnicity analysis. We conduct an additional analysis of accuracy by examining ethnicity in a 10-category breakdown for two scenarios: (i) the human majority assessment based on 200 portraits rated by three people (Figure 4) and (ii) GPT-5's evaluation of all 2,000 portraits (Figure 5). We present confusion matrices that highlight confusions specific to each class.

5.3. Takeaways

- Dialogue images achieve strong **coherence** (8.3/10) with solid **realism** and **consistency** (7.5/10 each), validating the turn-level prompting and identity control.

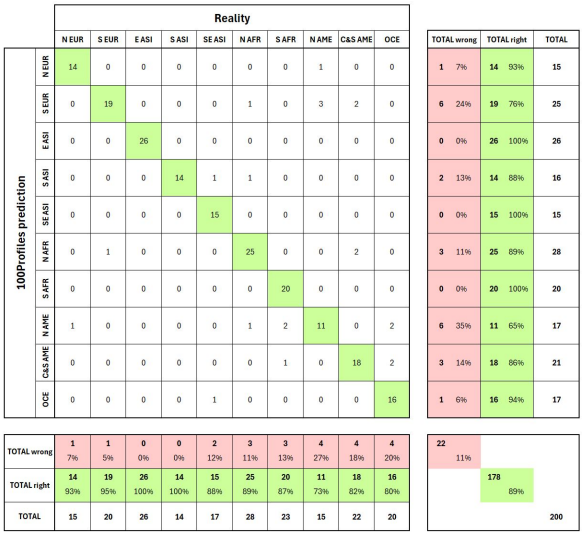


Figure 4: Ethnicity confusion matrix (human majority, 200 portraits).

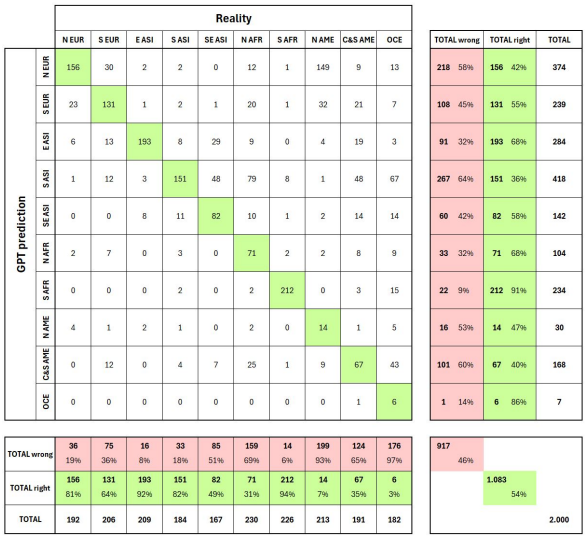


Figure 5: Ethnicity confusion matrix (GPT-5, 2,000 portraits).

- For profile portraits, **humans** outperform GPT-5 on **ethnicity** and **age group**, with near-ceiling **gender** accuracy in both cases.
- **Human (200 portraits, majority vote).** High per-class accuracy overall. The main weakness is *North American* (about 27% errors), most often confused with *European* clusters (Northern/Eastern and Southern/Mediterranean), suggesting close visual/photographic cues between North American and European appearances.
- **GPT–5 (2,000 portraits).** Performance varies widely across classes:
  - **Strong classes:** *Sub–Saharan African* (≈94%) and *East Asian* (≈92%). *South*

Asian (≈82%) and *European (Northern & Eastern)* (≈81%) are moderate.

- **Extreme underprediction of *Oceanian*.** The model outputs the *Oceanian/Pacific Islander* label only **7** times (6 correct) while there are **182** true instances; errors spread across many classes, with a notable drift to *South Asian*.
- **Collapse on *North American*.** Of **213** true cases, only **14** are predicted correctly; the dominant confusion is into *European (Northern & Eastern)* (**149** times).
- **Neighbor confusions.** Frequent mistakes follow appearance adjacency: *Southeast Asian* ↔ *South/East Asian*; *North African & Middle Eastern (MENA)* → *South Asian*; and *Central & South American* → *South Asian*/*European (Southern / Mediterranean)*.

5.4. Inter-Annotator Agreement

To address agreement for items rated by three annotators sampled from a larger volunteer pool (i.e., annotator identities vary across items), we report **Krippendorff's** α, which supports variable annotator assignment and multiple measurement levels. For profile portraits, we compute α on (i) *ethnicity* (nominal; using the primary selected cluster), (ii) *sex* (nominal), and (iii) *age group* (ordinal). For dialogue images, we compute α (interval) for the 0–10 ratings of realism, dialogue coherence, and character consistency. Agreement results are summarized in Table 9.

Target	Attribute	Krippendorff's α
Profile portraits	Ethnicity (nominal)	0.652
	Sex (nominal)	0.967
	Age group (ordinal)	0.920
Dialogue images	Realism (interval)	0.461
	Coherence (interval)	0.479
	Consistency (interval)	0.395

Table 9: Inter-annotator agreement measured with Krippendorff's α.

For ethnicity perception on the triple-rated portrait subset, annotators fully agreed in 55.5% of cases, had majority agreement in 40.0%, and total disagreement in 4.5%.

For dialogue image ratings, agreement is moderate. These results should be interpreted in light of the narrow distribution of scores: ratings are concentrated in the upper range of the scale (means between 7.5 and 8.3), and the variance for each dimension remains low (average variance < 1). Such limited dispersion reduces the expected disagreement term in Krippendorff's α, making the coefficient more sensitive to small rating differences between annotators.

6. Discussion and Use Cases

Why this resource matters. DREAM bridges persona-grounded dialogue and identity-consistent visual sequences within a single resource. By combining structured personas, profile portraits, and turn-level scene images, it enables controlled experimentation on multimodal dialogue grounding while maintaining character continuity across scenes. The unified JSON structure further facilitates reproducibility and direct integration into training and evaluation pipelines.

Benchmarking multimodal grounding and identity. The dataset facilitates controlled experiments on (i) *visual grounding of dialogue* at the turn level and (ii) *character persistence* across different scenes. Our human evaluation results (Sec. 5) demonstrate strong coherence and good, though not perfect, realism and consistency, indicating potential for improvement.

Fairness, bias, and representation. DREAM utilizes appearance-based demographic labels on synthetic portraits to facilitate *auditing* and *monitoring* of demographic shifts and confusions, such as those found in near-neighbor clusters. The analysis of confusions reveals systematic vulnerabilities, like those between North American and European, or Oceania and South Asian demographics, which should be addressed in future models through culture-aware rendering and perception.

Foundation for demographically aware agents. Apart from serving as a benchmark, DREAM can also be utilized as training data for *visual generation that preserves persona* and *assistants aware of demographic factors*, which ensure consistent visual identifiers throughout interactions. This encompasses personalized cultural mediation systems, such as adapting explanations, artifacts, and styles according to the user's inferred cultural cues, while avoiding the creation or retention of assertions regarding real identities.

7. Conclusion and Future Work

We presented DREAM, a multicultural multimodal resource that connects persona-based conversations with storyboard-style image sequences. It includes a two-fold evaluation method that involves both human assessments and automated portrait perception. This corpus aids research on dialogue

grounded in visuals, identity consistency, and demographic auditing based on appearance.

Future Directions. We outline several potential paths forward:

- **Temporal extension:** moving from static images toward short video sequences to better model motion, gaze, and temporal consistency.
- **Richer grounding:** extending dialogue diversity (languages, settings, and interaction types) and improving cultural and environmental specificity.
- **Baselines & reproducibility:** releasing reference baselines and generation scripts to facilitate comparison and community-driven extensions.
- **Real-world grounding:** exploring hybrid settings that combine synthetic generation with human-curated or real-world visual data.

DREAM is intended as a living benchmark: reproducible today, extensible tomorrow. Our aim is for it to inspire approaches that are not only more coherent and consistent, but also more equitable and transparent in their multimodal behavior.

8. Ethics and Limitations

Scope and intent. All dialogues, personas, and images in DREAM are synthetic or sourced from publicly available datasets and are intended solely for research purposes. This resource is designed for benchmarking and analysis purposes, such as studying representation, robustness, and culturally-aware rendering, and is not intended for use in identity verification or surveillance.

Privacy and consent. No actual portraits were gathered; the evaluators volunteered by providing a self-chosen, non-identifying nickname along with general demographics (such as age group, gender identity, and nationality) for the purpose of aggregate analysis. No contact information or persistent identifiers were retained.

Content safety. Generation followed provider safety constraints to avoid harmful or sexualized content, particularly involving minors. As expected for current text-to-image systems, some outputs contain visual artifacts (e.g., hands or text rendering), which are explicitly documented through human evaluation.

Bias and representational constraints. While we strive to ensure coverage across diverse appearance-based clusters and demographic groups, the dataset remains artificial and may not capture detailed cultural subtleties. The concept of "ethnicity" is employed in an operational sense as broad visual clusters for perception research, rather than as legal, genetic, or self-identification categories. We adhere to careful language in alignment with best practices for documenting sociotechnical attributes (Bender and Friedman, 2018). Synthetic narratives avoid stereotypes and limit culture-specific clothing to minimal, contextually appropriate indicators (further details provided below).

Diversity and Subjectivity in Evaluation. Human ratings naturally possess a degree of subjectivity, influenced by the cultural contexts of the annotators. Although we provide information on the size and diversity of our rater pool, expanding the demographic range of annotators would enhance the generalizability of the findings.

Closed Models and Reproducibility. The generation pipeline relies on closed LLM/VLM/T2I systems. To improve reproducibility, we release the final JSON resource together with generated portraits and scene images, and provide generation scripts and configuration details. Results should be interpreted as tied to the model versions available at the time of generation. Refer to Section 3 for further details.

8.1. Classification and Fallback Priors: Justification and Origins

Appearance-based ethnicity clusters. We categorize using ten broad, appearance-focused groups such as East Asian, Sub-Saharan African, Southern European/Mediterranean, Northern&Eastern European, MENA, North American, Central&South American, South Asian, Southeast Asian, and Oceanian/Pacific Islander. These groupings are explicitly *not* designed as legal or genetic categories, but instead serve as general visual classifications that align with UN M49 regional divisions. This alignment is intended to prevent arbitrary classifications and ensure that country lists remain consistent when sampling for nationality or residence (United Nations Statistics Division, 2024). Consequently, each group is associated with one or more M49 regions for nationality sampling and understanding diaspora dynamics.

Age bands. Our age categories include `<18`, `18–29`, `30–39`, `40–49`, `50–59`, and `60+`. These divisions are designed for research purposes and

serve to (i) separate minors; (ii) adhere to the common practice in social sciences of using decade-long intervals for adults; and (iii) ensure representation across both working-age and senior populations. Although international organizations might use different groupings, such as the UN's definition of "youth" as 15–24 or using 5/10-year intervals, decade-based categories are typical in numerous statistical reports and are suitable for studies on perception where age is a broad visual cue. It is important to note that these categories are used for modeling convenience and do not imply any inherent natural boundaries.

Gender identity categories. Our categories include *male*, *female*, *non-binary*, *transgender male*, and *transgender female*. These categories are aligned with widely accepted professional definitions: according to the APA Dictionary of Psychology, *gender identity* refers to an individual's internal sense of gender and acknowledges *nonbinary* identities. Similarly, the term *transgender* pertains to a gender identity that differs from the sex assigned at birth (American Psychological Association, 2023). We apply these five categories to encompass commonly recognized identities while maintaining a manageable label space for perception and evaluation tasks. **Safeguards for Minors:** For individuals younger than 13, we limit outputs in `profile_struct` to {male, female}, and we omit sexual orientation labels (detailed below) to ensure age-appropriate content.

Sexual orientation categories. The categories we acknowledge include *heterosexual*, *gay/lesbian*, *bisexual*, *asexual*, *pansexual*, as well as an *other/queer* category. U.S. public health directives recognize these classifications and their application in respectful terminology within federal communications. For instance, the CDC's "Preferred Terms" lists explicitly incorporate the terms pansexual, asexual, and non-binary (Centers for Disease Control and Prevention, 2024). Experts define sexual orientation as a consistent pattern of romantic or sexual attraction, with these terms widely utilized in both research and practical applications.

Priors (percentages) and balancing. Our priors are designed to prioritize *coverage* over representing the actual global distribution. Specifically: (i) We maintain roughly 10% for each of the ten appearance clusters to prevent collapse into a single mode and to allow for detailed cluster analysis, using the M49 mapping to ensure wide geographic representation (United Nations Statistics Division, 2024); (ii) For gender identity and sexual orientation, we set priors to (a) preserve heterosexual male/female as the majority (reflecting findings from large national surveys) while (b) ensuring sufficient sample sizes for minority groups (like non-binary, transgender, asexual, and pansexual individuals) for thorough analysis. As empirical benchmarks (not direct targets), we consider national data reporting measurable but minority LGBTQ+ populations (e.g., Gallup in the U.S.; ONS in the UK), with bisexuals often being the largest subgroup (Gallup, 2023; Office for National Statistics, 2023); (iii) Regarding *residence* versus *nationality*, we account for 25% cross-border residence to mirror global migration trends, with tens of millions living outside their home country, aligning with UN data on international migrants accounting for about 3–4% of the global population (higher in certain regions), and we oversample to ensure a sufficient number of diaspora cases for analysis (United Nations Department of Economic and Social Affairs, 2020).

Cultural attire cues. In a limited number of contexts, image prompts may include minimal attire elements (e.g., headscarves) when these are consistent with commonly observed public norms or context-specific constraints in the assigned setting. The goal is to reflect plausible real-world scenarios rather than to encode stereotypes or normative expectations. These cues are treated purely as contextual visual details and are not intended to imply religious affiliation or personal beliefs.

8.2. Limitations

Synthetic scope. The personas and images are fictional, and the resource does not represent self-identified ethnicity, nationality, religion, or language. Appearance-based clusters are broad and might not accurately reflect mixed heritage or variations.

Granularity and coverage. We focus on achieving a broad scope across 10 clusters and key demographics rather than detailed localities (such as specific Indigenous nations, smaller diasporas, or intersectional identities). Future efforts will aim to incorporate languages, dialects, and more detailed cultural contexts.

Model artifacts. We note common T2I artifacts like hand representations and text renderings, as well as occasional identity shifts across scenes. Qualitative error lists can be found in Section 5.

Evaluation design. The human raters, numbering 77 and representing 9 nationalities, are a convenience sample. Increasing the diversity of evaluators and including professional raters would enhance the conclusions, especially for detecting subtle cultural nuances.

External validity. Design choices such as demographic priors or diaspora sampling are heuristic and intended for controlled benchmarking. DREAM should therefore be used to study model behavior under structured diversity conditions rather than as a proxy for real-world demographics.

8.3. Responsible Use

Appropriate uses include research on multimodal grounding, robustness and bias analysis, identity-consistent generation, and evaluation methodology. Prohibited uses encompass surveillance, law enforcement targeting, or attempts to deduce protected characteristics of real individuals without their consent.

8.4. Data Documentation

We adhere to best practices for documentation (such as Data Statements for sociolinguistic attributes) and provide schema files along with dialogue-specific JSON files with clear keys and label sets to facilitate reproducibility and auditing (Bender and Friedman, 2018).

9. Data and Code Availability

Repository. All code for persona expansion, visual-turn extraction, prompt construction, batch generation, and evaluation is available in a public GitHub repository.[3]

Dataset release. We release (i) the final per-dialogue JSON files, (ii) the per-dialogue visual-turn specification files, and (iii) the generated portraits and scene images. A public subset of DREAM is hosted on Hugging Face Datasets.[4]

Versioning and reproducibility notes. The released artifacts correspond to the model versions available at generation time (LLM: `o4-mini`; T2I: `gpt-image-1`; VLM: `gpt-5`). We provide configuration files, prompts, and deterministic preprocessing scripts to maximize reproducibility under future model changes.

10. Acknowledgment

This work was supported by the European Commission through Project ASTOUND (101071191-HORIZON EIC-2021-PATHFINDERCHALLENGES-01), and by projects BRAINS (PID2024-155948OB-C52) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by “ERDF A way of making Europe”, by the “European Union” and by the Program ’ACTIVIDADES I+D PROCESOS HUMANOS Y SOCIALES’ of the Comunidad de Madrid, Spain, under Project PHS2024/PH-HUM-52 (Innovatrad-CM)". We also want to give thanks to Microsoft Azure services (in particular, to Irving Kwong) for their sponsorship to continue processing new datasets that could be interesting for the dialogue community.

[3] github.com/JuanMallodelaCalle/personalized-ai-visual-summaries

[4] huggingface.co/JuanMallo/dream-75pct

11. Bibliographical References

American Psychological Association. 2023. Gender identity (apa dictionary of psychology). https://dictionary.apa.org/gender-identity.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

Emily M. Bender and Batya Friedman. 2018. Data statements for NLP: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *UIST*.

Centers for Disease Control and Prevention. 2024. Preferred terms for select population groups & communities. https://www.cdc.gov/health-communication/php/toolkit/preferred-terms.html.

Abhishek Das, Satwik Kottur, Khushi Gupta, Amanpreet Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

Gallup. 2023. Lgbt identification in the u.s. https://news.gallup.com/poll/656708/lgbtq-identification-rises.aspx.

Mao Xun Huang and Hen-Hsen Huang. 2024. Integrating text-to-image and vision language models for synergistic dataset generation. In *IJCAI*.

Office for National Statistics. 2023. Sexual orientation, uk: annual estimates. https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/sexuality/bulletins/sexualidentityuk/2023.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. In *ACL*.

United Nations Department of Economic and Social Affairs. 2020. International migrant stock. https://www.un.org/development/desa/pd/.

United Nations Statistics Division. 2024. Standard country or area codes for statistical use (m49). https://unstats.un.org/unsd/methodology/m49/.

Hiromi Wakaki, Yuki Mitsufuji, Yoshinori Maeda, Yukiko Nishimura, Silin Gao, Mengjie Zhao, Keiichi Yamada, and Antoine Bosselut. 2024. Comperdial: Commonsense persona-grounded dialogue dataset and benchmark. *arXiv preprint arXiv:2406.11228*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

12. Language Resource References

Awsaf49. 2018. *PersonaChat (Hugging Face mirror) train_both_revised.txt*. Hugging Face. Hugging Face Datasets. PID https://huggingface.co/datasets/awsaf49/persona-chat. Mirror used in our experiments; original dataset described by Zhang et al. (2018).

Miller, Alexander H. and Feng, Will and Fisch, Adam and Lu, Jiasen and Batra, Dhruv and Bordes, Antoine and Parikh, Devi and Weston, Jason. 2017. *ParlAI: A Dialog Research Software Platform*. Meta (Facebook). GitHub. PID https://github.com/facebookresearch/ParlAI. Software framework hosting the original PersonaChat release.

Sony AI. 2024. *ComperDial (Hugging Face Dataset)*. Sony AI. Hugging Face Datasets. PID https://huggingface.co/datasets/Sony/ComperDial. Community dataset page.

Appendix A. JSON Artifacts and Schema Examples

A.1. Final dialogue JSON (evaluation & playback)

The final per-dialogue JSON integrates (i) two personas (`profile_struct`, `profile_narrative`, `profile_prompt`) and (ii) the dialogue sequence, where `image_id` entries are inserted between turns.

```
{ "dialogue_id": "<DIALOGUE_ID>",
  "profiles": {
    "A": {
      "profile_struct": {...},
      "profile_narrative": {...},
      "profile_prompt": "..."
    },
    "B": {
      ...
    }
  },
  "dialogue": [
    { "persona_id": "<DIALOGUE_ID>_A", "
       text_1": "..." },
    { "persona_id": "<DIALOGUE_ID>_B", "
       text_2": "..." },
    { "image_id": "<DIALOGUE_ID>_img_1"
       },
    { "persona_id": "<DIALOGUE_ID>_B", "
       text_3": "..." },
    ... ]
}
```

A.2. Visual-turn specification JSON (pre-render)

Prior to rendering, we generate a per-dialogue visual-turn specification file listing `visual_turns` (prompt + indices + labels). This supports prompt reuse and reproducible rendering.

```
{ "dialogue_id": "<DIALOGUE_ID>",
  "visual_turns": [
    {
      "image_id": "<DIALOGUE_ID>_img_1",
      "dialogue_indices": ["text_1","
          text_2"],
      "speaker_focus": "both",
      "scene_type": "shared",
      "prompt": "..."
    },
    {
      "image_id": "<DIALOGUE_ID>_img_2",
      "dialogue_indices": ["text_4"],
      "speaker_focus": "B",
      "scene_type": "cutaway",
      "prompt": "..."
    },
    ... ]
}
```

Appendix B. Visual Turn Examples

This appendix shows representative examples of the main `scene_type` categories used in DREAM visual turns. Images from Figures 6, 7, 8, and 9 are shown for illustrative purposes only, and prompts are presented in shortened form for readability.

Figure 6: **Shared scene.** Two speakers meeting in a professional office setting during an initial greeting.

Figure 7: **Cutaway scene.** A close-up cinematic insert of a red clutch abandoned on a cobblestone street.

Figure 8: **Memory scene.** Nostalgic black-and-white childhood memory at a baseball stadium.

Figure 9: **Imagined scene.** Dreamlike sky where clouds form the face of Elvis Presley.

Appendix C. Prompt Templates (Simplified)

This appendix summarizes the core prompt templates used throughout the DREAM pipeline. For readability, we provide simplified template versions that preserve the logical structure and design principles while omitting repetitive instructional details. The complete prompts are available in the public repository (Sec. 9).

C.1. Persona Expansion

The persona expansion prompt transforms short persona descriptions and dialogue turns into structured profiles suitable for visual generation. The template follows an *inference-first* strategy, using controlled fallback values only when attributes cannot be inferred from the character's own content. Additional safeguards enforce age consistency, cultural plausibility, and demographic coherence.

```
SYSTEM ROLE:
"You are an expert in psychological
    inference and character profiling."

INPUT:
- dialogue_id
- original persona sentences
- dialogue turns (A/B speakers)
- fallback demographic values

CORE RULES:
- infer attributes only from the speaker
    's own content
- use fallback values when not inferable
- adjust fallbacks if inconsistent with
    dialogue evidence
- minors safeguards and age-appropriate
    content
- nationality vs. residence constraints

OUTPUT (STRICT JSON):
{
  "dialogue_id": "...",
  "profiles": {
    "A": {
      "profile_struct": {...},
      "profile_narrative": {...}
    },
    "B": {...}
  }
}
```

Key design principles.

- Inference-first attribute assignment to reduce hallucination.
- Controlled fallback distributions for demographic coverage.
- Explicit safeguards for minors and culturally plausible contexts.
- Separation between structured fields and narrative enrichment.

C.2. Profile Portrait Prompt

This prompt generates identity-anchoring profile portraits used later to maintain visual consistency across scenes. The output consists of two concise, self-contained portrait prompts (A and B), each aligned with the expanded profile.

```
TASK:
Create two photorealistic portrait
    prompts from structured profiles.

PROMPT CONTENT:
- age, gender identity, ethnicity
- shoulders-up framing (1:1)
- neutral/soft lighting
- sharp facial focus
- subtle environment hints

CONSTRAINTS:
- strict consistency with profile_struct
- age-appropriate depiction
- cultural attire rules when required

OUTPUT:
A_PROMPT: ...
B_PROMPT: ...
```

Design goal. Portrait prompts define stable identity cues (face, hairstyle, accessories) while leaving pose and scene composition to later stages.

C.3. Visual Turn Selection

The visual-turn selection prompt converts dialogue text into storyboard-like visual scenes. Each selected moment becomes a structured entry defining scene type, speaker focus, and a cinematic scene description.

```
TASK:
Identify visually meaningful moments in
    dialogue.

OUTPUT:
{
  "dialogue_id": "...",
  "visual_turns": [
    {
      "image_id": "...",
      "dialogue_indices": [...],
      "speaker_focus": "A|B|both",
      "scene_type":
        "shared|memory|imagined|cutaway|
            montage",
      "prompt": "cinematic scene
          description"
    }
  ]
}
```

Key design principles.

- Storyboard-style scene extraction.
- Explicit scene taxonomy (`shared`, `memory`, `imagined`, etc.).
- Composition controlled via `speaker_focus`.

- Identity consistency enforced through later combination with portrait prompts.

C.4. Scene Rendering Prompt

Scene rendering combines portrait identity references with scene-level prompts. The scene description is treated as the primary instruction, while portrait prompts act only as identity anchors.

```
PRIORITY ORDER:
1) Scene prompt (primary instruction)
2) Character references (identity
    anchors)

STYLE:
- ultra realistic
- cinematic
- square 1:1 composition
- natural lighting

NEGATIVE CONSTRAINTS:
- do not reuse portrait backgrounds
- do not mix identities
- avoid text overlays or watermarks
```

Design goal. This separation allows scene diversity while preserving character identity across images.

C.5. Automated Portrait Evaluation Prompt

The automated evaluation prompt performs appearance-based demographic perception using a vision–language model. The model must output a strict JSON structure containing ethnicity, gender, and age.

```
TASK:
Estimate demographic attributes from
    portrait image.

OUTPUT (STRICT JSON):
{
  "ethnicity": "<one predefined label>",
  "gender": "<male|female|other>",
  "age": <integer>
}
```

Design goal. The prompt enforces single-label classification to enable direct comparison with human perception results under a stricter evaluation condition.

Appendix D. Qualitative Findings Taxonomy

We report representative examples of the eight recurrent issue categories used to summarize free-text comments in Sec. 5. Examples are shown verbatim and are intended to illustrate the taxonomy.

1. **Anatomy/hands/limbs.** Issues like extra fingers or limbs, and merged hands: “Four fingered left hand”, “Six fingers on left hand”, “Three arms lol”, “Strange fingers on right hand”, “Finger missing”.
2. **Object physics & floaters.** Instances of objects intersecting or floating, and impractical tool usage: “Floating sunglasses”, “The candy jar is floating”, “Laptop sitting on a wet painting”, “Strange way to hold the phone”.
3. **Text & iconography errors.** Errors in spelling, mirrored or visible-through-screen text, nonsensical logos: “Fitness is spelled wrong”, “The computer screen is visible behind the computer”, “Patagionia misspelled”, “HSA instead of NBA”, “Nirvana is missing an N”.
4. **Identity drift (face/age/gender/ethnicity).** Changes in the appearance of speakers between turns or compared to their portrait: “A looks like another person”, “B looks younger”, “A is not the person in the image”, “Different ethnicity”.
5. **Wardrobe/prop continuity.** Shifts or side changes in logos, clothing, eyewear, or tattoos: “Logo changes again”, “B’s shirt is very similar but different”, “Changing tattoo”, “Hair goes through hat”.
6. **Gaze/composition & interaction.** Characters not properly facing each other or the mentioned object: “Not looking at each other”, “She is not looking at the TV”, “He is not looking at the PC”.
7. **Scene-content mismatch.** The depicted scene does not match the dialog: “A is shown cooking but she was talking about restaurants”, “Dreams not very well depicted”, “Zero coherence, nothing about the TV show”.
8. **Object semantics & world knowledge.** Category-level errors involving instruments, vehicles, animals, and attire: “5 tuners on guitar”, “Nurses don’t usually carry stethoscopes”, “Too much helmets”, “That is not a golden”.

Appendix E. Full Dialogue Visual Example

This appendix presents a complete dialogue example from DREAM (Figure 10). The example illustrates how persona conditioning and dialogue grounding produce visually consistent characters across multiple scenes while maintaining coherence with the conversational content.

Speaker A

Speaker B

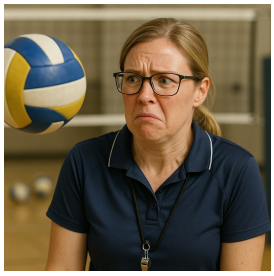
Dialogue transcript (`persona_chat_8310`).

A: Are you going out tonight?
B: Hi. What is with kids these days? I am fun but they do not agree.

Image 1

A: It always a nice time old or not.
B: My sixth graders are really getting out of hand lately.
A: Well I cant party with kids but nice job it seems.
B: I coach too but the girls throw their volley balls at me some times. On purpose!

Image 2



A: Are whoopee cushions still fun?

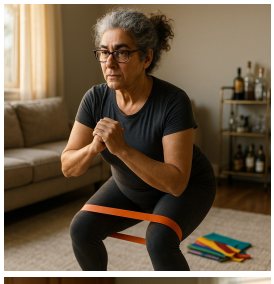
Image 3

B: No. It reminds me of all the pranks they pull on me. Do you like kids?
A: Yes but throw the ball at them?
B: I am too nice. What other teacher holds karaoke friday parties?

Image 4

A: I am too fat to move I am slimming down now.

Image 5



B: Good for you. Try some apples, I am allergic but eat the ones they give me.

Image 6

A: Yea fine I will come sing really loud.
B: Thank goodness. We do crosswords and crochet at the same time.
A: I will bring beer for when the kids leave.
B: Sounds like a plan. We can watch the football game when they are gone too.

Image 7

Summary. Speaker A (Laleh) is a 59-year-old Iranian-American woman with a relaxed, self-deprecating tone and a domestic, leisure-oriented lifestyle; Speaker B (Emma) is a 35-year-old U.S. teacher and volleyball coach, socially active and embedded in school/community routines. Intermediate cutaway scenes ground dialogue references while preserving identity consistency across the sequence.

Figure 10: Full dialogue example showing aligned storyboard frames and transcript.