

Más agentes, sin mejor resultado

Balance de los ensayos de modelado 3D, revisión independiente y paralelización.

Resultado: los flujos probados no produjeron un personaje aceptable más rápido. Dividir el trabajo tampoco resolvió los problemas de forma, proporción y coherencia visual.

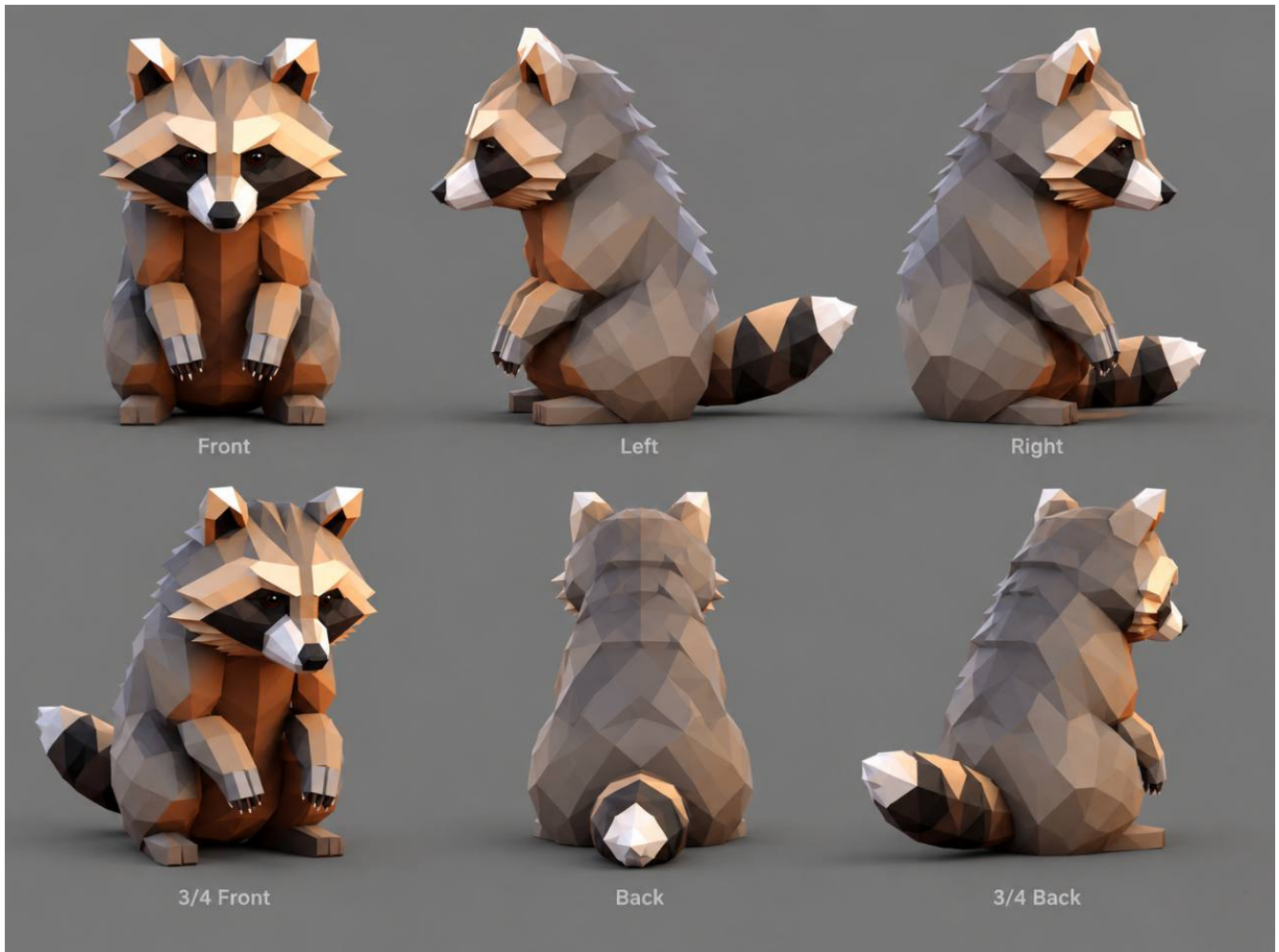


Figura 1. Referencia original aportada por el usuario: seis vistas del mismo mapache. Esta imagen es el objetivo, no un resultado generado durante el ensayo.

El objetivo era obtener un mapache sentado, estilizado y de baja complejidad geométrica. No se pidió rigging, animación ni un pipeline de producción completo. Aun así, el modelo debía preservar anatomía, volumen facial, proporciones y continuidad entre vistas.

El experimento se detuvo por solicitud del usuario. Ninguno de los mapaches debe registrarse como éxito de calidad, aunque algunos jueces hayan emitido una aprobación.

Lo que sí se implementó

La infraestructura funcionó; su funcionamiento no demostraba competencia artística.

Se construyó un MVP con un agente generativo que escribía operaciones de Blender, un selector de acciones acotadas y un crítico visual mediante Perplexity. Un controlador determinista mantenía la política, los permisos, los presupuestos, los umbrales y la evidencia. El constructor no podía modificar la política de aceptación.

La inspección se realizaba sobre revisiones guardadas: renderizado independiente, comprobaciones de geometría, exportación GLB y reimportación. También se incluyeron pruebas y modo simulado para verificar rechazo y reparación sin consumir créditos.

CALIBRACIÓN PREVIA: UN HUMANOIDE DE JUGUETE

La calibración inicial cubría una figura sintética mucho más sencilla que el mapache. Los ejemplos válidos y defectuosos eran fixtures preparados; no constituían una evaluación artística independiente de personajes arbitrarios.

Configuración	Válidos aceptados	Defectuosos aceptados por el sistema
v1: preguntas compuestas	0 de 3	0 de 9; rechazó todo
v2: forma y color separados	3 de 3	0 de 7
v3: primeros planos y auditoría del GLB	3 de 3	0 de 7

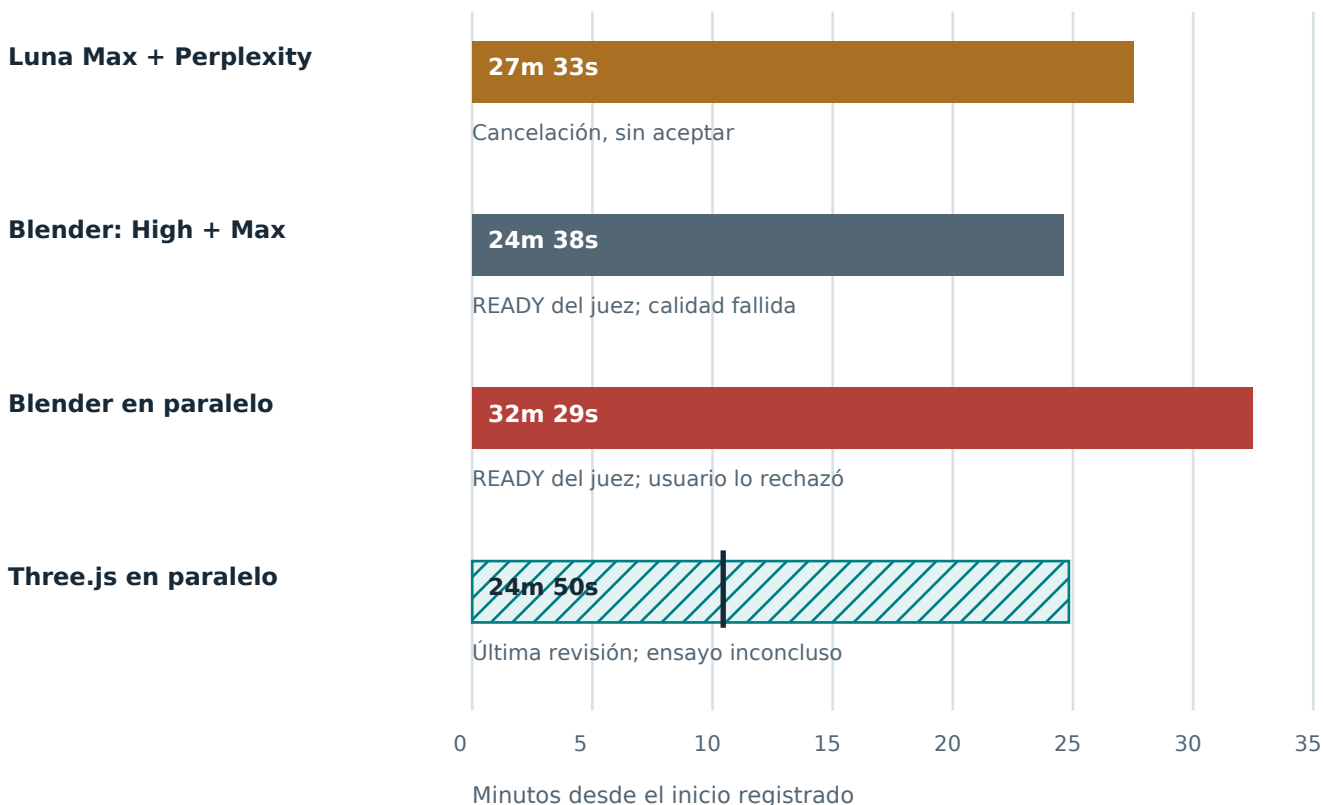
En v2 y v3, el crítico visual por sí solo aprobó algunos defectos. Las mediciones deterministas vetaron esos casos. La mejora de la tabla pertenece al **sistema combinado**, no demuestra que el crítico comprendiera por completo la forma.

Alcance y consumo: las tres calibraciones sumaron 435 solicitudes pagas y 772.229 tokens de entrada. Hubo llamadas adicionales en ensayos en vivo. No se reconciliaron costos monetarios. Los dos ensayos finales del humanoide terminaron en revisión humana requerida.

Fuente: modeling-assistant/CALIBRATION.md. Diez pruebas offline validaron el MVP inicial; la simulación probó el flujo, no la calidad visual del modelo.

El tiempo no mejoró

Tiempos de pared registrados; las barras terminan en eventos diferentes.



Three.js tardó **10m 27s en mostrar la primera vista** (marca negra). La barra rayada llega a la última revisión registrada, a los 24m 50s. No representa un modelo terminado ni incluye el intervalo posterior hasta la solicitud del reporte.

El ensayo paralelo de Blender consumió **7m 51s más, aproximadamente 32%**, que el ensayo anterior. La comparación no fue controlada: incluyó imágenes de partes y construcción desde cero, mientras que el anterior retomaba un modelo existente.

Las operaciones individuales podían ser rápidas: medianas de 0,305 s para selección de acciones y 0,662 s para crítica visual en la calibración v3; una ejecución de ensamblado, exportación y cuatro renders de Blender tomó unos 15 s. El flujo completo consumió decenas de minutos en planificación, preparación, construcción, coordinación, revisiones y reparaciones.

Límite de la comparación: no se midió una ejecución one-shot de referencia. No existe un factor cuantificado contra ella. Lo comprobado es que estos flujos no demostraron la ventaja práctica de velocidad buscada. El primer render de Three.js también incluyó preparación del entorno.

Aprobación no equivale a calidad

Comparación de renders conservados, sin mejorar ni reconstruir las imágenes.

Blender: High + Max



Blender en paralelo



Three.js: primer borrador



READY del juez tras dos revisiones. La aprobación no acreditó fidelidad.



Partes aprobadas y READY final. El usuario rechazó el aspecto y las partes.



Vista preliminar del conjunto. No hubo ensamblado aprobado ni juicio final.

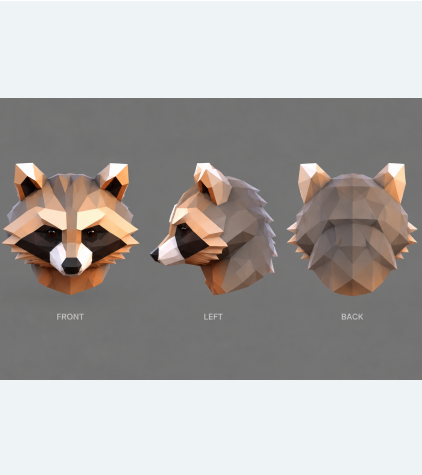
Frentes arriba, perfiles abajo. Los ensayos cambiaron de herramientas y de reparto del trabajo, pero persistieron errores de construcción: caras aplicadas como placas, hocicos desproporcionados, brazos mecánicos, patas poco naturales y diferencias de silueta.

La iluminación no está normalizada entre ensayos. Estas imágenes documentan cada resultado; no son una medición fotométrica ni una comparación controlada de materiales.

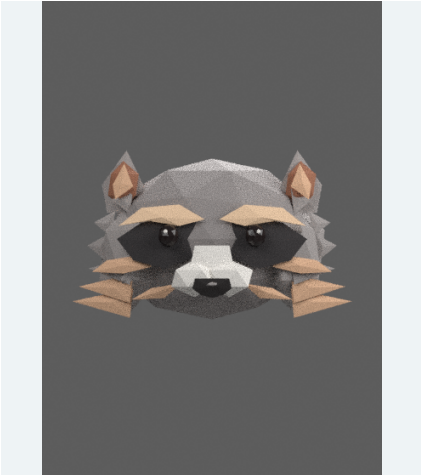
Las partes ya estaban mal

El problema no apareció solamente al juntar los componentes.

Referencia generada de cabeza



Cabeza Blender aprobada



Cabeza Three.js R2 rechazada



La hoja de cabeza de la izquierda fue generada para guiar al constructor; no es la referencia original ni un modelo 3D. Las piezas no tradujeron de forma fiable esas imágenes a volumen. Aparecieron mejillas como láminas, cejas postizas, ojos salientes y perfiles poco integrados.

Three.js al detener el ensayo	Decisión local	Consecuencia
Cabeza, revisión 2	REVISE	Seguía sin cumplir
Cuerpo, revisión 3	APPROVE	No equivale a éxito del personaje
Cola, revisión 1	APPROVE	No habilitó la aceptación final



Cuerpo Three.js: R1 rechazada (izquierda) y R3 aprobada localmente (derecha). Entre ambas hubo una revisión retirada por orientación invertida de caras, que hacía desaparecer superficies. Corregir esos errores permitió mejorar una parte; no produjo un mapache aceptado.

Los jueces más estrictos rechazaron defectos que antes pasaban. Eso mejoró el rechazo, pero no demostró competencia suficiente del constructor ni una mejora de velocidad.

Qué demostraron las pruebas

Conclusión práctica, límites de interpretación y trazabilidad.

HALLAZGOS

- 1. La validez técnica no resolvió la calidad.** Abrir una escena, comprobar vértices, exportar y reimportar un GLB acreditaba integridad del archivo. No acreditaba la anatomía ni la fidelidad del personaje.
- 2. Las aprobaciones anteriores fueron falsos positivos.** Los jueces dieron demasiado peso a rasgos reconocibles y archivos válidos. En el ensayo paralelo, el juez final incluso restó importancia al aspecto lavado por conservarse la paleta en los materiales. La apariencia visible debía seguir siendo un criterio de rechazo.
- 3. Dividir no sustituyó la capacidad.** Cabeza, cuerpo y cola eran encargos más pequeños, pero seguían exigiendo razonamiento espacial. Más agentes añadieron coordinación y revisiones sin eliminar la necesidad de rehacer formas. Las partes fallaron antes de ensamblarse.
- 4. Cambiar herramientas y endurecer prompts no bastó.** Three.js permitió un circuito local y los jueces fueron más exigentes. Hubo mejoras parciales, pero quedó una cabeza rechazada y no se alcanzó un resultado completo aprobado.

INTERPRETACIÓN Y LÍMITES

La hipótesis del usuario es que la comprensión del mundo y de la geometría 3D constituye una habilidad latente más desarrollada en modelos mayores, como Opus 5.5 o GPT-6 Astra. Es compatible con su interpretación de lo observado, pero aquí no se compararon esos modelos ni se investigaron datos o técnicas de entrenamiento. Tampoco se prueba incapacidad universal de todos los modelos pequeños.

El propósito era probar maneras concretas de lograr la tarea, no explicar por qué los modelos fallan. Para este objetivo, los GPT-6 Luna y los flujos ensayados **no cumplieron el estándar de calidad ni demostraron la mejora de velocidad**. La salida correcta es registrar el experimento como fallido, sin rebajar el criterio para obtener una aprobación.

FUENTES DEL EXPERIMENTO

Calibración: modeling-assistant/CALIBRATION.md.

Tiempos: timing.json en raccoon-luna-max, raccoon-normal-vm, raccoon-parallel-test y raccoon-threejs-test.

Veredictos: raccoon-normal-vm/judge/iteration-002; raccoon-parallel-test/final-review/revision-001; raccoon-threejs-test/reviews/{head,body,tail}.

Imágenes: referencia original, referencias de partes generadas y renders guardados de las revisiones identificadas en cada figura.

Rutas relativas a outputs/ dentro de la carpeta del proyecto. Las decisiones READY se conservan como evidencia histórica; este reporte las clasifica según el resultado de calidad. No se ejecutaron nuevos ensayos para elaborar el documento.