Generación multimodal: imagen, audio, vídeo, código

Difusión inversa y casos publicitarios reales.

Conceptos clave

- **Difusión**: la red aprende a deshacer un paso de ruido. En inferencia, parte de ruido puro y repite $T \approx 1\,000$ pasos.
- Stable Diffusion 3, Midjourney v7 (imagen); Sora, Veo 2, Runway Gen-3 (vídeo); Whisper (ASR); ElevenLabs (TTS).
- Riesgos prácticos: derechos de autor, deepfakes, marca corporativa, coste oculto.
- Pipeline canónico: *encarga — revisa — publica* con humano siempre en el último paso.

Ejemplo a mano

Coste de una campaña multimedia para apertura de tienda Mercadona:

Etapa	Modelo	$/pieza	Unidades	Total
Concepto creativo	GPT-5 (texto)	0,005	20	0,10
Storyboard visual	Stable Diffusion 3	0,04	30	1,20
Vídeo 15 s	Veo 2	0,15	5	0,75
Locución es-ES	ElevenLabs	0,03	3	0,09
Música original	Suno	0,02	2	0,04
Total campaña	—	—	—	**2,18 $**

Sin generativa, esta campaña costaría ∼ 4 000 € a una agencia tradicional.

Concepto → aplicación multimodal

Familia	Aplicación	KPI
Imagen	Creatividad publicitaria, mockups.	Tiempo a entrega.
Audio (TTS)	Voz para asistente telefónico.	Naturalidad MOS.
Vídeo	Spots de campaña, demos producto.	Tasa aprobación.
Código	Asistente programación interno.	Aceptación / PR.
ASR (Whisper)	Transcripción reuniones, atención cliente.	WER, tiempo manual.

Idea clave del Día 12. El proyecto AI-first debe incluir una sección de **gobierno de contenido multimedia**: indemnidad de licencia, consentimiento para voces clonadas, alineamiento con marca y control de coste.

Para llevarte: lectura Ho et al., *DDPM*; práctica: storyboard de 4 imágenes y voz off de 20 s con free-tiers (HF + ElevenLabs).