Neptune

BY PRAJWAL .
(INNOVATIVE CHALLENGER)

Concept Video: https://youtu.be/cEW4dLxoCrU
Github Repository: https://github.com/Prajwal115/neptune

At A Glance:

- AI adoption in content creation grew rapidly but suffers from acceptance, originality, and quality issues.

- Human creative depth is multi-layered; current AI pipelines try to simulate only one layer of it, hence produce shallow and raw outputs.

- A multimodal, discovery-focused system that integrates sources, analysis, and frame-level inspection improves quality, traceability, and human alignment.

- Prototype features include source harvesting, YouTube/style ingestion, frame-sliced analysis, audio separation and tone analysis, script assistance from multimodal models, and an editor (Drawboard) for consolidated references.

AI in Content Creation:

- Models moved from research to production quickly and began automating text, image, video, and audio tasks.

- Tools reduced friction for producing drafts, variations, and prototypes across media.

- Rapid tooling created volume but not guaranteed quality or human acceptance.

- AI tends to remix common patterns; outputs feel derivative, leading to Originality deficit.

- Shallow processing through single-pass generation yields raw artifacts that lack contextual depth.

- Audiences and creators cannot reliably trace what is synthetic, edited, or sourced.

- Creators and industries resist perceived theft or uncredited training on existing work.

- Discovery of assets like stock videos and credible references remains time-consuming.

- Creators need cross-format adaptation; existing methods force manual conversions.

Some Actions Taken:

- Several creative communities and markets have pushed back on large-model training and content reuse when datasets include their work without consent.

- Notable reactions have focused on animation and illustrative arts, with creators and some regional stakeholders calling for limits or redress when models were trained on copyrighted visual work.

Why AI outputs feel raw

- Humans process content through many implicit and explicit layers: lived experience, cultural context, iterative revision, multi-sensory memory, and value-driven selection.

- Most AI pipelines perform few processing layers: prompt → generate → minimal edit.

- Result: outputs have surface coherence but lack the layered conditioning that gives human work nuance, subtext, and believability.

Some Multimodal Approach Principals:

01 Increase processing depth by chaining analysis, transformation, and human-in-the-loop validation steps.

02 Treat discovery as a first-class multimodal problem: prompts, videos, links, images, audio and human annotations feed a consolidated reference set.

03 Make provenance and explainability visible at every stage.

04 Automate repetitive inspection (frame-slice, tone detection, subtitle checks) so humans focus on curation and interpretation.

The Features -

- **Discovery aggregator:** ingest search queries, prompts, images, videos, YouTube links, and audio; surface metadata and provenance.

- **Drawboard Reference Builder:** consolidated canvas for references with tagging, color-coding, and ordering.

- **YouTube style ingestion**: extract frames, audio, subtitles; produce pacing, palette, shot-length, and energy profiles.

- **Frame packing and analysis:** split video into intervals, pack groups of frames into wide images, run visual OCR/analysis, mapped results to timecodes and displayed.

- **Automated issue reporting:** flag missing subtitles, visual issues, odd objects, audio clipping, level issues, and placement/presentation anomalies with severity and timecodes.

- **Audio pipeline:** separate voice and background, generate transcripts, compute tone metrics, and infer mismatches between audio and visual intent.

- **Script integration:** multimodal model suggests structure, maintains order and clarity, tags scenes, and links Drawboard references.

- **Human review:** accept/reject edits, iterate quickly, and push approved changes to renderer.

Some Points to Work into:

- Frame slicing strategy must balance coverage and compute cost; frames are packed together to reduce API calls while preserving temporal locality.

- Audio separation accuracy varies by recording conditions and will need fallback manual controls.

- Provenance metadata should be immutable and exported with outputs for tracing compliance.

- Model latency and cost: heavy multimodal analysis requires batching and priority queuing for large uploads.

- UI needs to present complex diagnostics succinctly: timecode-linked issue cards and a visual summary strip.

Expected benefits

- Faster discovery and asset selection time.

- Clearer provenance and fewer copyright disputes.

- Higher perceived originality because outputs are filtered, annotated, and curated.

- Better human alignment through tone-visual synchronization and iterative review.

- Reduced manual QA time via automated issue detection.

Some Risks:

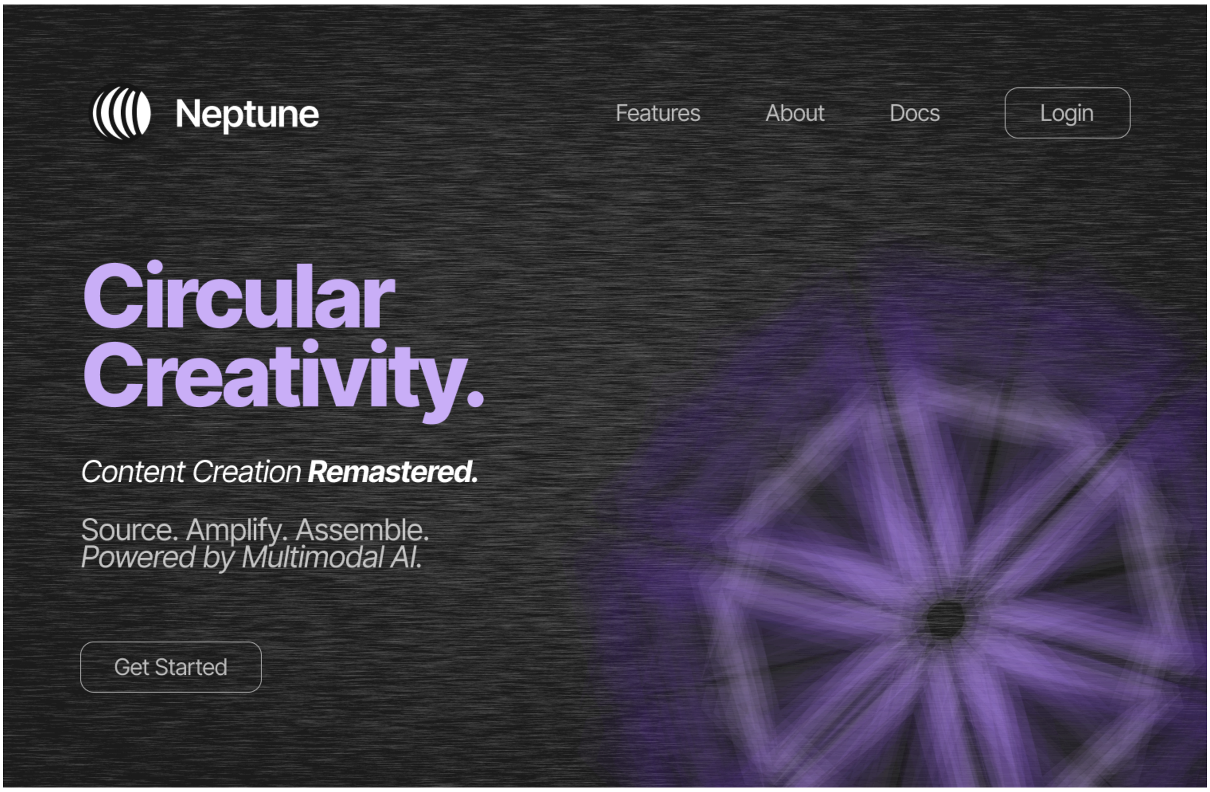
- Detection errors: false positives/negatives in visual and audio analysis.

- Licensing and compliance complexity when surfacing third-party assets.

- Community resistance if models reuse creator content without clear consent.

- Overreliance on automated suggestions can reintroduce shallow outputs unless human curation is enforced.

Current progress:



Circular Creativity.

Content Creation ***Remastered.***

Source. Amplify. Assemble.
Powered by Multimodal AI.

- Figma prototypes are being built for the Drawboard and editor flows.

- Frontend prototypes are being implemented in HTML and iterative UI tests are ongoing.

- Backend work: initial testing of Gemini-like multimodal models has started for style ingestion and analysis.

- API and service layer will be implemented with FastAPI for routing, model orchestration, and lightweight microservices.

- GitHub repository will be populated and updated continuously as tests complete.