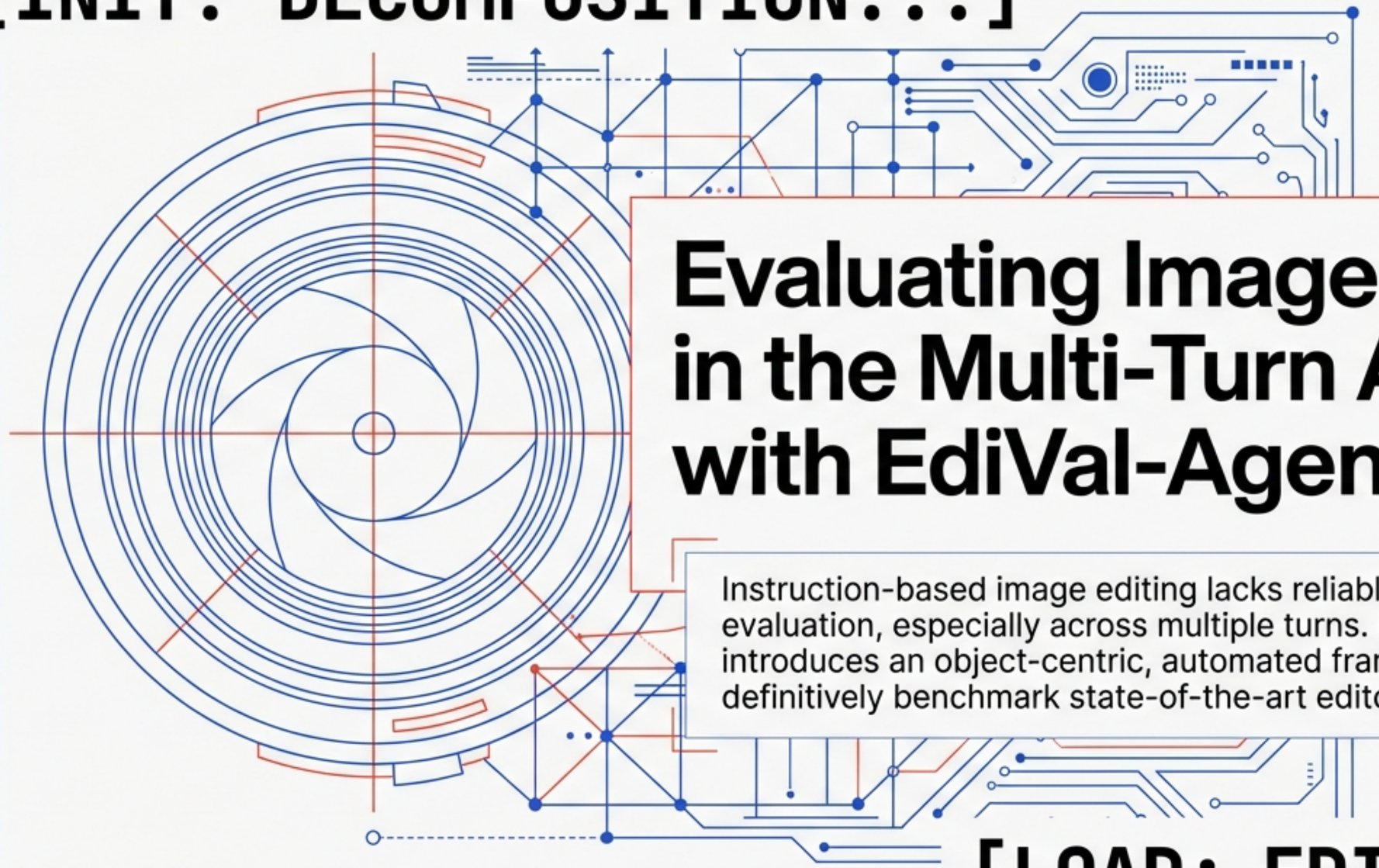
[INIT: DECOMPOSITION...]



Evaluating Image Editors in the Multi-Turn Arena with EdiVal-Agent

Instruction-based image editing lacks reliable, interpretable evaluation, especially across multiple turns. EdiVal-Agent introduces an object-centric, automated framework to definitively benchmark state-of-the-art editors.

[LOAD: EDIVAL-AGENT]

Current evaluation paradigms fail to capture true multi-turn capability

	Reference-Based (L1/CLIP)	VLM-Only (Zero-Shot)	EdiVal-Agent (Object-Centric)
Coverage	Single Realization	High	Dynamic & Compositional
Bias	Inherits Generator Bias	Natural Image Priors	Neutral Object Tracking
Spatial Reasoning	Pixel-Locked	Hallucinates Positions	Precise Geometric Checks
Pixel Sensitivity	Overly Rigid	Misses Subtle Edits	Feature-Level Preservation

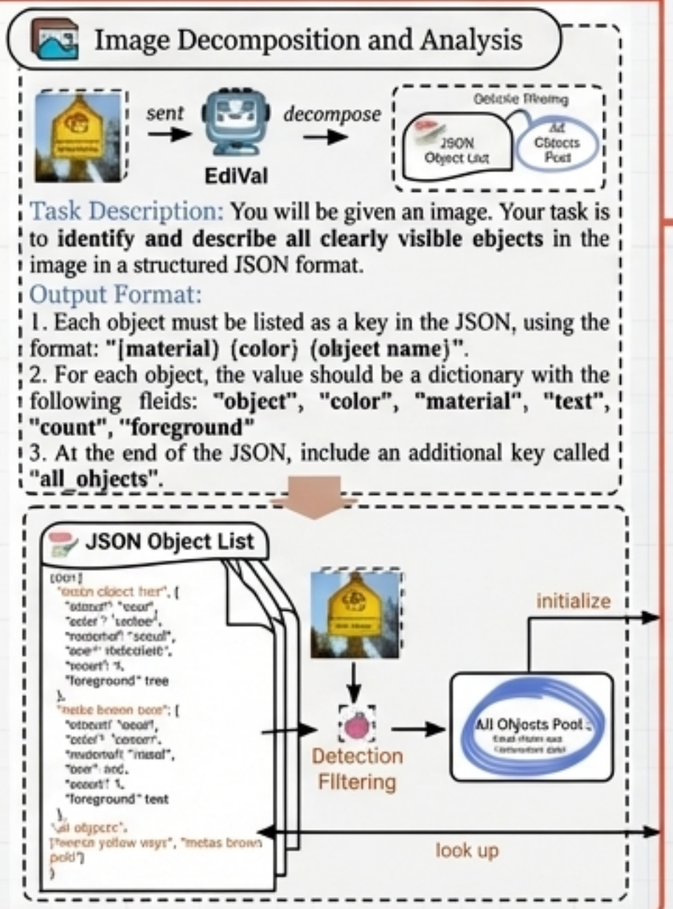
Anatomy of an automated, object-centric evaluation agent

Step 1: Parse scene into structured, symbolically trackable objects.

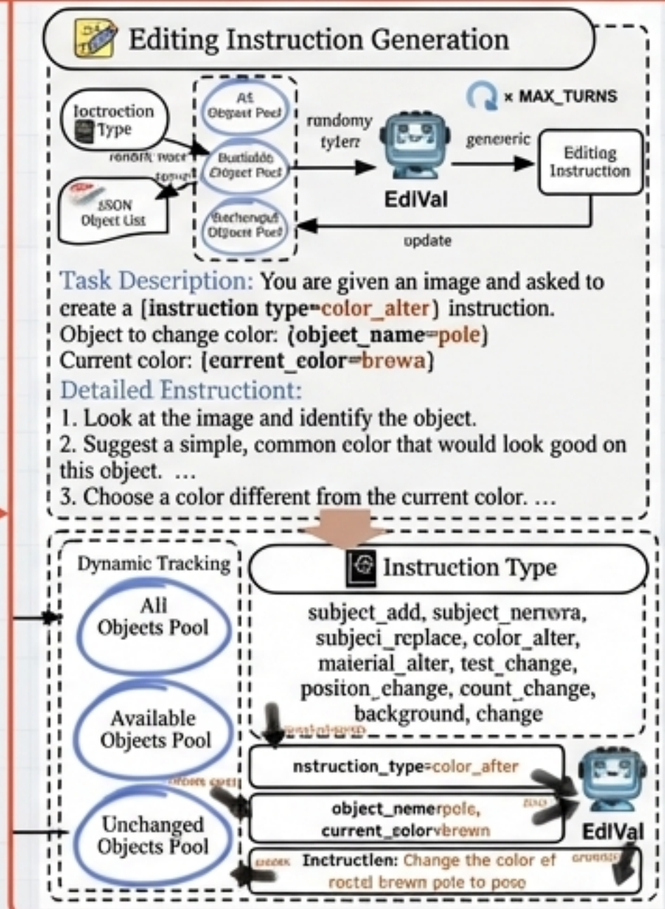
Step 2: Synthesize context-aware instructions and track state across turns.

Step 3: Measure precise success using specialized tools (Detectors + VLMs + Preference Models).

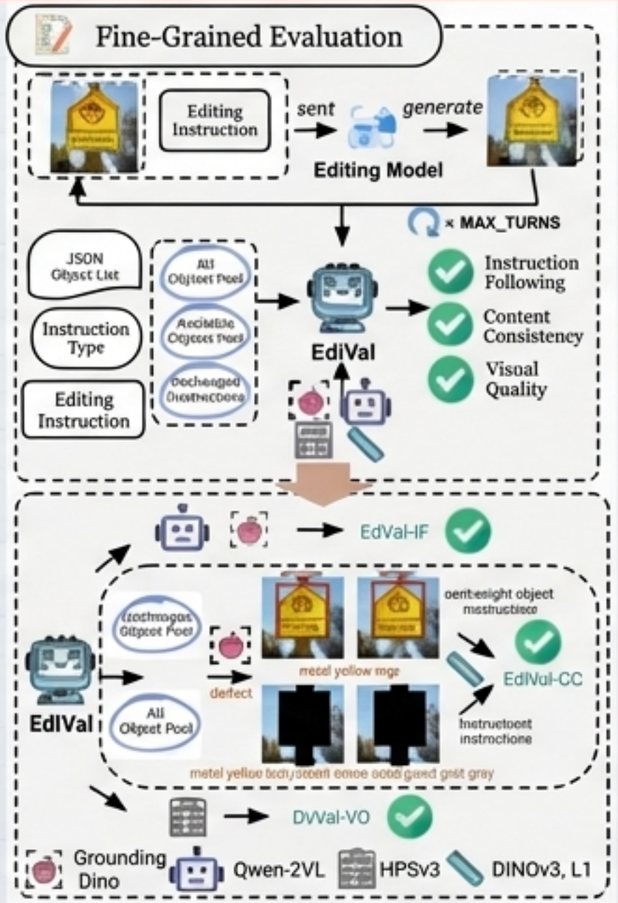
Decomposition



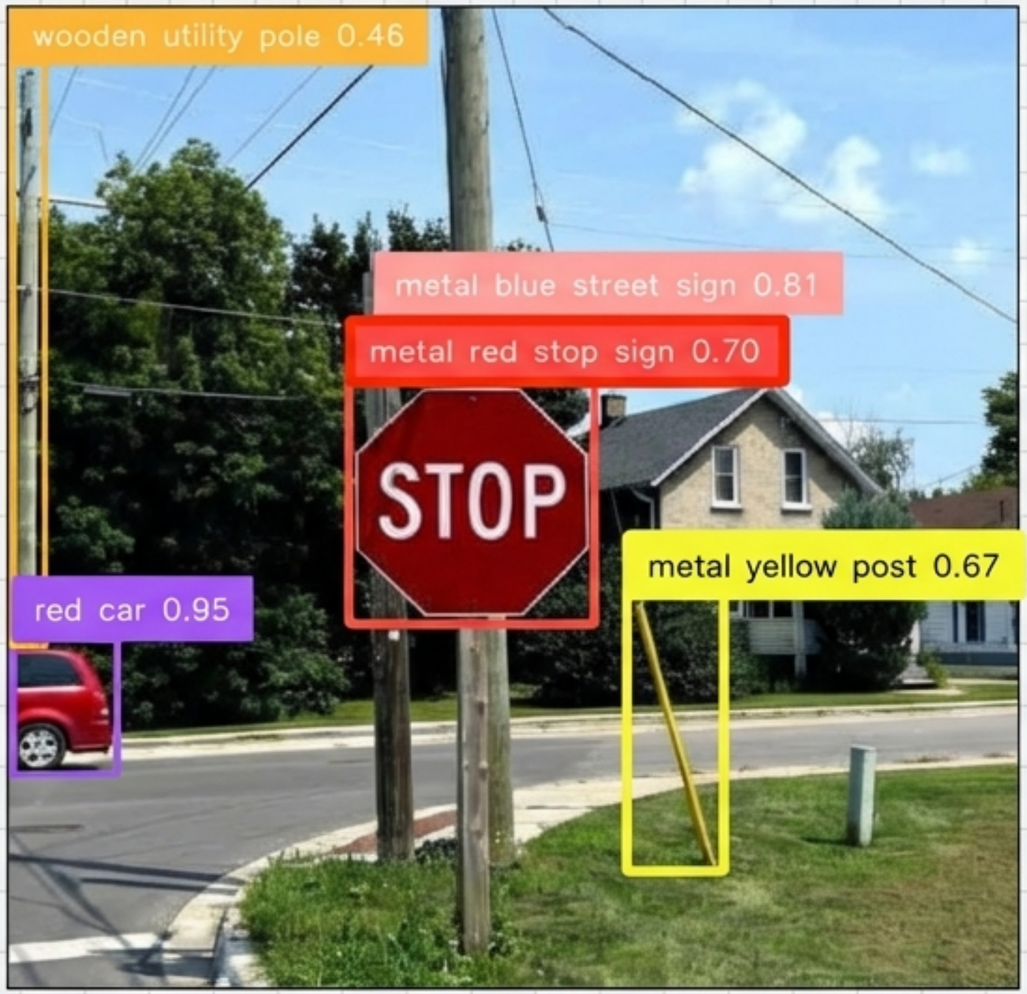
Instruction Generation



Evaluation



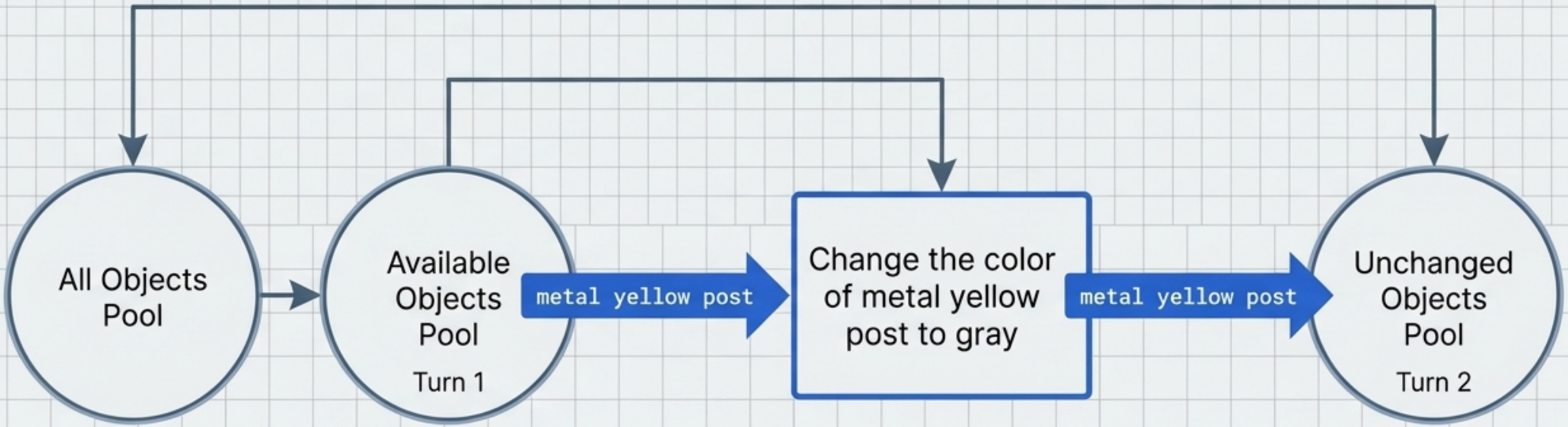
Step 1: Translating pixels into verifiable symbolic logic



Grounding-DINO validates VLM extraction, pruning low-confidence hallucinations.

```
{
  "metal yellow post": {
    "object": "post",
    "color": "yellow",
    "material": "metal",
    "count": 1,
    "foreground": true
  }
}
```

Step 2: Dynamic tracking prevents context collapse across turns

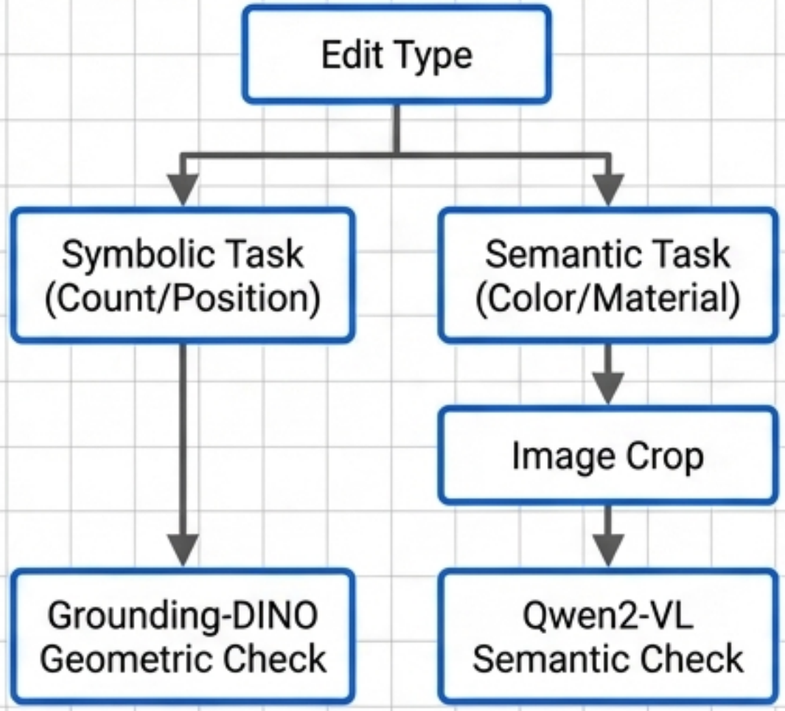


Core Mechanic:
At each turn t, the agent exp., expands or contracts the available object pool based on the intended edit, ensuring subsequent prompts never reference erased or physically impossible objects.

Scope:
9 instruction types across 3 sequential turns.

Step 3: A tripartite metric for comprehensive assessment

EdiVal-IF
(Instruction Following)



EdiVal-CC
(Content Consistency)

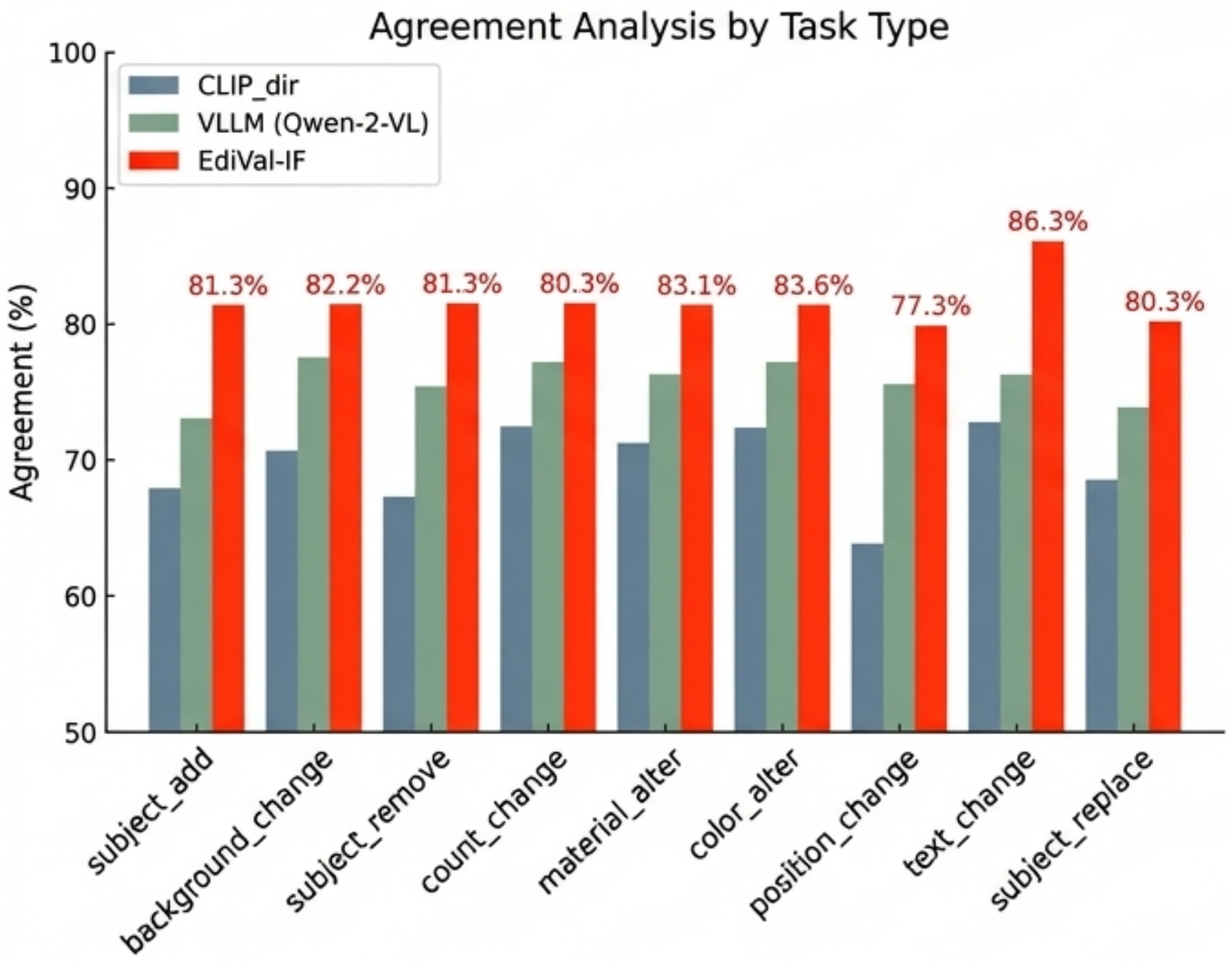
[Base Image Background Crop] + [Unchanged Object DINOv3 Features] + Semantic Preservation Score

EdiVal-VQ
(Visual Quality)

Quantifies overall artifact and quality shifts.

Grounding-DINO for geometry, **Qwen2-VL** for localized semantics, **DINOv3** for feature-level preservation.

The EdiVal-IF metric aligns with human judgment



Ranking (Accuracy)

1. EdiVal-IF: 81.3% Accuracy

2. VLM-Only: 75.2% Accuracy

3. CLIP Directional: 65.4% Accuracy

Takeaway:

By combining object detection with VLM reasoning, EdiVal avoids the spatial hallucination trap of zero-shot VLMs, fundamentally closing the gap to human baseline **(85.5%)**.

The Arena: Benchmarking 16 models across the multi-turn gauntlet

In-Context Autoregressive

Gemini 2.0 Flash
GPT-Image-1.5
Nano Banana 2

Flow Matching

FLUX.1-Kontext
Qwen-Image-Edit
Step1X-Edit

Diffusion

MagicBrush
AnyEdit
UltraEdit

The Gauntlet: 1,716 instructions, 9 constraint types, exactly 3 sequential turns.
The Objective: Survive Turn 3 with high instruction fidelity while preventing background collapse.

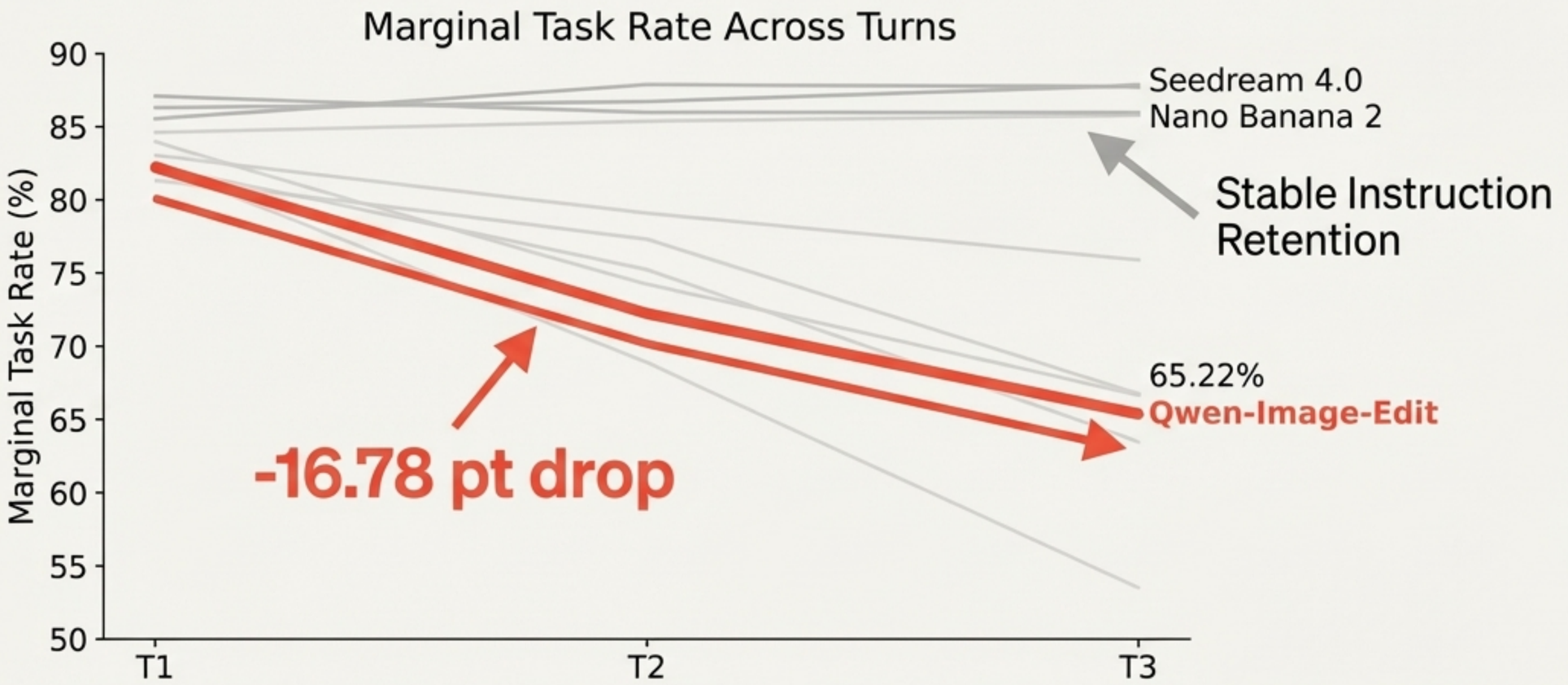
The Multi-Turn Leaderboard: Closed-source systems dominate

Rank	Model	Detail
01	Seedream 4.0	Highest overall EdiVal-O at T1/T3, strong cross-turn consistency.
02	GPT-Image-1.5	Massive CC improvement over previous GPT-Image-1.
03	Nano Banana 2	Consistent instruction following and low spatial degradation.
04	FLUX.2-max	Strongest flow matching baseline, outperforming pure diffusion.

Insight 1: A stark gap remains between closed-source API models and open-source models.

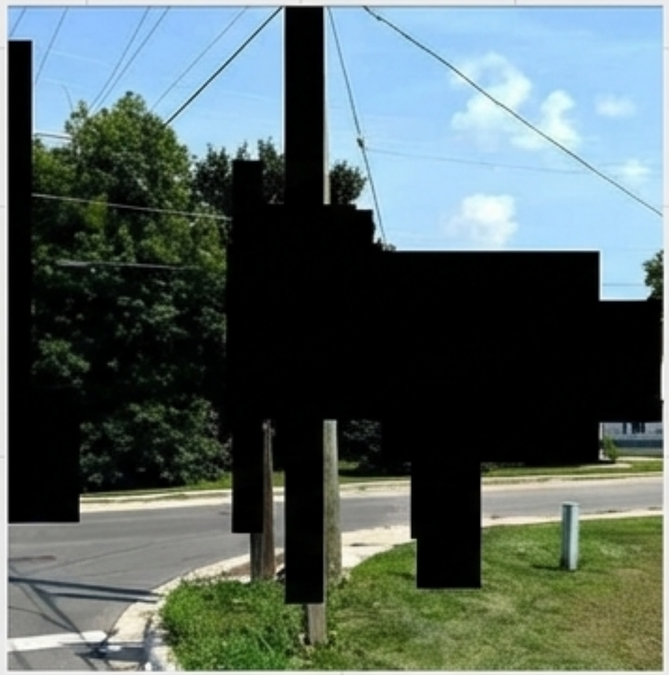
Insight 2: Latency Trade-off. Newer model versions (GPT-Image-1.5, Nano Banana 2) improve multi-turn robustness but incur significant speed penalties.

Error accumulation: The Turn 3 degradation cliff



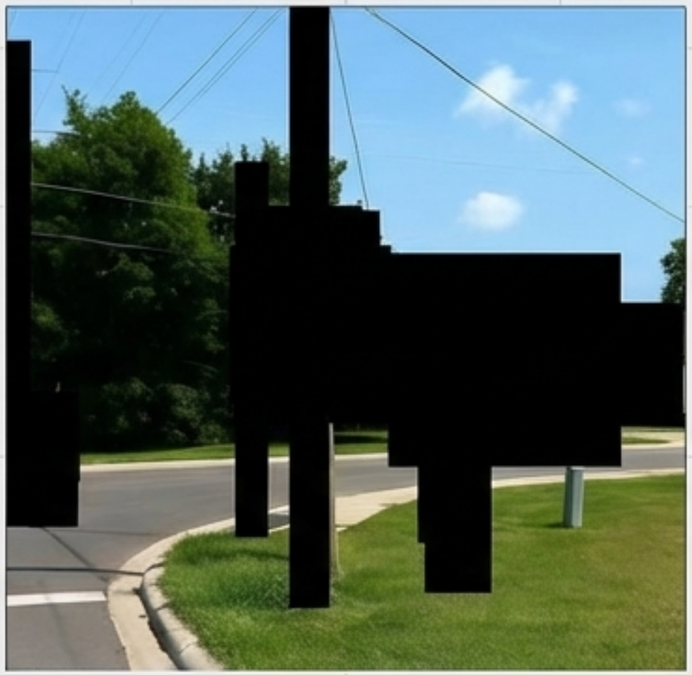
Models exhibit distinct failure profiles over time. While top systems maintain instruction retention, Qwen-Image-Edit leads open-source at Turn 1 but degrades rapidly by Turn 3.

Content Consistency: Separating targeted edits from collateral damage



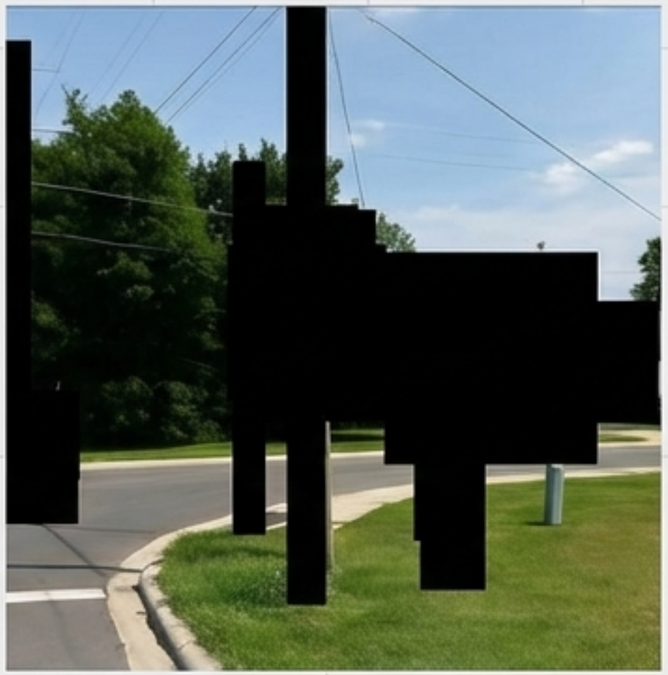
Nano Banana Output

High CC: 98.05



GPT-Image-1 Output

Low CC: 95.19



FLUX.1-Kontext Output

High CC: 97.82

The Test: Extracting unchanged background and object regions using DINOv3 features to measure semantic drift.

The Result: FLUX.1-Kontext and Nano Banana excel at localized edits. GPT-Image-1 exhibits notably poor consistency, allowing irrelevant background details to morph across turns.

Visual Quality: The tension between beautification and preservation

Base Image

(Neutral)

GPT-Image-1

(High Aesthetic, High Drift)

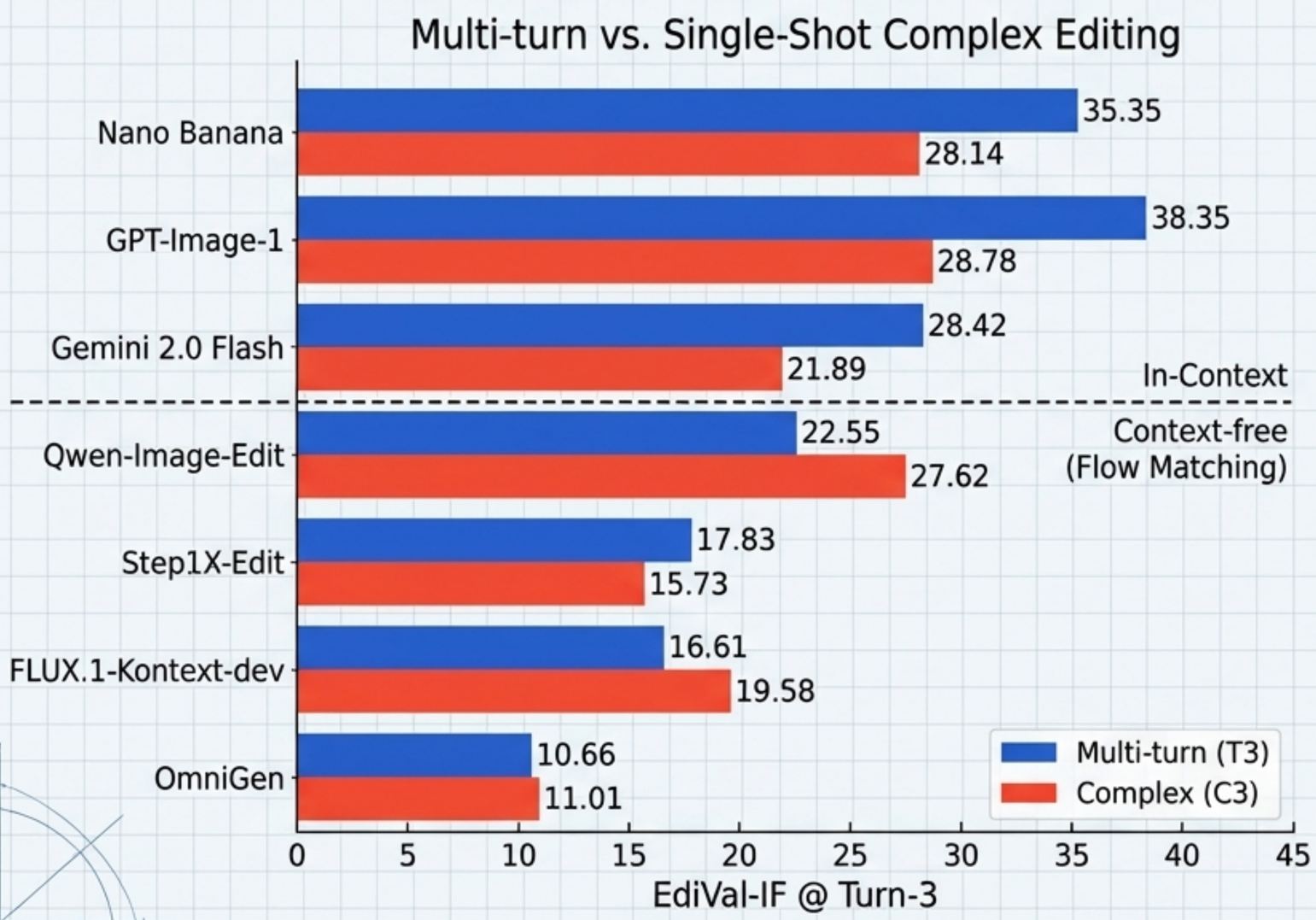
Gemini 2.0 Flash

(Strict Fidelity, Low Drift)

GPT-Image-1 consistently inflates Human Preference Scores (HPSv3) by aggressively beautifying the image, resulting in massive stylistic drift.

Gemini 2.0 Flash and FLUX.1 prioritize strict fidelity to the base image's original exposure and style, accepting lower absolute aesthetic scores.

The Multi-Turn Insight: Exposure Bias in Flow Matching



1. The Exposure Bias Trap:
Single-turn editors are trained on real images. In multi-turn, they are forced to edit their own previous outputs.

2. Small distributional mismatches compound per turn. Compression into a single 'complex prompt' actually saves these models from their own compounding errors.

3. In-context Autoregressive models (like Nano Banana) avoid this trap, successfully utilizing a 'chain of edits' history to maintain stability.

Blueprint for the next generation of image editors

[01]

Stop Relying on VLMs Alone

VLM-only evaluation is a dead end for spatial and subtle edits. Object-centric decomposition (EdiVal) is the required standard.

[02]

Solve the Turn 3 Cliff

Open-source architectures must address exposure bias. Training pipelines must include recursive, self-generated images to survive multi-turn degradation.

[03]

Define the Style Objective

Developers must explicitly choose between ‘Beautification’ (GPT-style) and ‘Strict Fidelity’ (FLUX-style)—current multi-turn metrics heavily penalize models caught in the middle.

[END_OF_REPORT // REPOSITORY: GITHUB.COM/TIANYUCODINGS/EDIVAL]