Grounding the Visual Axis of a Narrative World Model: Consistency-Aware Reference Generation for Long-Form Stories

Divyajot Singh
PocketFM
ext-divyajot.singh@pocketfm.com

Imon Banerjee
PocketFM
imon.banerjee@pocketfm.com

Shourya Gupta
PocketFM
shourya.gupta@pocketfm.com

Aaryaman Jain
PocketFM
aaryaman.jain@pocketfm.com

Shreyash Mishra
PocketFM
shreyash.mishra@pocketfm.com

Vasu Sharma
PocketFM
vasu.sharma@pocketfm.com

Abstract

Diffusion image generators are individually strong but stateless: each call has no memory of what a character looked like a panel ago. Producing imagery for a long-form serialized story—with hundreds of episodes, evolving characters, large wardrobes, and nested locations—by calling such models independently yields drifting identities, inconsistent outfits, and changing locations. We build on a Narrative World Model (NWM) [11]: a text/audio substrate that grounds a story's narrative and audio/voice axes over an episode→scene→beat structure, but contains no structured, beat-resolved, or renderable visual representation. We present VividWorld, a visual world-modeling layer that (i) separates noisy beat-level visual observations from a range-based continuity resolution model of base, persistent, conditional, and transient visual state; (ii) models distinct visual identities and derivable entities such as outfits, time-of-day variants, and interactions; and (iii) materializes a deterministic reference-image dependency DAG that covers the show's visual equivalence classes and conditions each render on the canonical references it must match. A pose/quality-adaptive face-embedding metric turns "is this the same character?" into a measurable gate. On seven production shows, the main gain is stability rather than raw attribute coverage: VividWorld cuts hard identity contradictions ∼3.3× (to 0.23 per 1k observations) versus a strong memory-bearing baseline, with zero within-scene wardrobe flicker in a whole-library audit. A small downstream panel test shows the same pattern at render time: adding the reference library improves VLM-rated identity consistency from 2.50 to 4.25 on a 1–5 scale and wins all four forced-choice comparisons.

1. Introduction

For serialized visual content, consistency is the product. Audiences immediately notice when a protagonist's eye color shifts, an outfit changes mid-conversation, or a recurring room is rebuilt between episodes. Yet dominant text-to-image and image-editing models have no intrinsic notion of "the same character as before." Two difficulties make a naive per-panel approach fail at show scale. Statelessness causes drift: independent sampling cannot guarantee that the hundredth image of a character matches the first; without an anchor, renders wander. Audio-first sources under-specify the visual: written for audio drama, they rarely state visual state. Clothing is described once and then assumed, physical traits are revealed gradually, and props appear without stage directions, so a faithful adaptation must recover latent visual state from prose and narrative logic, not merely transcribe it.

What is a visual reference library? A visual reference library is a structured set of canonical images—one per visual equivalence class—anchoring every panel render to the correct character identity, outfit, and location. Building one for a long-form show from text/audio is hard because: appearance must be recovered from prose rather than transcribed; characters may need multiple distinct references (child vs. adult, human vs. transformed); derived entities (outfits in different lighting, object interactions) must be generated in dependency order; and an upstream change should propagate to only affected entries. Fig. 1 illustrates the structure. To our knowledge, no prior work addresses reference-library construction at the scale of a long-running serialized show.

We build on a Narrative World Model (NWM) [11]—a text/audio substrate providing a global character/lo-

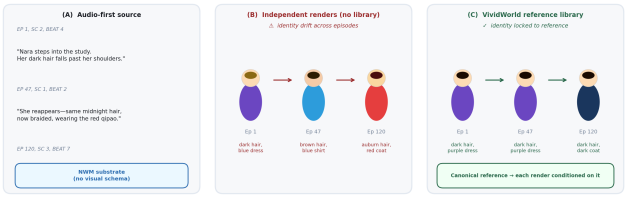

Figure 1. Without a reference library, independently sampled renders drift across episodes (panel B). VividWorld recovers latent visual state from text/audio and anchors every panel to canonical references (panel C).

cation registry and a hierarchical episode→scene→beat structure (a beat is the atomic dramaturgical unit, e.g., a self-contained exchange or action)—that intentionally contains no visual representation. We present VividWorld, which fills this gap and renders an internally consistent reference library. Contributions: (1) a visual-state extraction and continuity-resolution algorithm (Sec. 3.1) converting noisy beat observations into typed range lifecycles; (2) a visual identity / entity model (Sec. 3.2) detecting when one character needs multiple references and deriving outfits, lighting, and interactions as first-class entities; (3) a reference-library Directed Acyclic Graph (DAG) (Sec. 3.3) with prerequisite-gated generation, recoverable lineage, and cascade-correct regeneration without model fine-tuning; (4) a pose/quality-adaptive identity-consistency metric (Sec. 3.4) serving as both automated QC gate and evaluation measure, validated on seven production shows (Sec. 4).

2. Related Work

Subject-personalization methods (DreamBooth [10], Textual Inversion [6], IP-Adapter [17], InstantID [14], ConsiStory [13], StoryDiffusion [19], The Chosen One [1]) operate at the level of a single subject or short batch; we use such conditioning as a primitive and contribute the scheduling policy that decides which canonical references each render receives. Story-visualization pipelines [2, 3, 7–9, 12, 15, 16] emit a sequence of panels for a handful of characters from a short caption or story. Our problem is upstream and at greater scale: building a consistent reference library for an entire long-form show from a text/audio substrate, via beat-level visual-state resolution and a dependency-gated DAG. ControlNet [18] conditions generation on control signals natively. We measure identity via ArcFace-style [4] embeddings (InsightFace [5]).

3. Method

Input substrate (NWM). We take as input a Narrative World Model [11] (prior/concurrent work): per character it records a prose description, demographics, a role, narrative histories, per-episode state, and an audio/voice grounding. The narrative is organized as episodes → scenes → beats, where each beat is the atomic dramaturgical unit (a self-contained exchange or action) annotated with the characters present. Appearance appears only incidentally, as free prose or narrative state; there is no visual attribute schema, no per-beat visual record, and no images. VividWorld fills this gap in two phases: visual grounding (Secs. 3.1–3.2) and consistency-aware generation (Sec. 3.3). All intermediate artifacts are persisted, so the pipeline is idempotent and resumable. Full NWM schema details are in Appendix B.

3.1. Visual-State Extraction and Continuity Resolution

We separate observation from resolution. Observation extracts, per character per beat, a structured appearance schema (hair, eyes, skin, height, build, age, species-form), worn outfits (constrained to a wardrobe bank with ownership tracking), and active props. Every non-null value carries an evidence record: observed value, rationale, an evidence strength $\in \{\text{explicit}, \text{likely}, \text{inferred}\}$, and a change hint from a fixed taxonomy (base-revelation, persistent/conditional change, transient-damage, …). Ungrounded values are emitted as not_specified rather than guessed.

Resolution converts noisy, contradictory per-beat observations into a range-based continuity model. Each trait is typed as base, persistent (from a beat onward), conditional (active under a specific state), or transient (a wound, makeup), with explicit (episode, scene, beat) start/end coordinates. Two principles ensure completeness: an earliest-explicit-revelation rule (the first explicit reveal of a base trait overrides later, lower-confidence observations) and wardrobe gap-filling (unresolved beats back-filled from the nearest established outfit). Applying the resolved ranges back to every beat materializes a canonical per-beat visual state. Full algorithm and evidence-conflict rules are in Appendix C.

3.2. A Visual Identity / Entity Model

A narrative character is not the same as a visual identity. Beats with stable visual attributes are grouped into variant clusters (e.g., child vs. adult, human vs. transformed)—split on species/age/permanent-mark changes, never on emotion or clothing—so one reference can span non-contiguous ranges. Outfits and time-

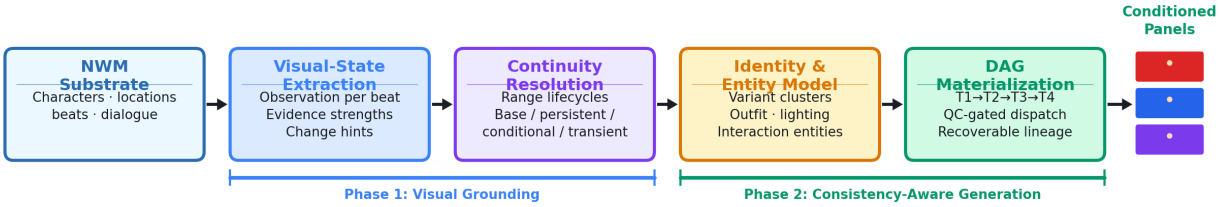

Figure 2. VividWorld pipeline. NWM substrate → visual-state extraction (per-beat observations with evidence strengths) → continuity resolution (base / persistent / conditional / transient ranges) → identity & entity model → QC-gated DAG materialization (T1→T4) → conditioned downstream panels.

Table 1. The tiered dependency DAG. Identity and style propagate down the lineage via conditioning.

Tier	Output	Conditioned on
T1	Char. sheet / loc. / object	Style references only
T2	Character variant	Parent variant (identity-lock edit)
	Sub-location / day-night	Parent background
T3	Look-state (dressed)	Identity sheet and garment ref.
T4	Interaction	Look-state and object ref.

of-day are derivable entities keyed by composite keys (e.g., char___outfit, tod___night), capturing combinatorics without enumerating panels. Each entity's prose is rewritten into a visual-only description that feeds image prompts.

3.3. Consistency-Aware Tiered Reference-Image DAG

Rather than generating each image independently, we render a minimal spanning set of canonical references—one per visual equivalence class—in dependency order, conditioning every render on the completed references it must match. Consistency comes from reference conditioning and prompt policy alone; no model fine-tuning is used. Tiers (Table 1) correspond to dependency depth: primaries (T1) render against style references only; each deeper tier conditions on the completed tier above it. Fig. 3 visualizes the resulting structure. Full DAG construction algorithm and equivalence-key definitions are in Appendix D.

Dependency-gated dispatch and lineage. The DAG is materialized at once as pending rows, each stamped with a prerequisite snapshot $prereq = \{e_i \mapsto v_i\}$; a row is eligible only when every upstream entity has a completed image at the recorded version. Tier completion triggers a quality-control (QC) review gate rather than auto-advancing, so errors are caught before propagating downstream. A content hash over prompt inputs makes identical regeneration a no-op; a version/active/-parent lineage keeps exactly one active image per group, enabling cascade-correct and crash-safe regeneration.

3.4. Identity-Consistency Metric

Against a cached ground-truth reference set, a candidate I is scored after CLAHE lighting normalisation and face-pose detection. For ground-truth embeddings $\{g_j\}$ and pose angles θ,

$$s = w(\theta)\ \max_j \cos\big(\phi(I),\, g_j\big), \tag{1}$$
$$w(\theta) = \max\big(0.7,\ 1 - \tfrac{0.3}{90^\circ}\max\big(|\theta_y|, |\theta_p|, |\theta_r|\big)\big)\,.$$

The image passes if $s \geq \tau(\theta, q)$; the verdict serves as an automated QC gate and evaluation metric. A human gallery complements it for wardrobe and location consistency.

4. Experiments

We evaluate the stage prior single-subject methods lack: visual-state extraction and continuity resolution from an audio-first substrate. The benchmark covers seven production shows (600–1000 episodes each), ∼29k per-(episode, character) observations, and ∼66k audited character-beats. All arms use the same instruction-LLM.

Baselines. Naive runs a stateless per-episode extraction prompt with no prior appearance. Rolling context prepends a flat accumulated character card (200–400 tokens/character) updated after each episode, with no evidence strengths or range lifecycles. Full prompt templates and context details are in Appendix A.

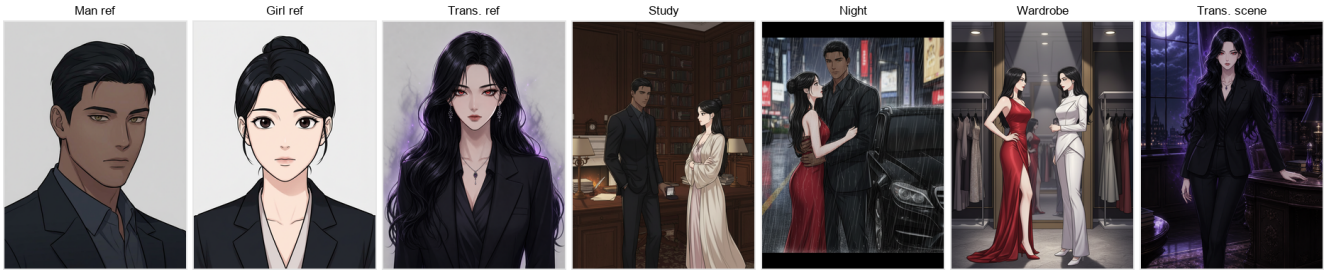

Figure 3. Reference scheduling methodology. Canonical reference nodes (base man, base girl, transformed-form variant) are selected by visual-identity equivalence keys; downstream panels condition on the active identity, outfit, location, night/rain lighting, or transformed-form parents along the DAG edges. See Appendix B for construction details.

Table 2. Continuity extraction across seven shows (∼29k observations per method, same backbone LLM). Coverage is higher-better; Tier-A contradictions per 1k are lower-better.

Method	Coverage (%)↑ Hair	Eyes	Skin	Build	Tier-A /1k↓
Naive (no memory)	2	2	1	4	2.69
Rolling context	29	24	28	48	0.76
VividWorld (ours)	30	26	29	58	0.23

Table 3. Blind forced-choice win rates (%↑), ∼340 images, 9 annotators.

Method	Char. id.	Style	Artifact
Naive	11	19	22
Rolling context	26	30	31
VividWorld	63	51	47

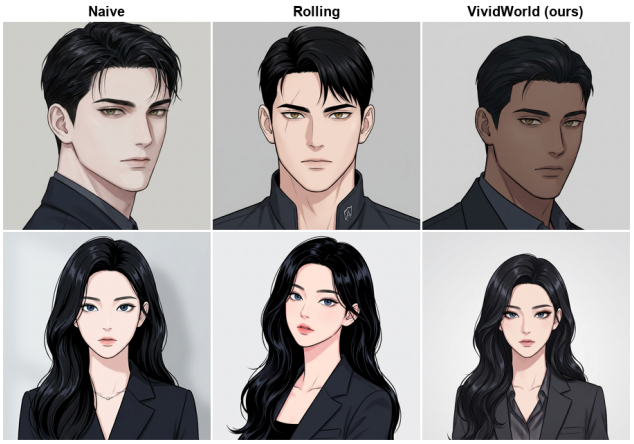

Figure 4. Qualitative comparison. Prompt fixed; visual-state source varies: Naive (no memory), Rolling (flat card), VividWorld (range-resolved, reference-conditioned). VividWorld locks identity to the canonical reference.

Continuity and faithfulness. Table 2 shows that memory is necessary: hair coverage rises from 2% for Naive to 29–30% for memory-bearing methods. The key comparison is VividWorld vs. Rolling: hair/eye/skin coverage is close, so the contribution is not more extraction, but better resolution. VividWorld improves build coverage (58% vs. 48%) and cuts Tier-A contradictions to 0.23/1k vs. 0.76 for Rolling and 2.69 for Naive; the full audit finds only ∼7 Tier-A contradictions and zero within-scene wardrobe flicker.

Stratified audit. Explicit traits are rarely contradicted; location, prop, and injury states are mostly latent fills rather than source-verifiable (Appendix E).

Rendered references and human study. VividWorld references score 65% attribute fidelity. Fig. 4 and Table 3: VividWorld wins at 63/51/47% on character identity, style, and artifact preference ($\kappa = 0.62$).

Ablations and downstream panels. Reference conditioning locks identity better than text-only arms (IP-Adapter: 0.75 vs. 0.25–0.36); VividWorld contributes the whole-show reference scheduling—panels improve from 2.50 to 4.25 identity consistency and win 4/4 forced choices (Appendices F–G).

5. Limitations and Conclusion

VividWorld can introduce unsupported visual details; evidence-strength gating reduces but does not eliminate this risk. The attribute schema and identity gate are anthropocentric—non-human characters require species-specific taxonomies and perceptual embeddings. VividWorld grounds the missing visual axis of a narrative world model: it extracts and resolves beat-level visual state, models distinct visual identities and derivable entities, and renders a consistent reference library along a dependency DAG with recoverable lineage.

References

[1] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In ACM SIGGRAPH, 2024.

[2] Junhao Cheng, Xi Lu, Hanhui Li, et al. AutoStudio: Crafting consistent subjects in multi-turn interactive image generation. arXiv preprint arXiv:2406.01388, 2024.

[3] Junhao Cheng, Baiqiao Yin, Kaixin Cai, et al. TheaterGen: Character management with LLM for consistent multi-turn image generation. arXiv preprint arXiv:2404.18919, 2024.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In CVPR, 2019.

[5] Jiankang Deng, Jia Guo, et al. InsightFace: 2D and 3D face analysis project. https://github.com/deepinsight/insightface, 2021.

[6] 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 ICLR, 2023.

[7] Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. DreamStory: Open-domain story visualization by LLM-guided multi-subject consistent diffusion. arXiv preprint arXiv:2407.12899, 2024.

[8] 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 CVPR, 2019.

[9] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. StoryDALL-E: Adapting pretrained text-to-image transformers for story continuation. In ECCV, 2022.

[10] 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 CVPR, 2023.

[11] Mohammad Saifullah, Thomas Kornmaier, Taaha Kazi, Vasu Sharma, Aditya Sanjiv Kanade, and Aanand Kumar Yadav. Narrative World Model: Narratology-grounded writer memory for long-form fiction. arXiv preprint arXiv:2607.05577, 2026.

[12] Jinlu Tang, Jiquan Wang, Yuze Yu, et al. StoryWeaver: A unified world model for knowledge-enhanced story character customization. arXiv preprint arXiv:2412.07375, 2024.

[13] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. ACM Transactions on Graphics (SIGGRAPH), 2024.

[14] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. InstantID: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519, 2024.

[15] Jianzong Wu, Chao Tang, Jingbo Wang, et al. DiffSensei: Bridging multi-modal LLMs and diffusion models for customized manga generation. arXiv preprint arXiv:2412.07589, 2024.

[16] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. SEED-Story: Multimodal long story generation with large language model. arXiv preprint arXiv:2407.08683, 2024.

[17] 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.

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, 2023.

[19] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. In NeurIPS, 2024.

A. Baseline Implementation Details

Naive baseline. Each episode is processed in isolation. The extraction prompt provides the NWM character registry (name, role, prose description) and the full episode text, then asks the LLM to return—for each character present—values for the same visual attribute keys used by VividWorld: hair color/length/style, eye color, skin tone, height, build, age band, outfits, and active props. No prior appearance state is supplied; every call produces an independent visual snapshot with no memory of preceding episodes.

Rolling context baseline. A running appearance registry (one character card per character) is maintained and updated after each episode. The extraction prompt is augmented with the serialised card prepended as: "Current known appearance of [name]: {card}. Update or confirm each field based on the episode below; return unchanged fields as-is." The registry is seeded from the NWM prose description and overwritten episode-by-episode in broadcast order. No evidence strengths, change-hint taxonomy, or range lifecycle tracking is applied. Context is bounded by the accumulated character card size (typically 200–400 tokens per character)—not a sliding window of past beats—so raw episode history is not re-consumed. Both baselines use the same per-episode call granularity as VividWorld's observation stage; the difference lies in what prior state is supplied and how contradictions are resolved downstream.

B. Narrative World Model (NWM) Details

The Narrative World Model [11] is a structured multimodal representation that grounds the narrative and audio axes of a serialized story. For each character, NWM records: a prose description capturing role and personality; demographics (age, species, occupation); a narrative history tracking major character arcs across episodes; a per-episode state snapshot; and an audio/voice grounding including samples, speech rhythm, vocabulary patterns, and verbal tics. Locations and objects similarly carry descriptors and their narrative lifecycles. The narrative hierarchy is formalized as episode→scene→beat, where beats are dramaturgical units (e.g., a conversation, an action sequence) annotated with the set of characters present, dialogue, stage directions, and symbolic narrative events. This structure enables reasoning about what characters know, who is present where, and narrative causality—the narrative grounding. However, NWM intentionally does not provide any visual representation: no attribute schema for appearance, no per-beat visual record, no wardrobe codification, and no images. This is by design, as NWM is sourced from audio-first scripts that leave the visual implicit. VividWorld fills this gap by recovering and rendering the missing visual axis.

Compact input schema. VividWorld only assumes four NWM record families, summarized in Table 4. This makes the interface independent of any particular NWM implementation.

Table 4. Compact NWM input records consumed by VividWorld.

Record	Required fields
Character	id, name, prose description, demographics, role/history, episode states, optional voice/audio cues.
Beat	episode, scene, beat, location_id, present characters, dialogue/stage directions, symbolic events.
Location	id, name, prose descriptor, parent/sub-location links, lifecycle or episode availability.
Object	id, name, prose descriptor, owner/holder when known, lifecycle or episode availability.


```
{"characters":[{"id":"c_nara","name":"Nara",
  "description":"guarded heir",
  "demographics":{"age":"adult"},
  "voice":{"rhythm":"controlled"}}],
 "locations":[{"id":"l_study","name":"study",
  "description":"wood-paneled room"}],
 "objects":[{"id":"o_ring","name":"ring",
  "description":"engagement ring"}],
 "beats":[{"episode":12,"scene":3,"beat":2,
  "location_id":"l_study",
  "present":["c_nara","c_gordon"],
  "stage":"Nara confronts Gordon",
  "events":["proposal_confrontation"]}]}
```


Qualitative figure construction. Fig. 3 is built from the same reference-library abstraction used by the generator. The first three tiles are entity-level references: two base visual identities and one transformed-form variant. The downstream tiles are panel-level composition targets whose beat state activates different parent sets: base identity plus location for the study scene, base identity plus a night/rain location variant, base identity plus wardrobe range for the boutique scene, and transformed identity plus location for the transformed scene. The generated transformed tiles are illustrative qualitative assets rather than additional quantitative test cases; the quantitative downstream panel ablation remains the four-case test in Appendix G.

C. Visual Resolution Details

Observation schema. For every beat, VividWorld records the characters present and a sparse visual observation tuple $o = (c, t, \ell, a, w, p, r)$, where c is a canonical character, t is the episode/scene/beat coordinate, ℓ is the resolved location when available, a contains physical attributes, w contains active clothing candidates, p contains active props, and r contains rationale strings or evidence strengths. Attribute keys are intentionally visual: hair color/length/style, eye color/shape/features, skin tone/scars/markings/tattoos, height, build, age band, and species/form. Momentary emotional expressions are not base attributes; when they imply a visible state (blood, bandage, pallor, disguise), they are routed into transient ranges or overrides.

Evidence strengths and conflict handling. The resolver separates three kinds of evidence. Direct observations come from explicit prose or visual-stage statements in a beat. Revelations are later statements that should update a base trait retroactively, such as a delayed first mention of eye color. Contextual fills are lower-confidence states such as default wardrobe implied by role, class, or prior range continuity. Conflicts are resolved by attribute-specific compatibility functions rather than string equality: “dark hair” is compatible with black or brown, “well-built” with muscular/athletic, and so on. A contradiction is emitted only when two known values are incompatible and no active override, transformation, disguise, or script-stated change explains the transition.

Range lifecycles. The output is a set of ranges rather than a single global character card. Base traits apply across all beats unless superseded. Persistent changes

start at a source coordinate and continue until another change. Conditional changes are activated only under a condition such as transformed form, disguise, or time skip. Transient ranges represent injuries, blood, bandages, illness, or soiling, and must expire. Clothing and prop ranges carry relation labels (wearing, holding, using, wielding, etc.) and continuity hints such as persistent_until_change. This range model is what makes downstream reference scheduling possible: the render planner can ask which identity, outfit, location, prop, and interaction references are active for a target panel.

D. DAG Materialization and Regeneration

Equivalence keys. The reference library is built from deterministic equivalence keys. A character identity key groups beats with the same stable body/species/age/permanent-mark state. A look-state key pairs an identity key with a primary outfit. A location key represents a parent or sub-location; time-of-day keys derive lighting variants from locations. Interaction keys combine a look-state, object, relation, and sometimes a location class. These keys define what "minimal" means: render one canonical reference for every key actually used by the show, not one image per beat.

Dependency edges. Edges point from a render to the references it must condition on. Character variants depend on parent identity references; child locations and day/night variants depend on base backgrounds; look-states depend on identity and garment references; interactions depend on look-states and object references. Each pending render row stores a snapshot of the upstream active image versions it was built against. A row is eligible only if all parents have completed active images at those versions. This makes dispatch crash-safe and regeneration local: when an upstream reference changes, only rows whose prerequisite snapshot is stale must be regenerated.

Operational ablations. Rolling context removes the range resolver but leaves a memory-augmented extractor. Grounding-only image generation keeps the resolved brief but removes all image-reference conditioning. Text-only downstream panels keep the same story prompt but remove the reference images. These ablations isolate the two central claims: range resolution reduces contradictions, and reference-conditioned DAG scheduling converts resolved state into stable downstream imagery.

E. Stratified Visual-State Audit

Table 5 gives the full stratified audit. Claim-side precision samples start from VividWorld outputs and ask a judge whether the source prose supports, contradicts, or omits each claim. The precision denominator excludes absent claims because script silence is not evidence of correctness or incorrectness. Source-side recall samples start from episodes, ask the judge to extract explicit visual-state facts, and then check whether VividWorld captures each fact in the resolved state.

Table 5. Stratified visual-state audit. Claim-side precision is supported/(supported+contradicted); script-silent absent claims are reported separately. Recall is over explicit source-side visual facts from sampled episodes.

Stratum	Claims	n_p	Prec.	Abs.	Rec. (n_r)
Explicit traits	39.4k	30	100.0	73.3	51.8 (168)
Inferred traits	3	3	0.0	66.7	— (0)
Wardrobe fills	40.4k	30	66.7	90.0	51.9 (104)
Transformations	1.3k	30	50.0	73.3	42.9 (7)
Injuries	1.9k	30	100.0	76.7	16.2 (68)
Locations	158	30	76.9	13.3	9.6 (136)
Props	5.4k	30	64.0	10.0	22.3 (188)

n_p=claim-side sample; n_r=source-side facts.

Interpretation. The audit supports a narrower and more defensible claim than "we recover all hidden visual state." Explicit traits are checkably precise in the sample, and injuries are usually not contradicted when the source says enough to adjudicate them. However, the large absent rates for wardrobe, transformation, and injury claims show that much of the library is operationally useful latent state rather than source-verifiable annotation. Location and prop recall are the clearest remaining weaknesses: many source-visible settings and objects are not materialized into the compact resolved state, even when the narrative prose states them.

F. Reference-Conditioning Primitive Benchmark

Table 6 evaluates image-side identity primitives on a common SDXL backbone: stateless text-to-image, a grounding-only VividWorld ablation with no reference DAG, face-embedding conditioning, shared-attention methods, and reference conditioning. We use 35 characters across the seven shows and three views per character. All target prompts use the same script-derived conditioning; the shared identity anchor for encoder methods is synthesized from script gold, not from VividWorld outputs.

Table 6. Reference-conditioning primitive benchmark. Ident: identity self-consistency. Face: face-detection rate. Style: art-style coherence. GoldFid: % of script-stated attributes read back.

Method	Type	Ident↑	Face↑	Style↑	GoldFid↑
Indep. (no mem.)	T2I	0.35	97.1	0.68	47.1
Rolling	T2I	0.36	97.1	0.66	53.3
Ground. only	T2I	0.25	98.1	0.63	50.0
InstantID	Embed	0.26	92.4	0.68	58.8
ConsiStory	Attn.	0.36	100	0.76	37.5
StoryDiff.	Attn.	0.65	100	0.83	37.5
IP-Adapter	Cond.	0.75	99.0	0.73	38.9

Takeaway. IP-Adapter and StoryDiffusion demonstrate that conditioning mechanisms can lock identity across views better than text-only arms, whereas the grounding-only VividWorld ablation improves script-attribute fidelity over independent text-only generation but still drifts without image references. Table 6 is therefore not a full VividWorld row. The full system uses this primitive operationally: it selects and schedules a whole show's reference library so each render receives the right identity, outfit, location, and object parents.

G. Downstream Panel Test

We render four story panels from show 2998 twice. Both arms receive the same panel prompt and character descriptions. The text-only arm receives no images. The VividWorld references arm additionally receives the relevant character reference portraits. A VLM judge scores identity, outfit, and location consistency on a 1–5 scale and chooses the better image for sequence use.

Table 7. Per-case downstream panel ablation detail. Scores are VLM ratings on a 1–5 scale.

	Text only			VividWorld refs			
Case	Id.	Out.	Loc.	Id.	Out.	Loc.	Winner
Study confrontation	2	4	5	5	5	5	VividWorld
Rain rescue	3	4	5	5	5	5	VividWorld
Gala rivalry	3	5	5	5	5	5	VividWorld
Boutique confrontation	2	5	5	2	5	5	VividWorld

The gain is concentrated in identity consistency because the prompts explicitly state outfit and location. The boutique case is informative: both arms collapse Nara and Yvonne into similar-looking women, so the VLM assigns equal identity scores, but still prefers the reference-conditioned panel for composition and lighting. This supports a cautious claim: references help downstream scene generation, but the test is small and should be scaled to more panels and human raters.