Google Cloud

PaperBanana: Automating Academic Illustration for AI Scientists

Dawei Zhu, Rui Meng, Yale Song, Xiyu Wei, Sujian Li, Tomas Pfister, Jinsung Yoon

June 04 2026

Motivation

The Missing Visual Communication Step

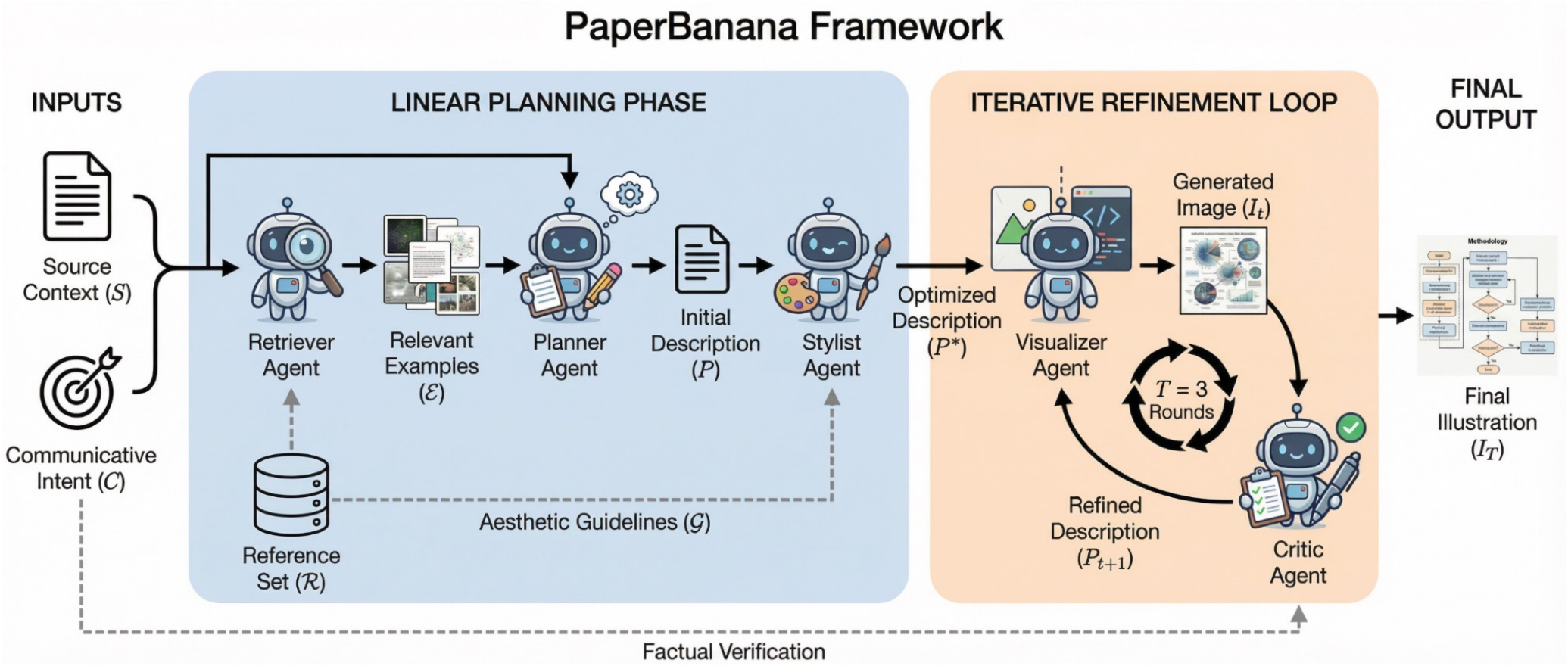
- Today's automated AI Scientists can read papers, generate ideas, and run experiment iterations
- But papers still need clear methodology diagrams and plots
- Making these figures is time-consuming and design-heavy

Automating Academic Figure Generation is Hard

- Requires both faithfulness and aesthetics
- Coding-based methods are editable but limited in expressiveness
- Image models are expressive but often unfaithful or messy

Our Proposal

Intuition: Let's apply an agentic workflow on top of image generation models to bridge the gap



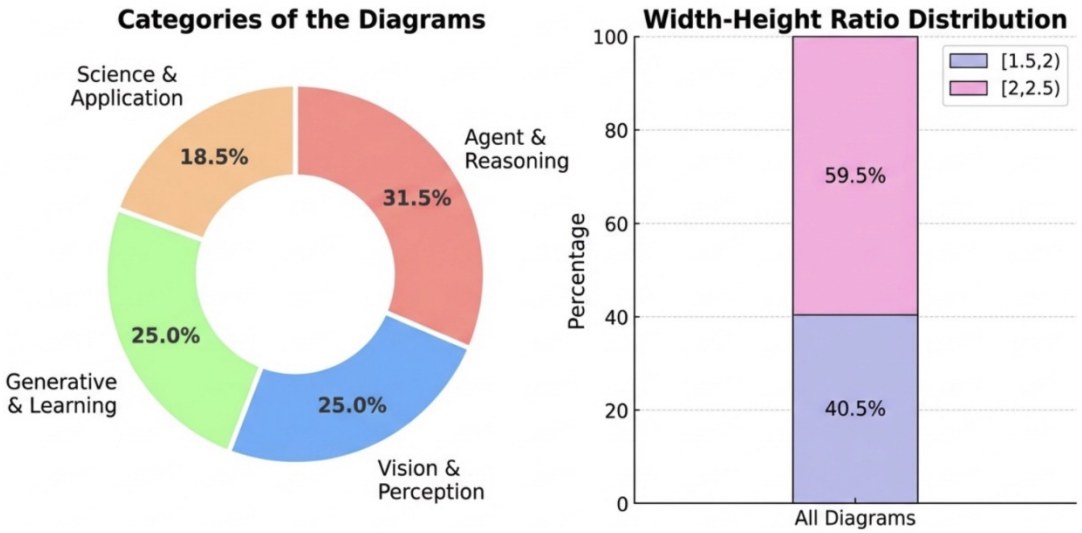
PaperBananaBench

Construction Steps

- Collection & Parsing
- Filtering
- Categorization
- Human Curation

Evaluation Dimensions

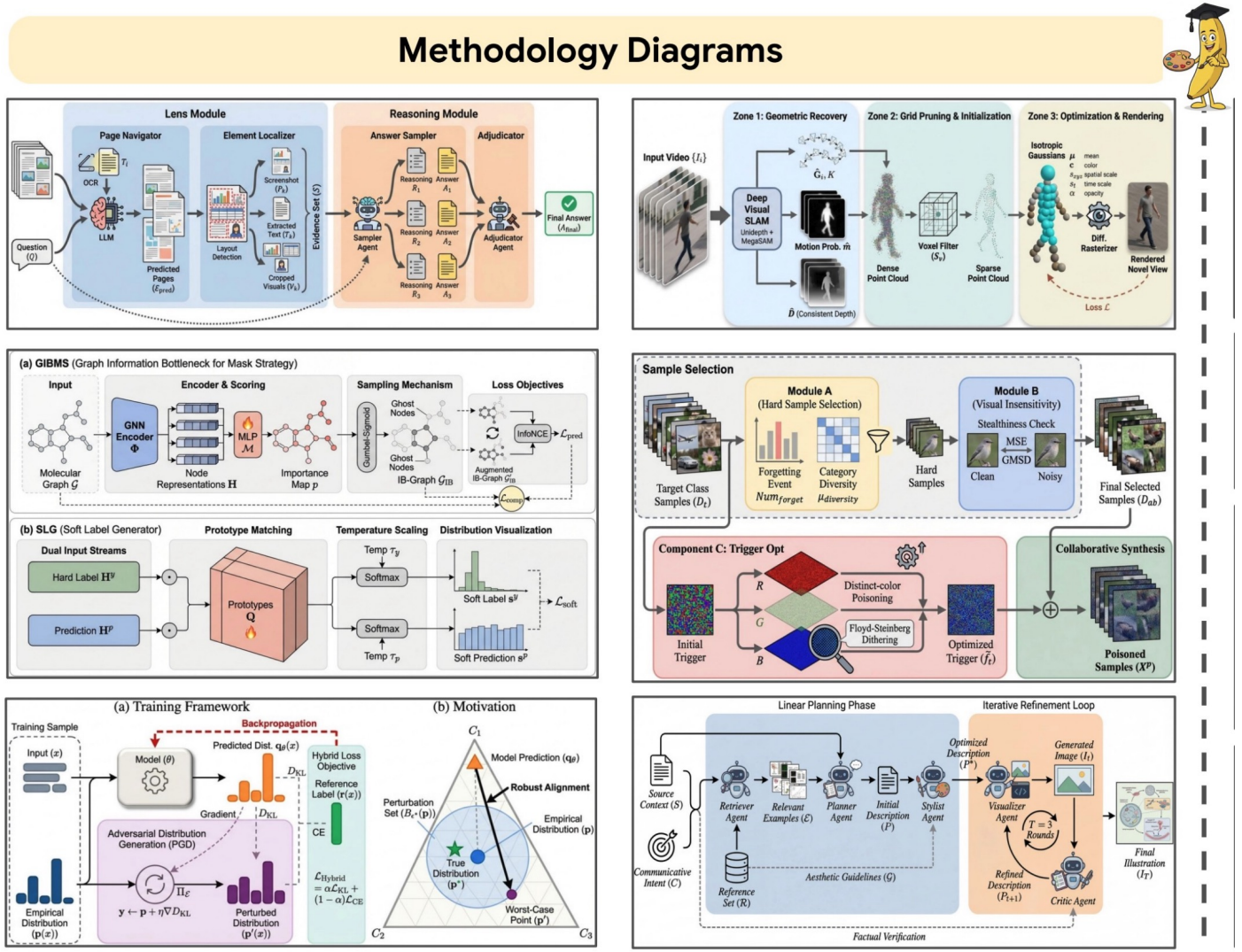
- Faithfulness
- Conciseness
- Readability
- Aesthetics



Main Results

Method	Faithfulness ↑	Conciseness ↑	Readability ↑	Aesthetic ↑	Overall ↑
Vanilla Settings					
GPT-Image-1.5	4.5	37.5	30.0	37.0	11.5
Nano-Banana-Pro	43.0	43.5	38.5	65.5	43.2
Few-shot Nano-Banana-Pro	41.6	49.6	37.6	60.5	41.8
Agentic Frameworks					
Paper2Any (w/ Nano-Banana-Pro)	6.5	44.0	20.5	40.0	8.5
AutoFigure†	37.0	10.2	10.0	14.0	8.3
PAPERBANANA (Ours)					
w/ GPT-Image-1.5	16.0	65.0	33.0	56.0	19.0
w/ Nano-Banana-Pro	45.8	**80.7**	**51.4**	**72.1**	**60.2**
Human	**50.0**	50.0	50.0	50.0	50.0

Some Cases


Methodology Diagrams
Lens Module
Page Navigator
Element Localizer
Reasoning Module
Answer Sampler
Adjudicator
Question
OCR
LLM
Predicted Pages
Layout Detection
Screenshot
Extracted Text
Cropped Visuals
Evidence Set
Sampler Agent
Adjudicator Agent
Final Answer
Input Video
Deep Visual SLAM
Motion Prob.
Dense Point Cloud
Voxel Filter
Sparse Point Cloud
Isotropic Gaussians
Diff. Rasterizer
Rendered Novel View
Loss
(a) GIBMS (Graph Information Bottleneck for Mask Strategy)
Input
Encoder & Scoring
Sampling Mechanism
Loss Objectives
Molecular Graph
GNN Encoder
Node Representations
MLP
Importance Map
Ghost Nodes
Gumbel-Sigmoid
InfoNCE
IB-Graph
(b) SLG (Soft Label Generator)
Prototype Matching
Temperature Scaling
Distribution Visualization
Dual Input Streams
Hard Label
Prediction
Prototypes
Softmax
Soft Label
Soft Prediction
Sample Selection
Module A
(Hard Sample Selection)
Module B
(Visual Insensitivity)
Stealthiness Check
Target Class Samples
Forgetting Event
Category Diversity
Hard Samples
MSE
GMSD
Clean
Noisy
Final Selected Samples
Component C: Trigger Opt
Initial Trigger
Distinct-color Poisoning
Floyd-Steinberg Dithering
Optimized Trigger
Collaborative Synthesis
Poisoned Samples
(a) Training Framework
(b) Motivation
Training Sample
Backpropagation
Input
Model
Predicted Dist.
Hybrid Loss Objective
Reference Label
CE
Gradient
Adversarial Distribution Generation (PGD)
Empirical Distribution
Perturbed Distribution
Model Prediction
Robust Alignment
Perturbation Set
Empirical Distribution
True Distribution
Worst-Case Point
Linear Planning Phase
Iterative Refinement Loop
Source Context
Communicative Intent
Retriever Agent
Relevant Examples
Planner Agent
Initial Description
Stylist Agent
Optimized Description
Visualizer Agent
Generated Image
Reference Set
Aesthetic Guidelines
Refined Description
Critic Agent
Final Illustration
Factual Verification

Ablation Study

Table 2. Ablation study on PAPERBANANABENCH. The shaded row indicates the default setting of PAPERBANANA. We systematically ablate each agent component to assess its contribution. The ◯ symbol denotes the Random Retriever which randomly selects 10 examples instead of performing semantic retrieval.

#	Module					Faithfulness ↑	Conciseness ↑	Readability ↑	Aesthetic ↑	Overall ↑
	Retriever	Planner	Stylist	Visualizer	Critic					
①	✓	✓	✓	✓	3 iters	**45.8**	**80.7**	**51.4**	**72.1**	**60.2**
②	✓	✓	✓	✓	1 iter	38.3	75.2	50.6	68.9	51.8
③	✓	✓	✓	✓	-	30.7	79.2	47.0	72.1	45.6
④	✓	✓	-	✓	-	39.2	61.7	47.9	67.4	49.2
⑤	◯	✓	-	✓	-	37.3	62.7	51.1	65.6	48.3
⑥	-	✓	-	✓	-	41.9	58.6	43.1	62.9	44.2

Statistical Plots Generation

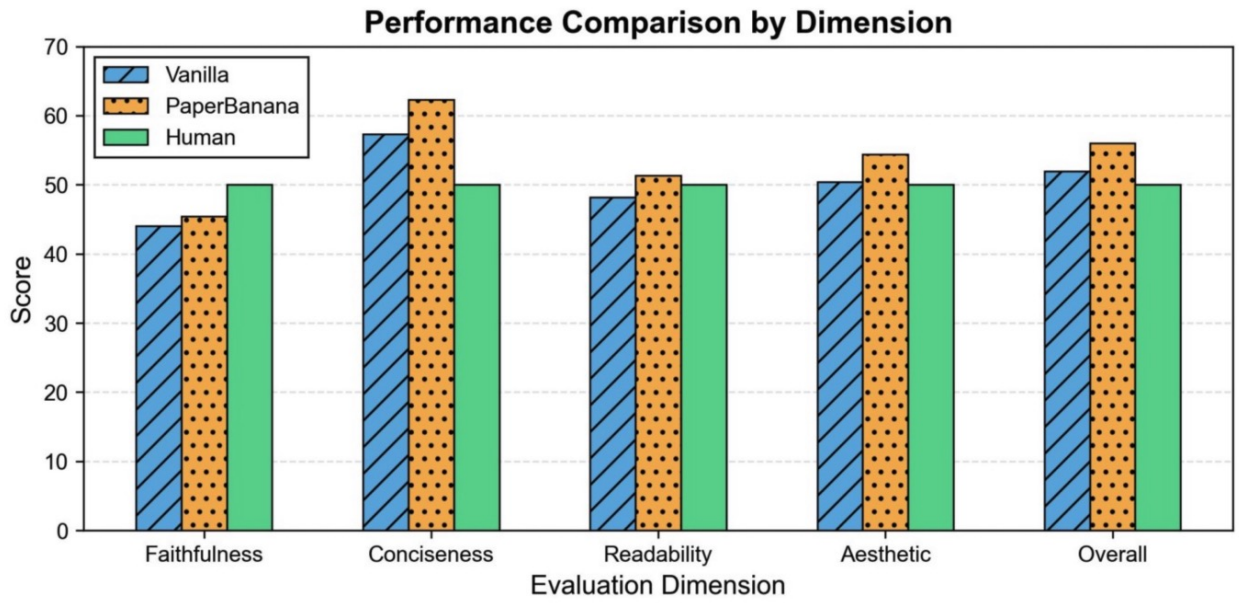
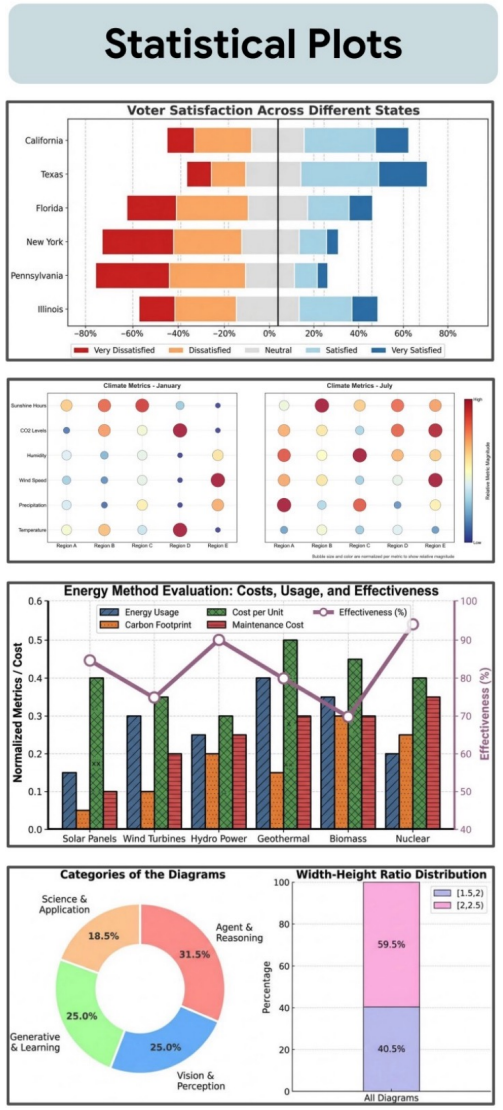
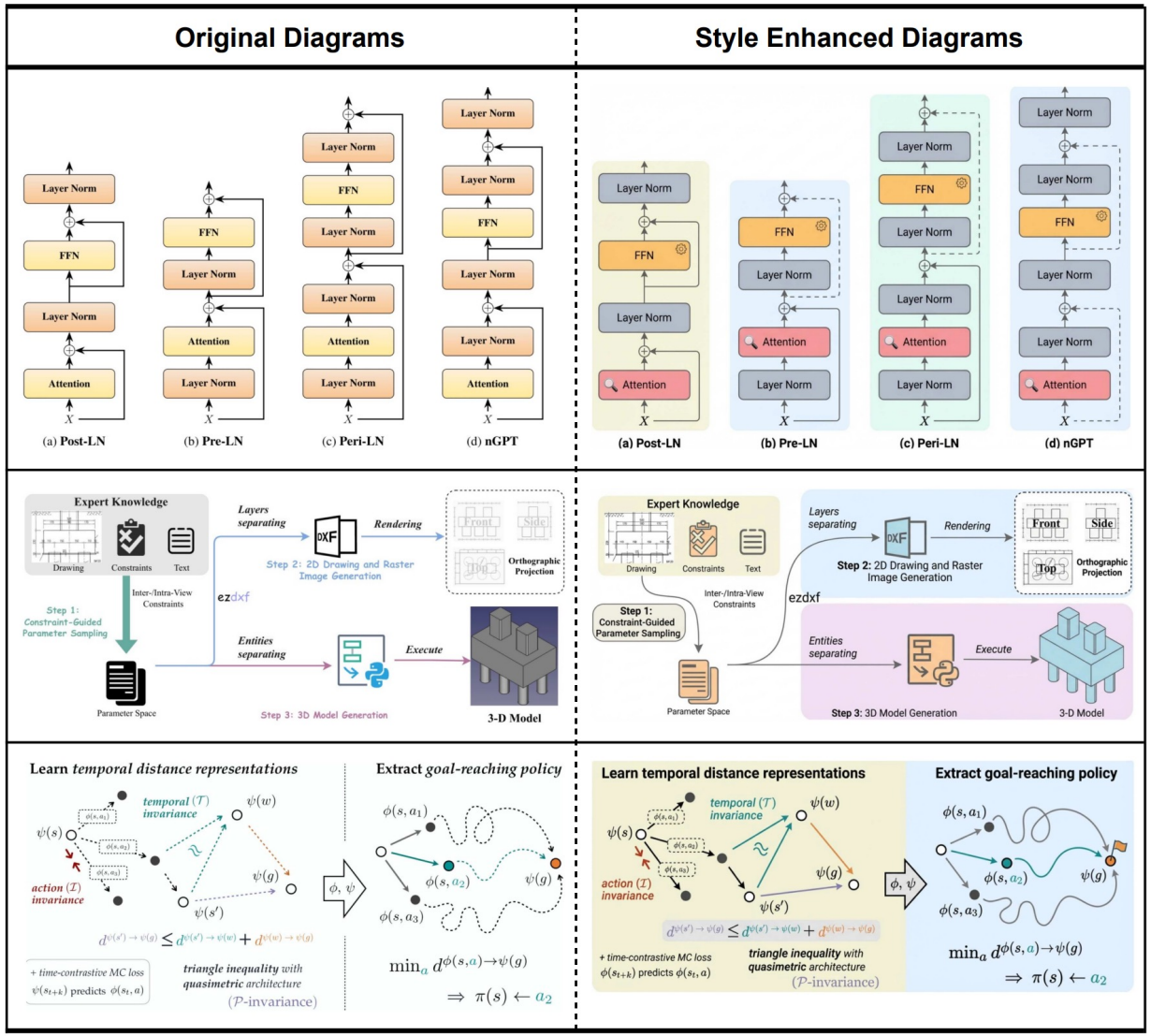


Figure 4. [Generated by 🍌] Vanilla Gemini-3-Pro vs. PAPERBANANA on the statistical plots generation test set. *F*, *C*, *R*, *A* is short for *Faithfulness*, *Conciseness*, *Readability*, and *Aesthetics*, respectively.



Discussion: Enhancing Aesthetics of Human-Drawn Diagrams



Discussion: Code vs Image Generation for Statistical Plots

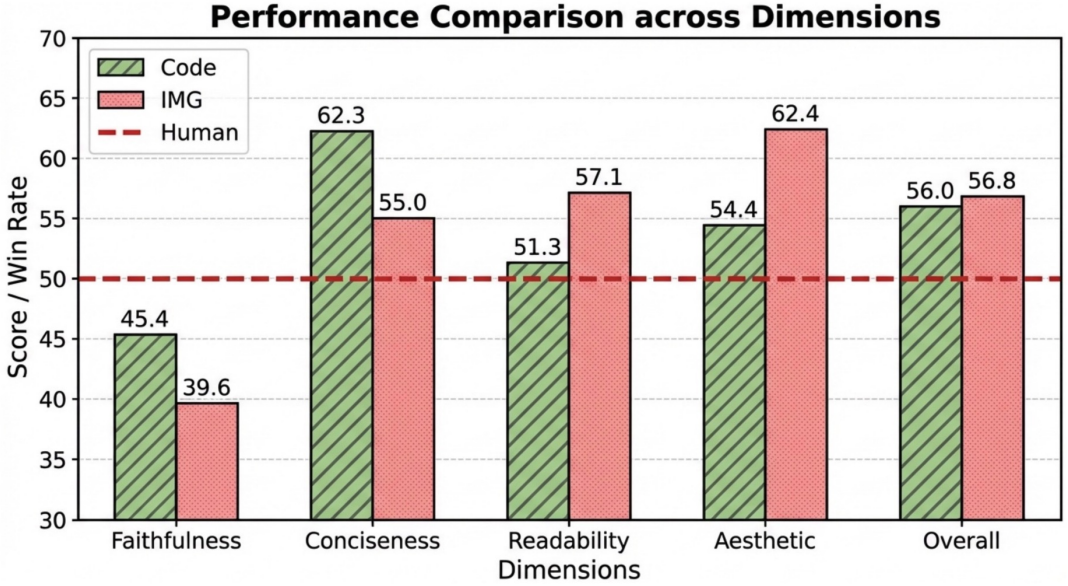
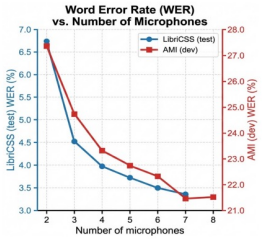
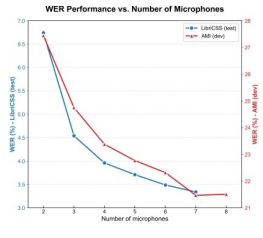
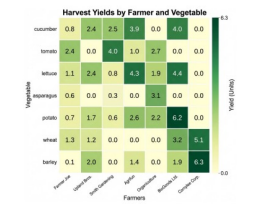
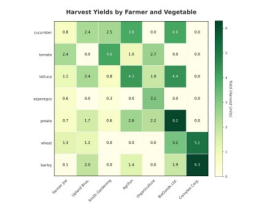
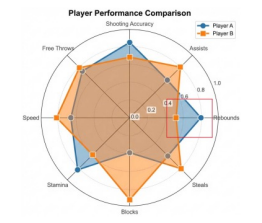
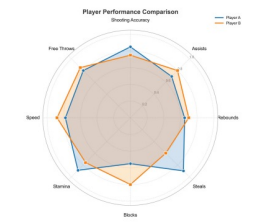
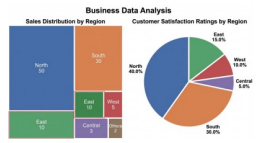
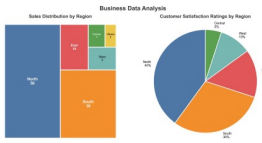
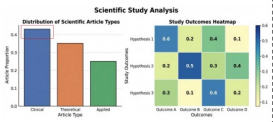
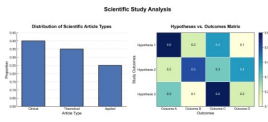
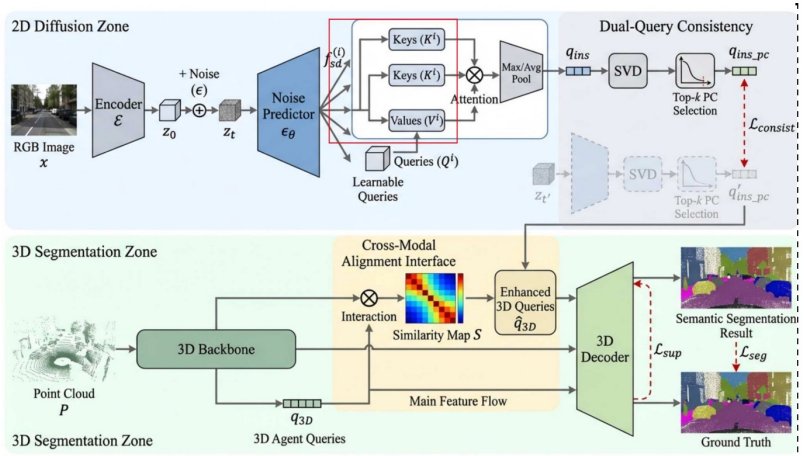
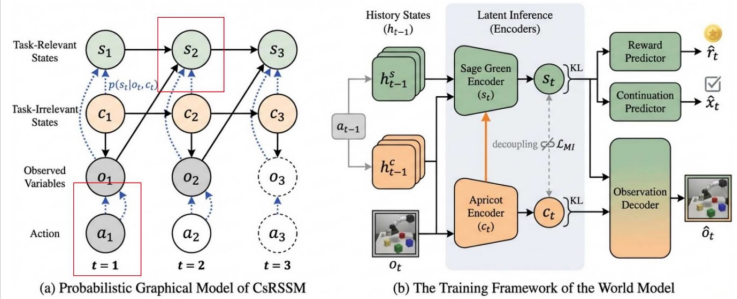
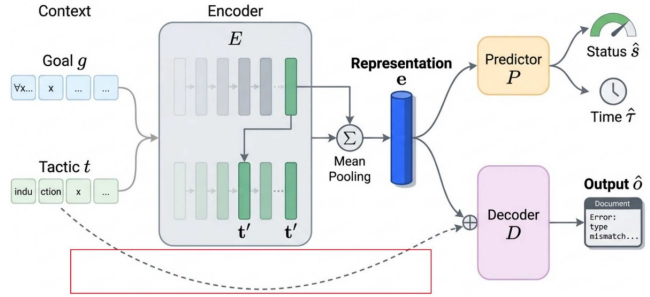


Figure 6. [Generated by 🍌] Coding vs. Image Generation for visualizing statistical plots.

Plots Visualized via IMG	Plots Visualized via Coding	Case Analysis
		Both plots are correct. The left plot looks more visually appealing
		Both plots are correct. The left plot looks more visually appealing
		The left plot looks better, but contains faithfulness issues: On the `Rebound` axis, it depicts Player A (Blue) as having a significantly higher value (~0.9) than Player B (Orange, ~0.6), effectively inverting the relationship and plotting the wrong values (likely duplicating the 'Steals' data).
		The left plot contains faithfulness issues: It duplicates the "East 10" category.
		The left plot contains faithfulness errors: The 'Clinical' data value is 0.4, but the Model draws the bar significantly taller than the 0.4 gridline and axis tick

Failure Cases

Diagrams from PaperBanana	Failure Reasons
	The attention block incorrectly draws connections from the Noise Predictor (f_{sd}) to three destination boxes, labeling two of them as 'Keys (K^i)' and one as 'Values (V^i)'. The text (Eq. 2) explicitly states that diffusion features map only to Keys and Values, while Queries (Q^i) come from learnable parameters.

Diagrams from PaperBanana	Failure Reasons
	It fails to draw generative connections (solid lines) from Actions (a) to Task-Relevant States (s). The method explicitly states (Eq. 3) that state transitions depend on previous actions ($h'_t = f(..., a_{t-1})$). By isolating the actions without driving the state dynamics, the diagram fails to represent the fundamental structure of a control/world model described in the text.
	The text explicitly states, "We note that the Decoder and Predictor can only access information of t through e" and defines the Decoder's input as the embedding and the goal (D(e,g))). However, the Model diagram draws a dotted skip connection from the "Tactic t" block directly to the Decoder, replacing the required "Goal" connection. This violates the functional constraints described in the methodology.

Limitations & Future Directions

- Towards Editable Academic Illustrations
- The Challenge of Fine-Grained Faithfulness
- Advancing Evaluation Paradigms
- Test-Time Scaling for Diverse Preferences
- Extension to Broader Domains

Google Cloud

Thanks!