Día 12 · Tema 4 — Difusión y modelos multimodales

Eduardo C. Garrido-Merchán · ecgarrido@comillas.edu Universidad Pontificia Comillas · ICADE

1. Modelos de difusión: la idea base

Un modelo de difusión aprende a *invertir* un proceso de ruido. El proceso *forward* corrompe una imagen x_0 añadiendo ruido gaussiano en T pasos:

$$q(x_t \mid x_{t-1}) \;=\; \mathcal{N}(x_t;\, \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t I), \tag{1}$$

con $\beta_t \in (0,1)$ creciente. Tras T pasos $x_T \sim \mathcal{N}(0, I)$, es ruido puro. El proceso *reverse* entrena una red ϵ_θ para predecir el ruido añadido y restarlo:

$$p_\theta(x_{t-1} \mid x_t) \;=\; \mathcal{N}(x_{t-1};\, \mu_\theta(x_t,t),\, \Sigma_t), \quad \mu_\theta(x_t,t) = \tfrac{1}{\sqrt{\alpha_t}}\Big(x_t - \tfrac{\beta_t}{\sqrt{1-\bar\alpha_t}}\epsilon_\theta(x_t,t)\Big), \tag{2}$$

x_t	imagen ruidosa en el paso t ($t=0$ original, $t=T$ ruido)
β_t	*schedule* de ruido; lineal o cosine
$\alpha_t = 1-\beta_t,\ \bar\alpha_t = \prod_{s\le t}\alpha_s$	factores acumulados
$\epsilon_\theta(x_t, t)$	red (U-Net o DiT) que predice el ruido

El entrenamiento minimiza $\mathbb{E}_{t,x_0,\epsilon}\big[\,\|\epsilon - \epsilon_\theta(x_t,t)\|_2^2\,\big]$.

2. Difusión condicionada y guidance

Para condicionar la generación con texto se entrena $\epsilon_\theta(x_t, t, c)$ con c el embedding del prompt. **Classifier-free guidance** interpola entre la predicción condicionada y la no condicionada:

$$\hat\epsilon \;=\; \epsilon_\theta(x_t,t,\varnothing) \;+\; w\,\big(\epsilon_\theta(x_t,t,c) - \epsilon_\theta(x_t,t,\varnothing)\big), \tag{3}$$

con $w > 1$ (típicamente 7,5 para Stable Diffusion). Más $w \Rightarrow$ mayor adherencia al prompt, menos diversidad.

3. Cinco familias multimodales

- **Texto → texto**: GPT-5, Claude 4.7, Gemini 2.5, Llama 4. Pago por token (ver Día 11).
- **Texto → imagen**: SD 3, Flux, DALL·E 4, Midjourney 7. Coste ~0.01–0.04 $ por imagen.
- **Texto → audio**: ElevenLabs, Suno, OpenAI Voice. Coste ~0.05–0.30 $/min.
- **Texto → vídeo**: Sora 2, Veo 3, Runway Gen-4, Kling. Coste ~0.50–2.00 $ por clip de 5–10 s.
- **Imagen → texto / VQA**: GPT-5V, Claude Vision, Gemini Vision. Coste por imagen *tokenizada* (orden de ~1k tokens por imagen).

4. Coste de una campaña multimedia (ejemplo)

Mercadona quiere 1 000 creatividades para campaña de Navidad: imagen, copy y voz. Estimación con precios de 2026:

Activo	Unidades	Coste estimado (€)
Imágenes (Flux Pro)	1 000 × 0.04 $	40
Copy (Claude Sonnet, 300 tok)	1 000 × 0.005 $	5
Voz (ElevenLabs, 10 s)	1 000 × 0.04 $	40
Total IA		~ 85 €
Equivalente agencia		~50 000 €

Idea clave. El factor ~500× no incluye el coste humano de selección, control de marca y revisión legal. Pero incluso multiplicando por 20, el ahorro es real y mide *escala*, no *calidad puntual*.

5. Riesgos del multimodal generativo

Deepfakes: la voz puede clonarse con 10 segundos de audio (ElevenLabs, OpenAI Voice Engine). **Atribución**: no hay garantía de que el modelo no haya memorizado una imagen con copyright; *watermarking* (Stable Signature, SynthID) es una mitigación parcial. **AI Act**: imágenes generadas por IA deben estar etiquetadas en su metadata si entran en el mercado europeo.

6. Concepto → aplicación

- Marca-creatividad masiva → Flux/SD3 + Claude para copy.
- Voz corporativa segura → ElevenLabs con consentimiento firmado del locutor; nunca clonar voces sin contrato.
- Vídeo de producto → Veo/Sora para borradores; rodaje humano para la pieza final.
- Análisis de catálogo → Gemini Vision para tagging masivo.