WORLDWEAVER: BENCHMARKING MULTIMODAL INTENT RECOGNITION AND USER MODELING IN INTERACTIVE VISUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized interactive continuous image generation is an important task for adaptive visual information distribution, in which visual content evolves with user intent and preferences. Existing benchmarks typically study final-image personalization, multimodal intent recognition, and user modeling separately; they rarely represent generated visuals as environments for later feedback or preserve alternative continuations from a shared history. We introduce WorldWeaver-Bench, an extensible benchmark for this setting. Its shared-prefix branching trees support joint evaluation of multimodal interaction-intent recognition, persistent user-profile modeling, and matched comparisons of alternative paths. Experiments show that story context improves the interpretation of current interaction intent, longer and broader interaction evidence improves profile recovery, and preference-compatible paths are not consistently the most diagnostic for user modeling. WorldWeaver-Bench provides a testbed for studying how generated visual content both responds to user preferences and reveals them through interaction.

1 INTRODUCTION

As interaction becomes more deeply integrated into image generation, personalized interactive continuous image generation plays an increasingly important role in adaptive visual information distribution. It generates a sequence of visual states, with each new state determined by the current context and user feedback. The sequence evolves with the user's changing intent and preferences, supporting scenarios such as interactive storytelling, virtual characters, digital education, and visual browsing. Its central challenge is to infer evolving intent, select suitable continuations, and preserve visual and narrative consistency as generated content influences subsequent behavior. Figure 1 illustrates this challenge through an interactive narrative comic, where different user profiles and topics lead to different interaction patterns and visual paths.

Existing work separately addresses personalized image generation (Ruiz et al., 2023; Ye et al., 2023), multimodal intent and grounding (Zhang et al., 2022; Yu et al., 2016), interactive narratives (Riedl & Bulitko, 2013; Zhou et al., 2024), and user modeling (Brusilovsky & Millán, 2007). However, existing benchmarks usually evaluate a final image, an isolated intent, a fixed trajectory, or a logged behavior sequence. They rarely treat generated visual states as environments that drive later feedback, preserve alternative continuations from a shared history, and expose an explicit long-term user-profile target. These gaps motivate an interactive personalized image-generation dataset that records evolving visual states together with user feedback, interaction histories, and user profiles, while preserving alternative continuations for controlled evaluation. Table 1 summarizes these structural differences.

In this paper, we propose WorldWeaver-Bench, an extensible benchmark for personalized interactive continuous image generation. We build it with an automated data-generation flow that initializes user–topic settings, generates feedback-conditioned visual states, expands alternative branches from shared prefixes, and validates annotations and continuity. Interactive narrative comic generation serves as a representative setting, with each user–topic process organized as a shared-prefix branching tree. The benchmark supports multimodal interaction-intent recognition, persistent user-profile modeling, and matched comparisons of alternative continuations under a common history.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

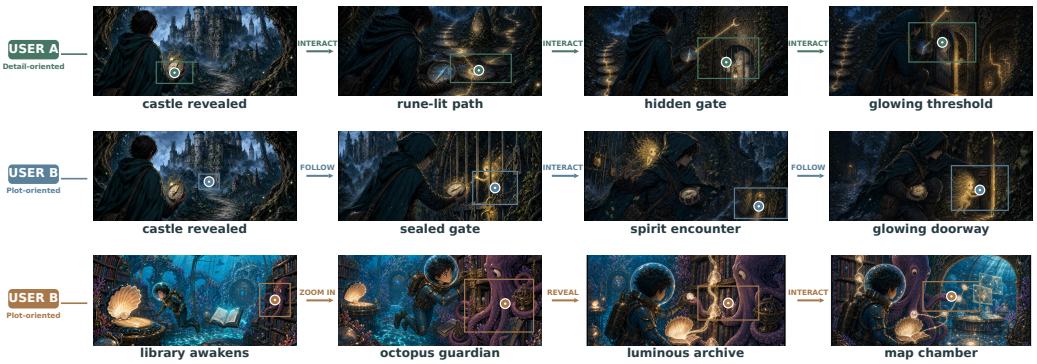

Figure 1: Interactive narrative comic examples of personalized continuous image generation. User A explores Hidden Castle in a detail-oriented manner, while User B follows plot-oriented progressions in Hidden Castle and Underwater Library. Arrows indicate interaction intents.

Table 1: Benchmark comparison. PER.: personalization; INT.: recorded interaction evidence; GEN.: generated visual states used for subsequent interaction; CONT.: multi-turn or multi-step histories; BR.: shared-prefix alternative continuations; PROF.: explicit user profiles; LAB.: localized-target, interaction-intent, or action annotations. △ indicates partial support; for PIPBench, INT. refers to preference feedback, not sequential interaction.

Dataset	PER.	INT.	GEN.	CONT.	BR.	PROF.	LAB.
RefCOCO (Yu et al., 2016)	×	×	×	×	×	×	✓
MIntRec (Zhang et al., 2022)	×	✓	×	×	×	×	✓
DreamBench (Ruiz et al., 2023)	✓	×	×	×	×	×	×
Mind2Web (Deng et al., 2023)	×	✓	×	✓	×	×	✓
MIntRec2.0 (Zhang et al., 2024)	×	✓	×	✓	×	×	✓
PIPBench (Wu et al., 2026)	✓	△	×	×	×	✓	×
WorldWeaver-Bench(Ours)	✓	✓	✓	✓	✓	✓	✓

Our contributions are summarized as follows: (1) We introduce a benchmark and an automated construction pipeline for continuous generated environments, structured user states, and shared-prefix interaction trees. (2) We provide reproducible methods for interaction-intent recognition and user-profile modeling in this setting. (3) We evaluate representative visual encoders, multimodal intent models, and vision–language models, and discuss how alternative interaction paths may differ in their alignment with user preferences and the evidence they provide for user modeling.

2 Related Work

2.1 Interactive Visual Narrative Systems

Interactive narrative research models user actions as transitions in an evolving story world (Riedl & Bulitko, 2013). VIST pairs photo sequences with narrative text (Huang et al., 2016), while Pororo-SV supports story visualization (Li et al., 2019). StoryDALL-E and StoryDiffusion generate coherent visual sequences (Maharana et al., 2022; Zhou et al., 2024). In parallel, DreamBooth, Textual Inversion, and IP-Adapter personalize image generation from subject references (Ruiz et al., 2023; Gal et al., 2023; Ye et al., 2023), while PIPBench evaluates preference-aware generation (Wu et al., 2026). Against this background, WorldWeaver-Bench treats each generated image as both a personalized output and the visual environment for the next interaction, while organizing user–topic processes as shared-prefix branching trees for joint intent recognition and user-profile modeling.

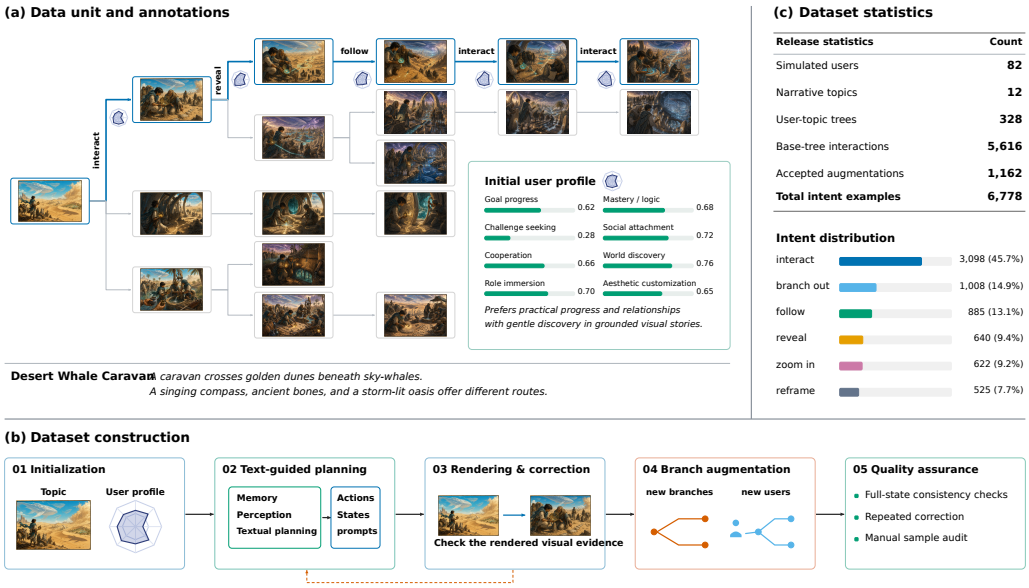

Figure 2: WorldWeaver-Bench overview: (a) a shared-prefix user–topic tree; (b) text-first construction with image correction and branch augmentation; (c) release statistics and intent distribution.

2.2 Multimodal User Understanding

MIntRec and MIntRec2.0 formulate intent understanding as multimodal classification (Zhang et al., 2022; 2024). RefCOCO, Visual Genome, and Flickr30k Entities instead focus on linking linguistic or pointing cues to image regions (Yu et al., 2016; Krishna et al., 2017; Plummer et al., 2015). CLIP supplies a general contrastive visual representation (Radford et al., 2021), and instruction-tuned vision–language models support visual instruction following and multimodal reasoning (Liu et al., 2023). These capabilities are complementary: grounding determines what the user selected, whereas intent recognition asks what change the user expects. Among recent intent methods, CAGC combines contextual augmentation with local and global contrastive learning (Sun et al., 2024), while MVCL-DAF uses dynamic fusion and multi-view supervision (Hu et al., 2025). We adapt these mechanisms to narrative localized interactions.

Beyond turn-level intent, user modeling aims to infer stable preferences and changing states from observed behavior (Brusilovsky & Millán, 2007). Our long-term dimensions are informed by the game-motivation framework of Yee (2006), while the short-term dimensions draw on PAD (Russell & Mehrabian, 1977) and SAM (Bradley & Lang, 1994). These theories motivate the design but do not directly define our scene-specific variables. For computational baselines, PINet provides modality-aware feature interaction (Tang et al., 2025), and MPRNet emphasizes temporal and cross-modal encoding (Li et al., 2025). Most recommendation and personality datasets estimate profiles from a fixed behavior log. By contrast, the shared-prefix structure of WorldWeaver-Bench supports matched comparisons of the evidence revealed by alternative generated paths.

3 The WorldWeaver Dataset

WorldWeaver-Bench represents personalized continuous image generation as a structured interaction process in which visual states, user feedback, story context, and user representations evolve together. An individual session follows a linear sequence, while the dataset retains multiple possible continuations from shared states, forming a branching narrative tree. This section describes the dataset's interaction structure, intent and user-state annotations, automated construction process, and overall statistics. Figure 2 gives an overview of the data organization, construction process, and release statistics.

3.1 INTERACTION STRUCTURE

A user u denotes a persistent identity associated with long-term preferences and motivations, together with short-term states that evolve during interaction. A topic τ specifies the narrative premise, characters, environment, and initial situation. Users assigned to the same topic share an initial visual scene, providing a common starting point for subsequent interactions. At each turn, the system interprets the user feedback using the current image, story state, interaction history, and estimated user state, then generates the next image.

For each user–topic pair (u, τ), the dataset organizes alternative interaction histories as a rooted directed tree

$$\mathcal{T}_{u,\tau} = (V_{u,\tau}, E_{u,\tau}, r_\tau). \tag{1}$$

Here, $V_{u,\tau}$ is the set of nodes in the tree, $E_{u,\tau} \subseteq V_{u,\tau} \times V_{u,\tau}$ is the set of directed parent–child edges, and $r_\tau \in V_{u,\tau}$ is the root node representing the initial scene of topic τ.

Each node $v \in V_{u,\tau}$ represents a visual and narrative state and contains the image I_v, story state s_v, and associated user-state annotations. Each edge $e = (v_i, v_j) \in E_{u,\tau}$ represents an interaction that leads from the parent state v_i to the child state v_j. It records the user feedback and target, intent annotation, state changes, and generation metadata. A path $p = (v_0, e_1, v_1, \ldots, e_L, v_L)$, where $v_0 = r_\tau$ and $e_t = (v_{t-1}, v_t)$, contains L successive transitions and represents one possible session history.

Paths with a shared prefix contain identical nodes and edges up to a branching node v_k (Figure 2(a)). Their common user identity, visual states, and interaction history provide matched starting conditions for comparing alternative continuations.

3.2 ANNOTATION SCHEMA

Intent annotations draw on multimodal intent classification and grounding (Zhang et al., 2022; 2024; Yu et al., 2016); user-state dimensions draw on Yee's motivation framework, PAD, and SAM (Yee, 2006; Russell & Mehrabian, 1977; Bradley & Lang, 1994). We also specify a first-order Markov model for state transitions.

3.2.1 INTENT TAXONOMY

A localized selection specifies where the user is attending, while intent specifies the requested visual or narrative change. We use six categories: *zoom in* (inspect visible detail), *reveal* (expose hidden information), *branch out* (move to a new narrative direction), *reframe* (change the viewpoint or composition), *follow* (continue an existing target or story thread), and *interact* (modify a target or its environment). Each edge also stores a ranked Top-3 list, a natural-language description, and structured slots for the action, target, goal, affect, continuity, and scope; the complete label table is retained in the appendix.

3.2.2 USER-STATE REPRESENTATION

We distinguish relatively stable preferences and motivations from states that may change from one turn to the next. User u at turn t is represented by an eight-dimensional long-term vector $\mathbf{p}_{u,t} \in [0,1]^8$ and a seven-dimensional short-term vector $\mathbf{a}_{u,t} \in [0,1]^7$. These variables are operational definitions for the visual-narrative setting, not general personality traits or psychological diagnoses.

The long-term dimensions reorganize achievement-, social-, and immersion-related factors from the framework of Yee into preferences for goal progress, logical mastery, challenge, social attachment, cooperation, world discovery, role immersion, and visual expression. The short-term dimensions retain pleasure, arousal, and dominance from PAD/SAM and add tension, curiosity, empathy, and cognitive load to describe immediate responses to a narrative scene.

3.2.3 STATE TRANSITION MODEL

For user u at turn t, the joint state is $\mathbf{x}_{u,t} = (\mathbf{p}_{u,t}, \mathbf{a}_{u,t}, I_t, s_t, b_t)$, where I_t is the image, s_t the story state, and b_t the localized interaction and intent. Under the first-order Markov assumption, this state summarizes the history needed for the next transition.

Short-term states retain part of their previous deviation from a common reference while responding to the current event:

$$\mathbf{a}_{u,t+1} = \Pi_{[0,1]^7} \left[\mathbf{a}_{\text{ref}} + \rho \left(\mathbf{a}_{u,t} - \mathbf{a}_{\text{ref}} \right) + \Delta \mathbf{a}_{u,t} \right]. \tag{2}$$

Here, ρ controls persistence and $\Delta \mathbf{a}_{u,t}$ is the scene- and outcome-dependent increment. Long-term preferences change more slowly:

$$\mathbf{p}_{u,t+1} = \Pi_{[0,1]^8} \left[\mathbf{p}_{u,t} + \beta g_p \left(b_t, I_t, s_t, \mathbf{a}_{u,t} \right) \right], \tag{3}$$

where β limits the influence of any single interaction. The simulated behavior policy depends on both components of the user state and on the current environment:

$$P \left(b_t \mid \mathbf{p}_{u,t}, \mathbf{a}_{u,t}, I_t, s_t \right). \tag{4}$$

3.3 Construction Pipeline

WorldWeaver-Bench is constructed offline with simulated users. The pipeline initializes users and topics, expands interaction trees, renders successor images, augments underrepresented branches, and applies structural and visual quality checks. It preserves shared starting conditions, branch provenance, and consistency between planned transitions and rendered states. Figure 2(b) summarizes the construction process.

3.3.1 User-Topic Initialization

Each simulated user is initialized with an eight-dimensional long-term preference and motivation profile, a natural-language profile description, and a seven-dimensional short-term state. The long-term profile is shared across all topics assigned to the same user, whereas the short-term state is updated independently along each interaction path.

Each topic specifies a narrative premise, characters, environment, initial story state, and root-image conditions. Users assigned to the same topic begin from the same non-personalized root scene. Their trajectories subsequently diverge through profile-dependent localized interactions, intents, and narrative choices. This shared initialization provides controlled starting conditions for comparing different users and alternative branches.

3.3.2 Text-Guided Generation

Data generation proceeds from textual planning to image rendering. The Memory Module supplies profile, short-term state, story history, and continuity constraints, while the Perception Agent identifies visible entities and scene-supported interaction targets. Conditioned on these observations, gpt-5.5 produces a structured transition specification containing the target, ranked intents, action slots, before/after story states, affective changes, and the rendering prompt. User preferences select among scene-supported options, but cannot authorize incompatible actions.

The Rendering Agent uses gpt-image-2 with the textual plan and parent image as conditions. The child image is stored as the next node and becomes the visual environment for the following interaction; branches copy the parent memory state and evolve independently after divergence. After rendering, the Perception Agent checks the parent–child pair against the plan and continuity constraints. Because short-term affect depends on the realized image, planned and image-grounded state increments are fused. Inconsistent records are corrected, regenerated, or rejected.

3.3.3 Branch Augmentation

The base construction process naturally favors common actions because some intents are more frequently supported by ordinary narrative scenes. We use two complementary augmentation strategies to improve the coverage of rare intent categories without assigning implausible labels to unsuitable images.

First, we add targeted branches to existing nodes. Candidate nodes are selected according to their visible affordances, and new outgoing branches are generated for underrepresented intents and difficult category boundaries. Each augmented branch remains linked to its source user, topic, parent node, and construction batch, allowing its provenance to be traced.

Second, we construct complete trees for a rare-intent-focused user cohort. We search a pool of candidate profiles for users whose preferences and motivations make underrepresented intents naturally likely, and pair them with topics that contain the corresponding visual affordances. In the current release, ten such users are assigned four topics each, producing 40 additional connected trees. Unlike isolated branch augmentation, these trees provide complete multi-turn paths in which rare intents occur repeatedly while remaining consistent with both the simulated user and the evolving visual environment.

3.3.4 Quality Assurance

Quality assurance combines privileged model context, iterative correction, programmatic validation, and manual inspection. During construction, gpt-5.5 receives an oracle-style view of the user state, ancestor history, story states, continuity constraints, annotations, and rendering conditions; this context is used only to check consistency and is unavailable to evaluated models. Automated checks validate roots, parent–child links, path connectivity, images, required fields, intent labels, numerical ranges, augmentation provenance, and data splits. Ambiguous records are clarified, corrected, regenerated, or rejected, and sampled parent–child records are manually audited for intent, narrative, user compatibility, and continuity.

3.4 Dataset Statistics

Figure 2(c) summarizes the release: 82 simulated users, 12 topics, 328 user–topic trees, 5,616 base interactions, and 6,778 intent examples after 1,162 accepted augmentations. We split by persistent user into 58/8/16 train, validation, and test users. All records from a user remain in one split.

4 Experiments

We evaluate two primary tasks: multimodal interaction-intent recognition and long-term user-profile modeling. We then use the shared-prefix structure for a separate retrospective analysis of how the selected path affects profile recovery.

4.1 Multimodal Interaction Intent Recognition

4.1.1 Experimental Setup

Given the current image I_t, localized interaction information c_t, and the preceding story state s_t, the model predicts one of the six intent categories. The story-context ablation removes s_t while retaining the current image and localization.

We compare CLIP, SAM-CLIP, and ViLT baselines (Radford et al., 2021; Kirillov et al., 2023; Kim et al., 2021), adapted multimodal intent-recognition models, and Qwen3.5, InternVL3.5, and Ministral 3 (Qwen Team, 2026; Wang et al., 2025; Mistral AI, 2025). The models are evaluated with prompt-only inference where applicable and with LoRA fine-tuning for supervised adaptation (Hu et al., 2022). All model and checkpoint choices use the validation split.

Prompted models return a primary category and an ordered Top-3 list. Macro-F1 is the primary metric (Sokolova & Lapalme, 2009). Top-1 measures the final decision, whereas Top-3 measures whether the reference intent remains among the leading alternatives.

4.1.2 Main Results

Table 2 reports the complete comparison. The most consistent result is the value of task-specific supervision: supervised adaptation improves intent recognition across the evaluated model families.

The two intent-specific adaptations also improve on the simpler visual baselines. This comparison suggests that contextual aggregation and adaptive fusion are useful in the narrative setting. The Top-1–Top-3 gap distinguishes candidate coverage from first-choice decision quality.

Table 2: Performance comparison for multimodal interaction intent recognition.

Method	Protocol	Top-1 ↑	Macro-F1 ↑	Top-3 ↑
CLIP	Frozen + MLP head	0.5285	0.4377	0.8832
SAM-CLIP	Frozen + MLP head	0.4774	0.3411	0.8384
ViLT	End-to-end fine-tuning	0.5186	0.4293	0.8501
CAGC-adapted	Context + contrastive	0.5835	0.4949	0.8847
MVCL-DAF-adapted	Dynamic fusion + contrastive	0.5730	0.4906	0.8819
Qwen3.5-9B	Prompt-only	0.4023	0.1847	0.6878
Qwen3.5-9B	LoRA fine-tuning	**0.6405**	**0.5753**	0.8893
InternVL3.5-30B	Prompt-only	0.2626	0.1945	0.6565
InternVL3.5-30B	LoRA fine-tuning	0.6214	0.5399	**0.8924**
Ministral-3-8B	Prompt-only	0.2603	0.1461	0.7382
Ministral-3-8B	LoRA fine-tuning	0.5908	0.4884	0.8840

Model	Input condition	Top-1 ↑	Macro-F1 ↑	Top-3 ↑
Qwen3.5-9B	Image + localization	0.3954	0.1685	0.6511
	+ Story context	0.4023	0.1847	0.6878
InternVL3.5-30B	Image + localization	0.1962	0.1343	0.6397
	+ Story context	0.2626	0.1945	0.6565
Ministral-3-8B	Image + localization	0.1824	0.0912	0.6916
	+ Story context	0.2603	0.1461	0.7382

Table 3: Story-context ablation for multimodal interaction-intent recognition.

4.1.3 STORY-CONTEXT ABLATION

To isolate the contribution of story context, we evaluate the same image–interaction pairs with prompt-only inference under two conditions: current image and localization alone, and the same inputs with preceding story context added. The image, localization marks, model, split, and decoding protocol remain unchanged. Here, context refers to narrative information available before the current interaction, not an explicitly supplied user history or profile.

Table 3 shows improvements across all reported metrics for all three models when story context is included. This supports the value of narrative information in disambiguating localized feedback that is visually compatible with several intents. One possible explanation is that prior choices reflected in the story carry forward preference-related cues.

Figure 3 illustrates how story context can carry forward a user's narrative tendencies, providing an implicit expression of the user profile through prior choices. Without these cues, the continuation tends to follow the direction suggested by the original story premise. With them, the inferred user intent can take precedence over that direction and redirect the plot. In Neon Hospital, a preference for detail discovery favors *reveal*: activating a wall mechanism to uncover a hidden passage, instead of following the diagnostic conduit deeper into the corridor. In Snowbound Beacon, an exploration-oriented user leaves the cabin to investigate the snowstorm, potentially bypassing clues on the map that would otherwise guide the continuation. These examples suggest that story context can help interpret current intent in light of persistent user tendencies, allowing personalized continuations beyond the progression implied by the synopsis.

4.2 USER PROFILE MODELING

4.2.1 EXPERIMENTAL SETUP

Given a set of observations $\mathcal{H}_u$ for user u, the model predicts the initial long-term profile $\hat{\mathbf{p}}_u \in [0,1]^8$. The main setting provides four topics and up to three interactions per topic. For each user–topic tree, we select a deterministic longest connected path; ties are resolved with a fixed random seed and never by consulting the target profile or test error. The target is the initial profile shared across the user topics, not a terminal short-term state. Future images, rendering prompts, affect labels, profile increments, and generated profile summaries are excluded from the input.



Figure 3: Qualitative story-context examples.

Table 4: Performance comparison for user profile modeling.

Method	Added evidence	MAE ↓	Pearson ↑	CCC ↑	W-Spr. ↑	Top-2 ↑
Training mean	n/a	0.1235	n/a	0.0000	0.4477	0.5938
Training median	n/a	0.1257	n/a	0.0000	0.4477	0.5938
PINet-adapted	Story + intent	0.0942	0.5807	0.4633	**0.7041**	**0.7188**
MPRNet-adapted	None	0.0967	0.5895	0.4406	0.6843	0.7083
Qwen3.5-9B (Zero-shot)	Story	0.2058	0.5970	0.3912	0.3656	0.3214
Qwen3.5-9B (LoRA)	Story	0.0836	**0.7432**	0.6031	0.7014	0.6875
Qwen3.5-27B (Zero-shot)	Story	0.2248	0.5753	0.3310	0.3088	0.2813
Qwen3.5-27B (LoRA)	Story	**0.0801**	0.6705	**0.6269**	0.7040	0.6250

We compare constant training-profile priors, lightweight supervised profile models over frozen visual features, and general vision–language models. The profile models receive visual and localized interaction evidence, with ablations that add story context or the intent produced by the perception stage. Qwen models directly predict the profile dimensions from serialized interaction evidence, with LoRA checkpoints selected using validation User MAE.

User MAE, based on mean absolute error (Willmott & Matsuura, 2005), gives each user equal weight after averaging over the eight dimensions. Pearson correlation (Pearson, 1895) measures whether a model distinguishes users along each dimension, while concordance correlation (CCC) also penalizes bias in location and scale (Lin, 1989). Within-user Spearman (W-Spr.) (Spearman, 1904) evaluates the ordering of the eight dimensions for each user, and Top-2 recall measures whether the two strongest dimensions are recovered.

4.2.2 Main Results

Table 4 reports the main profile-modeling comparison. Supervised adaptation substantially improves numerical profile recovery. The larger supervised Qwen model gives the strongest absolute-error and concordance results, whereas the smaller supervised variant best distinguishes users across dimensions. PINet-adapted, despite its compact trainable component, performs best on within-user ranking and Top-2 recall. The methods therefore excel at different aspects of the task: calibration, cross-user discrimination, and within-user preference ranking are related but not interchangeable.

Zero-shot Qwen models capture some cross-user variation but have higher MAE than constant priors, suggesting poor numerical calibration. LoRA substantially reduces this error.

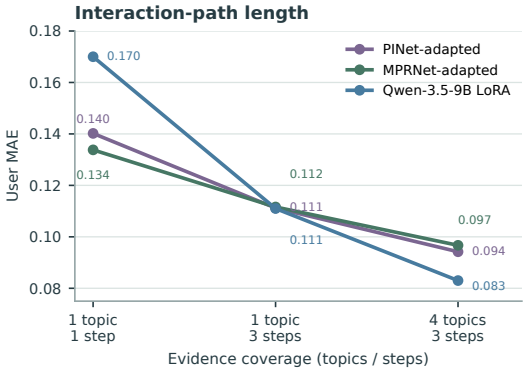

Figure 4: Interaction-evidence ablation for user profile modeling.

4.2.3 Effect of Interaction-Path Length

We examine how interaction-path length affects user-profile modeling. The first two conditions in Figure 4 show that extending the history within one topic reduces profile error across all three models. The third condition additionally increases topic coverage; its improvement cannot be attributed to path length alone.

4.3 Preference–Diagnosticity Analysis

We use the shared-prefix structure for a complementary analysis of whether a continuation that is more compatible with a user preference is also more informative for recovering the user profile. Each comparison pairs continuations that share the same prefix but differ in the continuation selected by the construction-time behavior policy. One continuation has higher simulated preference propensity and the other has lower propensity.

For each path p, we measure the eight-dimensional profile error $E(p)$. We define the preference–diagnosticity gap as

$$\Delta = E(p^{\text{high}}) - E(p^{\text{low}}), \tag{5}$$

where p^{high} and p^{low} denote the higher- and lower-propensity continuations, respectively. A positive value means that the lower-propensity path produces lower profile error. The analysis indicates that preference-compatible paths are not necessarily the most diagnostic: a less preferred continuation can expose stronger evidence about the user's profile. Detailed quantitative results for this analysis are reported in Appendix C.3.

5 Conclusion

We introduced WorldWeaver-Bench, a benchmark for personalized interaction in systems that generate continuous image sequences. Its shared-prefix narrative trees treat each image both as the result of one turn and as the environment for the next, allowing intent recognition, user-profile modeling, and matched branch analysis within a common data structure. Across both primary tasks, task-specific supervision provides the clearest improvement, while story context consistently improves intent recognition. Longer interaction histories generally improve profile recovery, but equally long branches can reveal different amounts of user information. These findings motivate future work on real-user collection and on estimating the expected modeling value of a candidate continuation before it is generated, while preserving narrative quality and user experience.

AI Use Statement

GPT-5.5 was used as an assistive tool during the preparation of this work. In the data-generation pipeline, it helped produce candidate visual-state descriptions, interaction-intent annotations, user-state records, and related construction materials, which were subsequently filtered and checked using the specified validation procedures. It was also used to support literature research, including locating relevant studies. During manuscript preparation, GPT-5.5 assisted with English-language polishing,

sentence revision, and improving the clarity and organization of the exposition. The authors verified the cited sources, experimental descriptions, numerical results, and final claims, and remain fully responsible for the data, analyses, interpretations, and submitted manuscript.

ETHICS STATEMENT

WorldWeaver-Bench is a synthetic benchmark: its users, profiles, interactions, and visual narratives are generated or simulated rather than collected from human participants. The release therefore contains no participant recruitment, personal identifiers, or private behavioral records, and the reported profiles are not intended as clinical, psychological, or individual-level assessments. Generated images, text, and model-assisted annotations may nevertheless inherit biases, stereotypes, or other artifacts from the underlying models. The release documentation will describe these limitations, the provenance of the generation models, and applicable usage restrictions. The benchmark is intended for research on interactive intent recognition and user modeling, not for automated decisions about real users; future studies involving human participants should obtain appropriate consent and ethics approval.

REPRODUCIBILITY STATEMENT

The dataset construction code has been released through an anonymous repository at `https://anonymous.4open.science/r/WorldWeaver-Construction-CE18/`. The dataset and supporting evaluation artifacts will be made publicly available in a subsequent release. This release will include the user–topic interaction trees, images and annotations, fixed train/validation/test splits, and the prompts, validation rules, evaluation scripts, model configurations, random seeds, software versions, and per-run metadata needed to reproduce the reported analyses. This paper describes the task inputs, metrics, and evaluation protocols. Release documentation will additionally provide checksums, data-format descriptions, and reproduction instructions.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Margaret M. Bradley and Peter J. Lang. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

Peter Brusilovsky and Eva Millán. User models for adaptive hypermedia and adaptive educational systems. In *The Adaptive Web*, pp. 3–53. Springer, 2007.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using Textual Inversion. In *International Conference on Learning Representations*, 2023.

Bo Hu, Kai Zhang, Yanghai Zhang, and Yuyang Ye. Adaptive multimodal fusion: Dynamic attention allocation for intent recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17267–17275, 2025. doi: 10.1609/aaai.v39i16.33898.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239, 2016. doi: 10.18653/v1/N16-1147.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5583–5594, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Jian Li, Huawei Zhang, Junyi Chen, Junhui Gong, Yufan Wang, Shifeng Wang, and Yuliang Zhao. MPRNet: A temporal-aware cross-modal encoding framework for personality recognition. *IEEE Transactions on Affective Computing*, 16(4):3342–3353, 2025. doi: 10.1109/TAFFC.2025.3601134.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. StoryGAN: A sequential conditional GAN for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6329–6338, 2019.

Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989. doi: 10.2307/2532051.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. StoryDALL-E: Adapting pretrained text-to-image transformers for story continuation. In *Computer Vision – ECCV 2022*, pp. 70–87, 2022. doi: 10.1007/978-3-031-19836-6_5.

Mistral AI. Introducing Mistral 3. Mistral AI technical blog, December 2025. URL https://mistral.ai/news/mistral-3.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015. doi: 10.1109/ICCV.2015.303.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen technical blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Mark O. Riedl and Vadim Bulitko. Interactive narrative: An intelligent systems approach. *AI Magazine*, 34(1):67–77, 2013.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.

James A. Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26963–26973, 2024. doi: 10.1109/CVPR52733.2024.02546. URL https://openaccess.thecvf.com/content/CVPR2024/html/Sun_Contextual_Augmented_Global_Contrast_for_Multimodal_Intent_Recognition_CVPR_2024_paper.html.

Bin Tang, Ke-Qi Pan, Miao Zheng, Ning Zhou, Jia-Lu Sui, Dandan Zhu, Cheng-Long Deng, and Shu-Guang Kuai. Pose as a modality: A psychology-inspired network for personality recognition with a new multimodal dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (2):1538–1546, 2025. doi: 10.1609/aaai.v39i2.32145.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1): 79–82, 2005. doi: 10.3354/cr030079.

Yuhang Wu, Shuxiang Zhang, Wee Hian Ching, Chi Zhang, and Miao Liu. PIPBench: A profile-inclusive framework for personalized image generation evaluation. In *Computer Vision – ECCV 2026*, pp. 456–473, 2026. doi: 10.1007/978-3-032-36969-7_25.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Nick Yee. Motivations for play in online games. *CyberPsychology & Behavior*, 9(6):772–775, 2006.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pp. 69–85, 2016. doi: 10.1007/978-3-319-46475-6_5.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. MIntRec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1688–1697, 2022. doi: 10.1145/3503161.3547906.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2403.10943.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/c7138635035501eb71b0adf6ddc319d6-Abstract-Conference.html.

A Annotation Details

This appendix specifies the intent labels, user-state dimensions, and record structure used in WorldWeaver-Bench. The user-state dimensions are operational annotations for simulated visual-narrative interactions, not psychological measurements of real participants.

A.1 Interaction Intents

Table A1 defines the six intent categories. Each edge also records a ranked Top-3 list and structured action, target, narrative goal, affect, continuity, and scope slots. The localized target identifies where the interaction occurs; the intent describes the requested change.

Table A1: Intent categories and interpretations.

Intent	Interpretation
Zoom in	Examine finer details of a visible target.
Reveal	Expose hidden or unavailable information.
Branch out	Move to a new location, character, clue, or direction.
Reframe	View the same event from a different composition.
Follow	Continue an existing character, object, or story thread.
Interact	Act on a target to change an object or environment.

A.2 User-State Dimensions

Table A2 separates eight long-term preferences from seven short-term states, with all values in $[0, 1]$. The initial long-term profile is shared across a user's topics and is the profile-modeling target; path-dependent states are stored separately.

Table A2: Long-term preferences and short-term states.

Type	Dimension	Interpretation
Long-term	Goal progress	Preference for clear goals and progress.
	Mastery/logic	Interest in rules, causality, and problem solving.
	Challenge seeking	Preference for difficulty, risk, and uncertainty.
	Social attachment	Interest in bonds with characters.
	Cooperation	Preference for helping and shared outcomes.
	World discovery	Interest in places, clues, and world knowledge.
	Role immersion	Preference for character perspectives and roles.
	Aesthetic preference	Preference for style and customization.
Short-term	Pleasure	Valence of the current experience.
	Arousal	Current activation and excitement.
	Dominance	Perceived control over the situation.
	Tension	Current pressure or uncertainty.
	Curiosity	Desire to explore unknown information.
	Empathy	Emotional resonance with characters.
	Cognitive load	Mental effort required by the interaction.

A.3 Record Organization

Topic records contain the narrative premise and shared root scene. Node records store images, story states, and user-state annotations. Edge records link parent and child nodes and store localization, intent, state changes, and generation metadata. Branches retain their source user, topic, and parent node, together with augmentation provenance where applicable. Construction annotations are not automatically exposed as benchmark inputs; the input boundaries are summarized in Appendix C.

B TOPIC ROOTS AND STORY SYNOPSES

Table A3 presents all 12 topics. Each root image provides a shared initial visual state, while its synopsis establishes the narrative premise and possible directions for interaction.

Table A3: Root images and story synopses for the 12 topics in WorldWeaver-Bench.

Root image	Topic and synopsis	Root image	Topic and synopsis
	**Abandoned Space Elevator** At the base of a silent orbital elevator, a maintenance badge suddenly reactivates, offering paths through the control deck, cable shaft, and derelict station ruins above.		**Desert Whale Caravan** A caravan crosses a sea of golden dunes beneath enormous sky-whales, where a singing compass, ancient bones, and a storm-lit oasis suggest different routes forward.
	**Hidden Castle** A mist-veiled castle appears only when moonlight strikes an antique compass, inviting a traveler to choose between the main gate, a side path, and the secrets hidden in the forest edge.		**Miniature City Inside a Watch** Inside an opened pocket watch, a tiny city runs on gears and bells, and a newly arrived repairer must choose which district of the living mechanism to touch first.
	**Neon Hospital** In a rainlit future clinic, a diagnostic AI detects an impossible vital sign, opening choices between the patient room, the glowing medical console, and a sealed emergency corridor.		**Ocean Festival on Paper Lantern Island** On an island lit by floating paper lanterns, the tide carries a message in a bottle, inviting choices among the festival pier, a lantern boat, and a moonlit reef path.
	**Rainy Cyberpunk Alley** In a neon alley during a storm, a courier receives a scrambled signal that points toward a vending shrine, a rooftop route, and a shadowed figure under the rain.		**Snow Mountain Radio Tower** At a remote mountain radio tower, a signal from beneath the snow begins to answer back, pointing toward the antenna, an old rescue map, and lights moving in the storm.
	**Talking Animal Courtroom** A woodland courtroom convenes after a stolen moonberry pie vanishes, and the first scene offers witnesses, evidence, and a suspicious judge's gavel as possible leads.		**Time Loop Train Station** A train station clock repeats the same minute, and a traveler must decide whether to board the impossible train, inspect the clock, or follow echoes from previous loops.
	**Underwater Library** A submerged library protected by glass domes awakens when a diver opens a shell-shaped catalog, revealing books, currents, and sea creatures that may each guide the next discovery.		**Volcanic Flower Market** A market of fireproof flowers blooms beside a gentle volcano, where a rare ember seed reacts to heat, music, and rival gardeners competing for its next form.

C Supplementary Evaluation Details

C.1 Input Boundaries and Evaluation Protocol

The split is defined by persistent user identity: 58 users for training, eight for validation, and 16 for testing. All records associated with a user remain in the same split. Model and checkpoint selection uses the validation split. Table A4 summarizes the two primary tasks.

Table A4: Inputs and targets for the primary evaluation tasks.

Task	Observed evidence	Prediction target
Intent recognition	Current image, localization, and preceding story context.	Primary intent and ranked Top-3 intents.
Story-context ablation	The same image and localization, with or without preceding story context.	The same intent targets.
Profile modeling	Connected interaction histories with visual and localized interaction evidence; added story or predicted-intent evidence follows each model setting.	The initial eight-dimensional long-term profile.

Story context is preceding narrative information, not an explicitly supplied user profile or user-history record. The intent ablation holds the image, localization, model, split, and decoding protocol fixed. For profile modeling, the main setting uses four topics and up to three interactions per topic, selected along a deterministic longest connected path. Ties use a fixed seed, without consulting target profiles or test errors. Future images, rendering prompts, affect labels, profile increments, and generated profile summaries are excluded from profile-model inputs.

C.2 Interaction-Evidence Conditions

Figure 4 compares one topic with one interaction, one topic with three interactions, and four topics with three interactions per topic. The first comparison varies within-topic history length; the second expands topic coverage at a fixed per-topic interaction budget. These conditions distinguish the depth of an observed path from the breadth of evidence across topics.

C.3 Preference–Diagnosticity Analysis

For the quantitative analysis summarized in Section 4.3, we construct 64 matched pairs from 16 unseen users and four topics. Within each pair, the two five-step continuations share their prefix and differ only in the continuation selected by the construction-time behavior policy. One path has higher simulated preference propensity and the other has lower propensity. MM denotes multimodal input, combining images and interaction text; ST denotes serialized-text input, representing the interaction history as text only. The gap is defined in Section 4.3; positive values favor the lower-propensity path in profile recovery.

Table A5: Preference–diagnosticity gaps with 95% intervals. Positive gaps indicate that the lower-propensity path yields lower profile error.

Model	Input	Gap Δ	95% interval
Qwen3-VL-8B	MM	0.0255	$[0.0116,\ 0.0402]$
Qwen3-8B	ST	0.0083	$[-0.0043,\ 0.0224]$
Qwen3-VL-32B	MM	0.0076	$[-0.0019,\ 0.0172]$
Qwen3.8-27B	MM	0.0061	$[-0.0060,\ 0.0194]$
Qwen3.8-27B	ST	0.0098	$[-0.0019,\ 0.0218]$
InternVL3.5-30B	MM	0.0001	$[-0.0183,\ 0.0187]$

Table A5 reports the matched-path results. Only Qwen3-VL-8B (Bai et al., 2025) has a confidence interval excluding zero, and its lower-propensity path yields lower profile error. The remaining comparisons are inconclusive, reinforcing the qualitative conclusion that preference compatibility

and profile diagnosticity are not equivalent. This is a retrospective comparison under a simulated behavior policy, not evidence that less-preferred content improves real-user experience.