The Crucible

A Multi-Agent System for procedural content generation. It semantically extracts game concepts using a Vision-Language Model, logically fuses their identities via an LLM reasoning engine, and generates a hybrid artifact using Latent Diffusion.

Concept

- How do we combine two game items to make a new one?

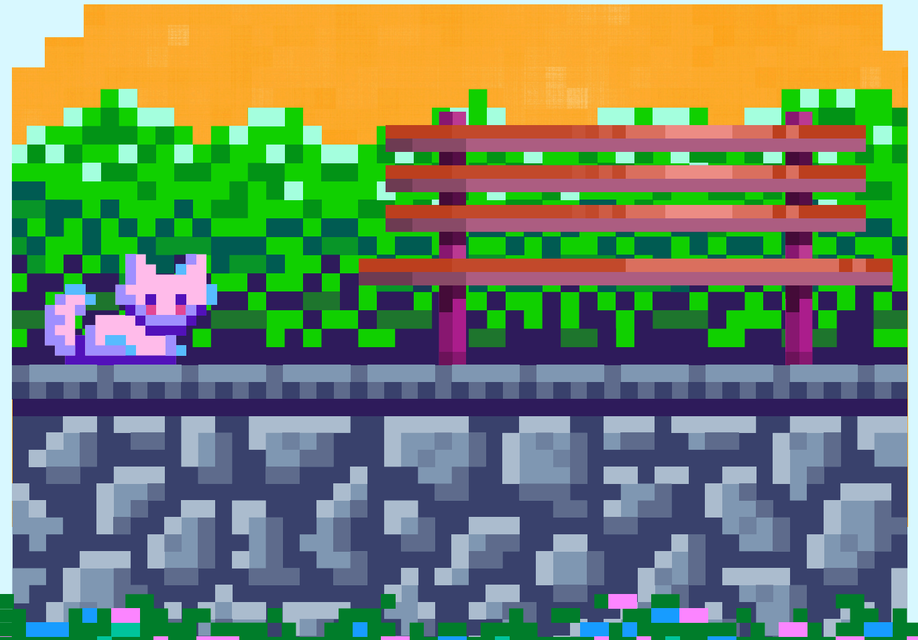
The Bad Way

- If you just mathematically average the pixels of a sword and a green gem, you get a blurry, brown mess.

The Goal

- We want to combine the concepts of the items, not their raw pixels.

WHAT ARE WE ACTUALLY TRYING TO DO?

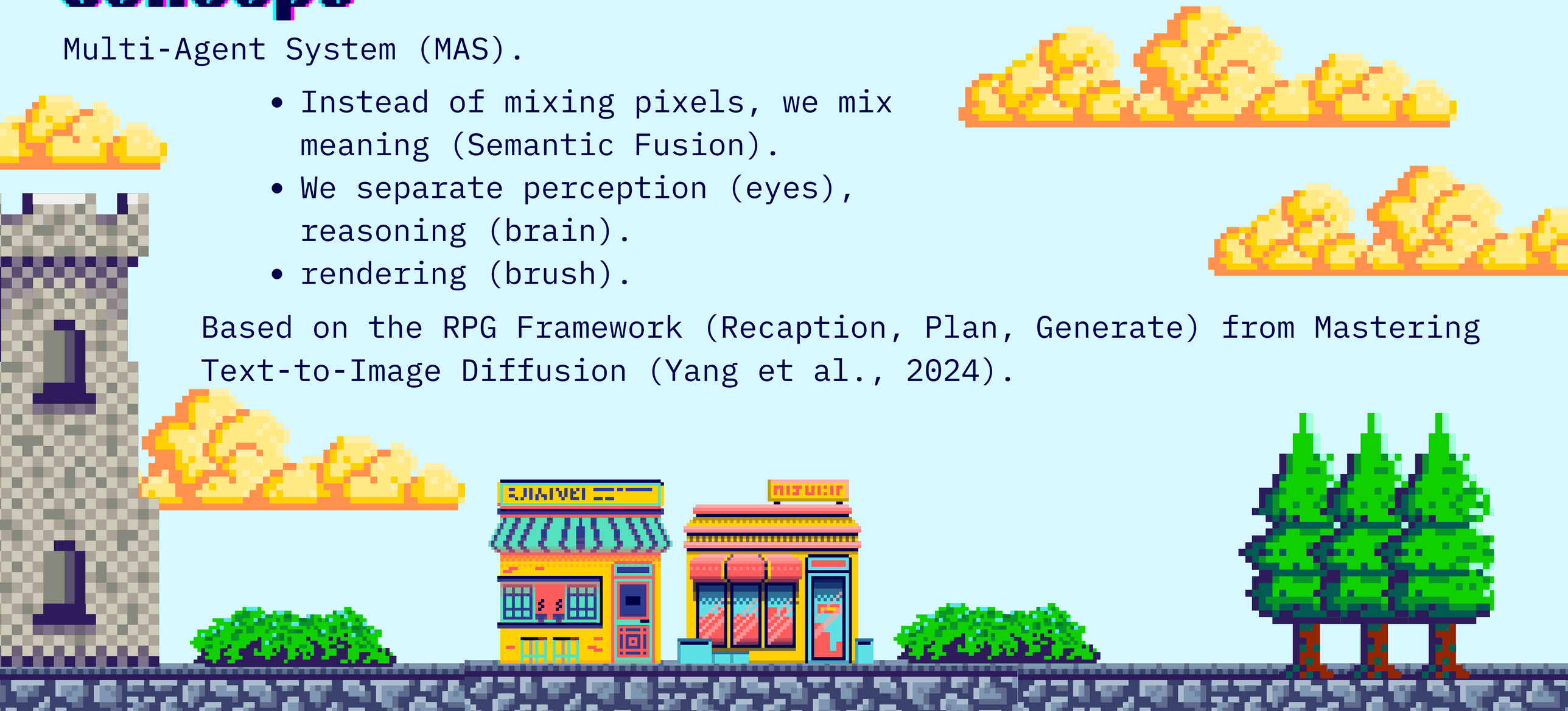


The Solution (Pipeline Decomposition) Concept

Multi-Agent System (MAS).

- Instead of mixing pixels, we mix meaning (Semantic Fusion).
- We separate perception (eyes), reasoning (brain).
- rendering (brush).

Based on the RPG Framework (Recaption, Plan, Generate) from Mastering Text-to-Image Diffusion (Yang et al., 2024).



Pillar 1 – The Appraiser (Phase 1: Reception)

Model
- Moondream2 (1.8B parameters)

Role
- Cross-Modal Encoder-Decoder

What it does:

Extracts semantic meaning from the base items. It analyzes the original 32x32 pixel art and outputs a discrete array of text tags (e.g., ['green', 'gem']).

Reception:
- The system cannot "plan" using raw pixels. Moondream acts as the eyes, translating visual data into a readable text state for the logic engine.

Pillar 2 - The Master Smith (Phase 2: Plan)

Model
- Gemini 3.1 Flash-Lite

Role
- Large Language Model

What it does:

Acts as the deterministic logic gate. It ingests the text tags from the Appraiser and applies Zero-Shot Contextual Reasoning to combine the concepts.

Planner

It acts as the "Global Planner," using logic to figure out how items physically and magically combine. It outputs the final blueprint: JSON lore and a synthesis prompt.

The Display Layer (Phase 3: Generate)

Model
- Flux.1 (via Pollinations API)

Role
- Diffusion Model

What it does:

Takes the text prompt generated by the Master Smith and runs a reverse-diffusion process to render the final pixel art sprite.

How it fits RPG:

The diffusion model doesn't have to think, reason, or plan. It strictly follows the highly detailed blueprint created in Phase 2 to draw the final artifact.

What We Did: Pipeline

- Dataset: Kaggle 16-bit RPG Item Collection.

The Orchestration Flow

- **Recaption (Moondream2)**: Translates the raw pixels of Image A and Image B into a semantic array of text tags.
- **Plan (Gemini 3.1 Flash Lite)**: Ingests the tags to logically reason out a fusion, outputting a precise text blueprint and game lore.
- **Generate (Flux.1)**: Executes a latent diffusion render strictly guided by the LLM's blueprint to manifest the final pixel art.

What We Did: Prompt Engineering

- **Domain Adaptation (Style Prefix):**

  Every generation is strictly wrapped with "pixel art, isolated on white background, 8-bit RPG icon".

- **Purpose**

  - This forces the diffusion model (which naturally wants to generate 4k photos) to behave strictly like a pixel art game engine.

Connor Hamilton

What We Did: Key Parameters

DETERMINISTIC_SEED

Hardcoded in the Flux API. We use a fixed seed number so the exact same item combo will always draw the exact same image (Reproducibility).

PRECISION = float16

Moondream2 was quantized to fit within a local 6GB VRAM constraint so it could run locally on standard computer hardware without crashing.

MAX_NEW_TOKENS = 40

Constrained the LLM generation speed and context length to keep the pipeline lightweight.

The System in Action

Item A

Forged Item

Item B

egis of the Sacred Frost

hen the radiant sanctity of the crusader's shield meets the pristine chill of the ice crystal, their opposing energies do not clash, but coalesce, forging an impenetrable barrier of divine frost. The crystal embeds itself, glowing with an ethereal blue light, hancing the shield's protective aura with chilling purity.

Item A

Forged Item

Item B

Aegis of the Verdant Vitality

By forging the essence of the life-restoring leaf potion into the core of a holy golden shield, I have created a relic that mends the warrior's wounds while they deflect incoming strikes. The metal itself pulses with a legendary green glow, drawing strength from the forest's eternal spirit.

The Output (Semantic Fusion)

SUMMARY

Chaining specialized models (Vision -> Text -> Image) creates much better game assets than trying to use a single AI to do everything at once.

REFERENCES

- Yang et al. (2024). Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. (https://arxiv.org/abs/2401.11708)
- Qin et al. (2024). DiffusionGPT: LLM-Driven Text-to-Image Generation System. (https://arxiv.org/abs/2401.10061)
- Zhai et al. (2023). Sigmoid Loss for Language Image Pre-Training (SigLIP). (https://arxiv.org/abs/2303.15343) - (Reference for the Moondream2 ViT encoder)
- Chen, Y.-C., & Jhala, A. (2025). GameTileNet: A Semantic Dataset for Low-Resolution Game Art in Procedural Content Generation. (https://arxiv.org/abs/2507.02941)
- Kaggle Dataset. 16-bit RPG Item Collection. (https://www.kaggle.com/datasets/ebrahimelgazar/pixel-art)

THANK YOU