

Deep Learning para Business Analytics

Día 12 · Tema 4 — Imagen, audio, vídeo, código

Difusión, generación y casos publicitarios.

Eduardo C. Garrido-Merchán

Profesor Propio Adjunto, Métodos Cuantitativos, ICADE

`ecgarrido@comillas.edu`



Madrid, Día 12

amazon

NETFLIX

Spotify

AMCAREX

OpenAI



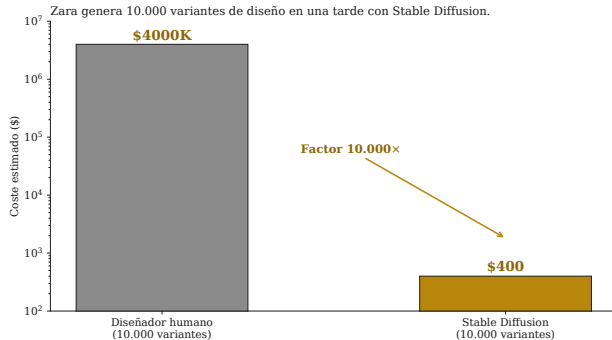
ejemplos del día

Lo que vamos a cubrir hoy.

1. Modelos de difusión
2. Cinco familias multimodales
3. Casos reales y coste

Al final del día: sabes calcular el coste de una campaña multimedia y elegir entre cinco familias de generadores según el caso.

Zara genera 10.000 variantes de diseño en una tarde .



Antes. Un diseñador tarda 2 días por variante. 10.000 variantes = 20.000 días = imposible sin un ejército creativo.

Ahora . \$0,04/imagen $\times$ 10.000 = \$400. En una tarde, Stable Diffusion genera todo el catálogo de prueba.

El humano selecciona, no dibuja. El diseñador pasa de productor a curador: elige las mejores variantes, ajusta prompts, aprueba la línea final.

IDEA CLAVE. La generación masiva de variantes elimina el cuello de botella creativo en fast fashion.

Parte 1 / 3

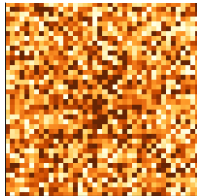
Modelos de difusión



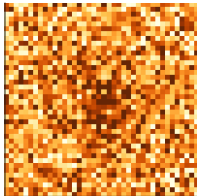
Difusión inversa 🎨: ruido $\rightarrow$ imagen, paso a paso.

Difusión inversa: el modelo aprende a quitar ruido paso a paso.

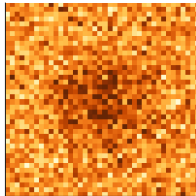
$t=5$
 $\sigma = 1.0$



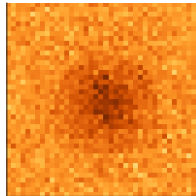
$t=4$
 $\sigma = 0.8$



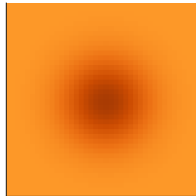
$t=3$
 $\sigma = 0.5$



$t=2$
 $\sigma = 0.2$



$t=1$
 $\sigma = 0.0$



Difusión en una frase y dos imágenes 🎨.

Forward. A una imagen 📷 real se le añade ruido gaussiano hasta convertirla en ruido puro tras $T \approx 1\,000$ pasos.

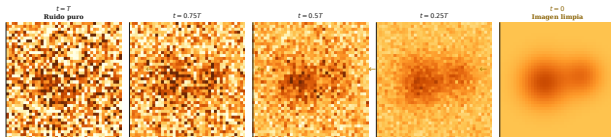
Backward. La red aprende a *deshacer* un paso de ruido. En inferencia, partes de ruido puro y aplicas la red T veces hasta obtener una imagen nueva.

Condicionar. El modelo recibe además un *prompt* 💬 (texto) que guía qué quiere generar. Esa es la magia de Stable Diffusion 3 y Midjourney v7.

IDEA CLAVE. La difusión es una RNN inversa sobre ruido, no un modelo autorregresivo.

Difusión = aprender a limpiar una foto borrosa, paso a paso .

Difusión = aprender a limpiar una foto borrosa, paso a paso.



Clave. El modelo nunca aprende a generar. Aprende a QUITAR RUIDO. Si sabes quitar ruido, puedes partir de ruido puro y llegar a una imagen limpia.

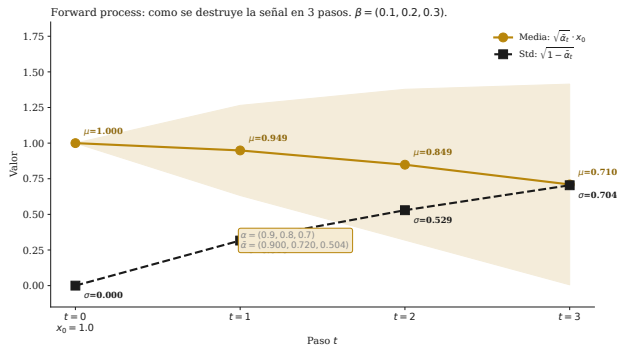
Forward (entrenamiento). Añade ruido gaussiano progresivamente hasta destruir la imagen. Esto es fácil y no necesita red neuronal.

Backward (generación). La red aprende a deshacer un paso de ruido. En inferencia, aplicas la red T veces desde ruido puro.

IDEA CLAVE. Generar una imagen = ejecutar T veces un limpiador de ruido aprendido.



Forward process: cómo se destruye una señal en 3 pasos.



Parámetros. $\beta = (0.1, 0.2, 0.3)$, $\alpha_t = 1 - \beta_t$,

$$\tilde{\alpha}_t = \prod_{s=1}^t \alpha_s.$$

Cálculo. $\alpha = (0.9, 0.8, 0.7)$. $\bar{\alpha} = (0.900, 0.720, 0.504)$.

Media de x_t . $\mathbb{E}[x_t] = \sqrt{\tilde{\alpha}_t} \cdot x_0$. Con $x_0=1,0$:
(0,949, 0,849, 0,710).

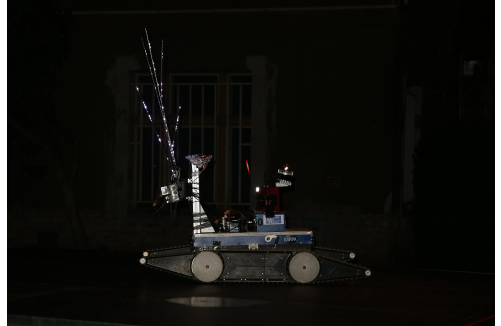
Desviación. $\text{std}(x_t) = \sqrt{1 - \tilde{\alpha}_t}$: (0,316, 0,529, 0,704).

Resultado. En 3 pasos, la señal ya es más ruido que información: $\text{std} = 0,704 \approx \text{media} = 0,710$.

IDEA CLAVE. El schedule β_t controla lo rápido que se destruye la señal. Elegirlo bien es crucial para la calidad.

Parte 2 / 3

Cinco familias multimodales



Cinco familias y sus modelos punteros en 2026.

Cinco familias multimodales y sus modelos punteros en 2026.

Imagen

Stable Diffusion 3,
Midjourney v7

Audio

Whisper (ASR),
Suno,
ElevenLabs

Vídeo

Sora, Veo 2,
Runway Gen-3

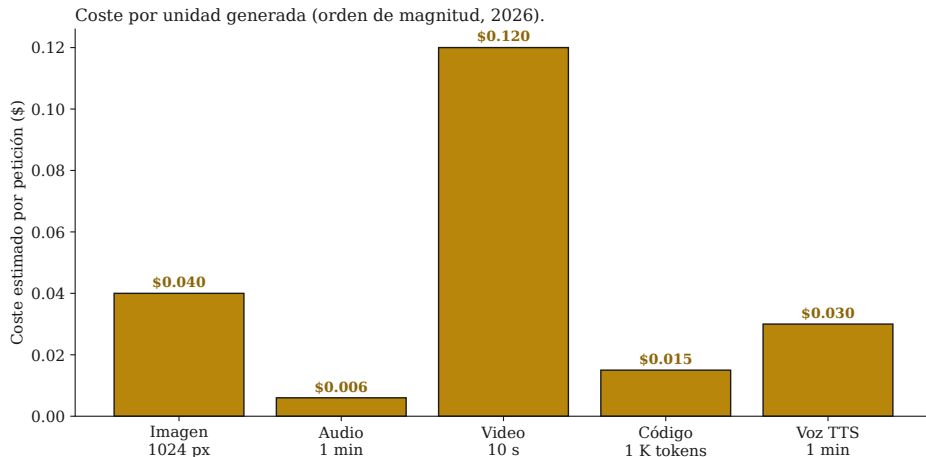
Código

GPT-5,
Claude Sonnet,
DeepSeek-Coder

Voz/TTS

ElevenLabs,
OpenAI TTS,
Google WaveNet

Coste 💰 por unidad generada (orden de magnitud).



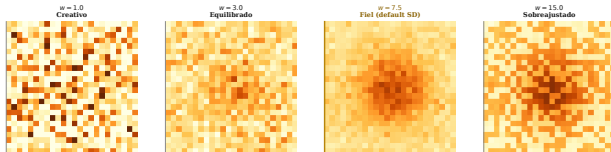
Concepto → aplicación: qué modelo y para qué caso.

Familia	Aplicación de negocio	Entradas / Salidas	KPI
Imagen 	Creatividad publicitaria, mockups de producto.	Texto/imagen → imagen.	Tiempo a entrega.
Audio 	Voz para asistente telefónico (Telefónica).	Texto → audio (wav).	Naturalidad MOS.
Vídeo 	Spots de campaña, demos producto.	Texto → vídeo (10–60 s).	Tasa de aprobación.
Código 	Asistente de programación interno (BBVA).	Lenguaje natural → código.	Aceptación / PR.
ASR (Whisper) 	Transcripción de reuniones, atención cliente.	Audio → texto.	WER, tiempo manual.

IDEA CLAVE. La misma pipeline de “encarga – revisa – publica” sirve para los cinco casos.

🎯 Classifier-free guidance: el dial de creatividad vs fidelidad 🎨.

Classifier-free guidance: el dial de creatividad vs fidelidad.



$\hat{z} = \varepsilon_\theta(x_t, t, \emptyset) + w \cdot (\varepsilon_\theta(x_t, t, c) - \varepsilon_\theta(x_t, t, \emptyset))$ $w = 1$: estándar. $w > 1$: amplifica texto. Default SD: $w = 7.5$.

Ecuación.

$$\hat{z} = \varepsilon_\theta(x_t, t, \emptyset) + w \cdot (\varepsilon_\theta(x_t, t, c) - \varepsilon_\theta(x_t, t, \emptyset))$$

$w = 1$: **estándar**. Genera con la distribución aprendida. Máxima diversidad, adherencia moderada al texto.

$w = 7.5$: **default SD**. Amplifica la señal del texto. El equilibrio habitual entre calidad y fidelidad al prompt.

$w > 10$: **sobreajuste**. Imágenes sobresaturadas con artefactos. Los colores se queman y los detalles se pierden.

En producción. $w = 5-8$ para publicidad. $w = 2-3$ para arte libre. Nunca $w > 15$ salvo para depuración.

IDEA CLAVE. CFG es a difusión lo que la temperatura es a LLMs: un dial continuo entre variedad y control.

Parte 3 / 3

Casos reales y coste



Ejemplo a mano: coste 💰 de una campaña 🛒 para Mercadona.

Ejemplo a mano: coste de generar una campaña multimedia para una apertura de Mercadona.

Etapas	Modelo	Coste/pieza	Unidades	Total
Concepto creativo	GPT-5 (texto)	\$0,005	20 ideas	\$0,10
Storyboard visual	Stable Diffusion 3	\$0,04	30 imágenes	\$1,20
Vídeo de 15 s	Veo 2	\$0,15	5 versiones	\$0,75
Locución (es-ES)	ElevenLabs	\$0,03	3 voces	\$0,09
Música original	Suno	\$0,02	2 tracks	\$0,04
Tienda nueva en Marbella	—	—	—	\$2,18

Riesgos ⚠ del multimodal generativo.

Derechos de autor. Modelos entrenados con datos masivos pueden reproducir estilos protegidos. Para uso comercial, exige modelos con *indemnity* contractual o entrenados con datasets limpios.

Deepfakes y voz clonada ⚠. Tecnología accesible para suplantación. La política interna debe exigir consentimiento explícito antes de clonar una voz.

Marca corporativa. Modelos genéricos pueden contradecir el manual de identidad. Cuando importa, fine-tunar el modelo con la guía de marca (Tema 6).

Coste oculto 💰. Las imágenes y vídeos consumen mucho cómputo. Una campaña con 100 vídeos puede llegar a 50 \$ fácilmente.

IDEA CLAVE. El proyecto AI-first debe incluir una sección de gobierno de contenido multimedia.

Lo que el alumno se lleva a casa de este día.

Concepto 1. La difusión inversa 🎨 es el modelo dominante para imagen, audio y vídeo en 2026.

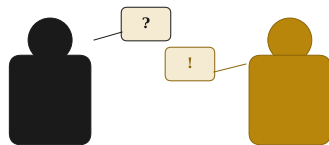
Concepto 2. Las cinco familias multimodales se eligen por caso de negocio, no por entusiasmo técnico.

Concepto 3. El coste 💰 se calcula por unidad (imagen, segundo de audio, token) y se acumula rápidamente en campañas grandes.

Para la práctica. Generar un mini-storyboard publicitario (4 imágenes + voz off) con APIs gratuitas o servicios free-tier.

Próximo día. Día 13 (Tema 5.1): agentes con Claude Code, el cambio de paradigma.

Pausa para debatir: ¿Voz clonada sin consentimiento: prohibición o regulación?



La pregunta. Una IA clona la voz de un actor para un anuncio sin autorización explícita. ¿Prohibición total o regulación con consentimiento?

Posición A. Prohibición: la voz es propiedad inalienable.

Posición B. Regulación: permitido con consentimiento documentado y trazabilidad.

Reglas. 5 minutos, dos turnos de palabra por bando, móviles boca abajo. Yo modero, no participo.

Hasta el Día 13.

Tu primer agente con Claude Code.

Eduardo C. Garrido-Merchán

ecgarrido@comillas.edu

The Gemini logo is displayed on a light beige rounded square background. It features the word "Gemini" in a sans-serif font, with the "G" in blue and the remaining letters in a gradient of purple to pink. A small, four-pointed star icon is positioned above the "i".