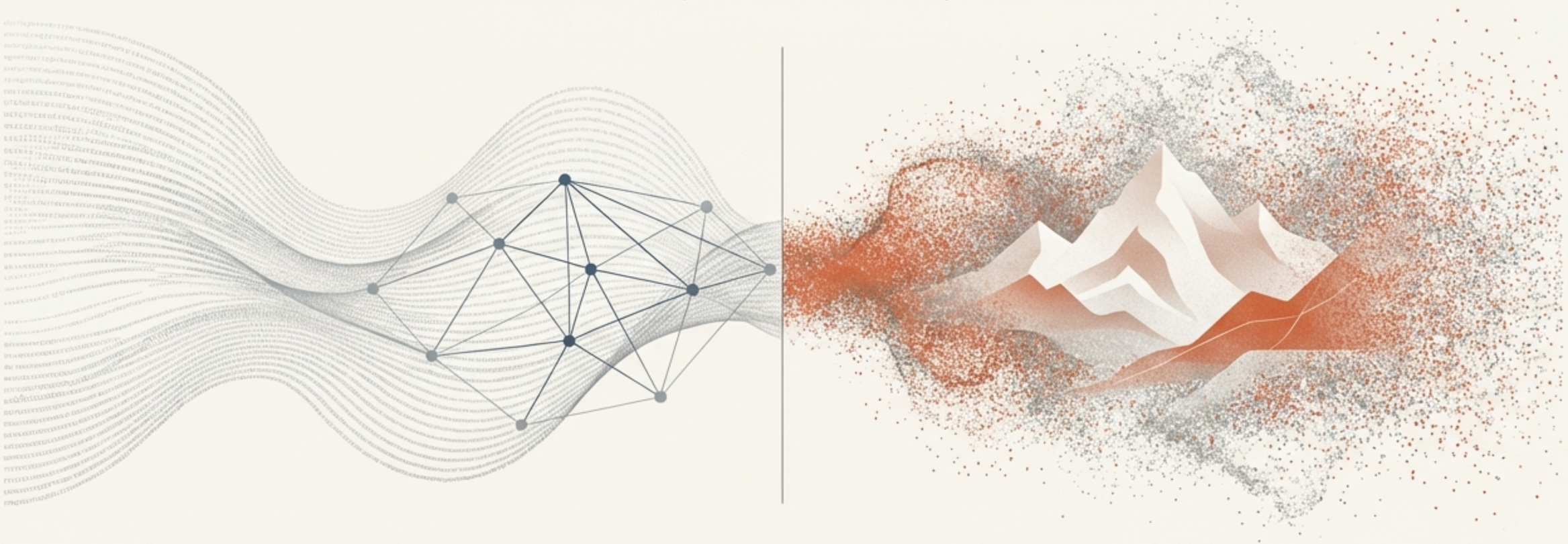
Unha historia de dúas IAs: Dos mundos das palabras aos mundos dos píxeles

Como funcionan realmente as IAs que xeran imaxes, explicado a través da súa comparación coas IAs que xeran texto.

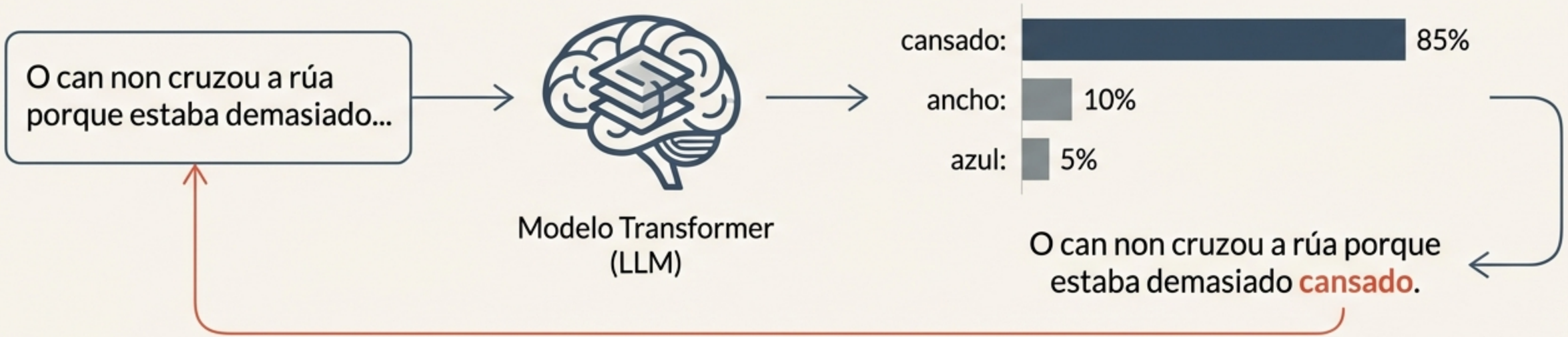


A IA xerativa ten dúas modalidades principais: a que **fala** (Modelos de Linguaxe ou LLMs, como ChatGPT) e a que **ve** (Modelos de Difusión, como Midjourney ou Stable Diffusion).

Aínda que ambas parecen máxicas, operan baixo principios radicalmente distintos.

Para entender como se '**debuxa**' unha imaxe a partir de texto, primeiro debemos entender como se '**escribe**' a seguinte palabra.

O Reino das Palabras: A lóxica secuencial dos LLMs



Como funciona?

Os LLMs son fundamentalmente motores de autocompletar avanzados. Descompoñen o texto en pezas discretas chamadas **tokens** (palabras, anacos de palabras, ou mesmo emojis).

O Proceso

Operan de forma **autorregresiva**: predín o seguinte token máis probable baseándose na secuencia de tokens que o preceden.

O Mecanismo Clave

Utilizan a **Auto-Atención (*Self-Attention*)** para entender o contexto interno. No exemplo, 'cansado' refírese a 'can', non a 'rúa'. A Auto-Atención constrúe esta coherencia lingüística.

En Resumo

É un proceso secuencial, probabilístico e que opera sobre un vocabulario discreto e finito.

O Reto dos Píxeles: Por que xerar imaxes é un problema diferente

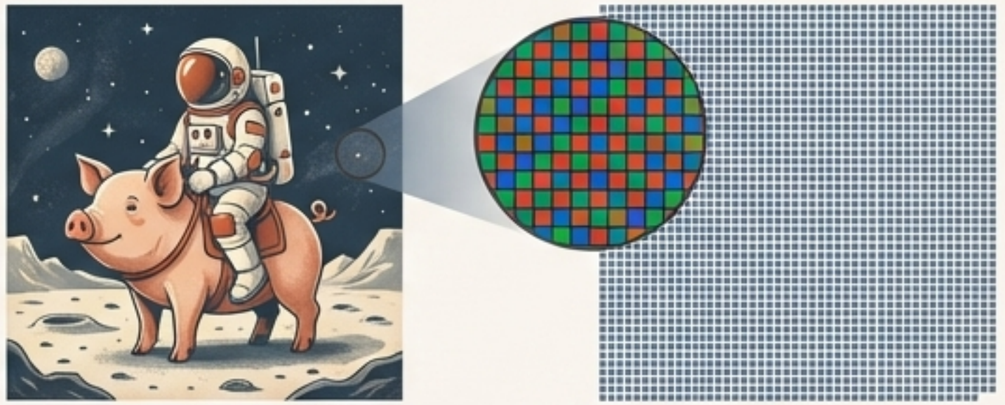
O Mundo dos Tokens (Texto)

un astronauta montando un porco

[un] → [astro] → [nauta] → [montando] → [un] → [porco]

Poucas decenas de tokens discretos.

O Mundo dos Píxeles (Imaxe)



512x512 píxeles = 262.144 píxeles.

Cada un con 3 valores de cor (RGB) = **786.432 valores numéricos**

Centos de miles de valores continuos.

A xeración de texto elixe entre unhas poucas decenas de miles de tokens. A xeración de imaxes debe definir o valor de centos de miles (ou millóns) de píxeles.

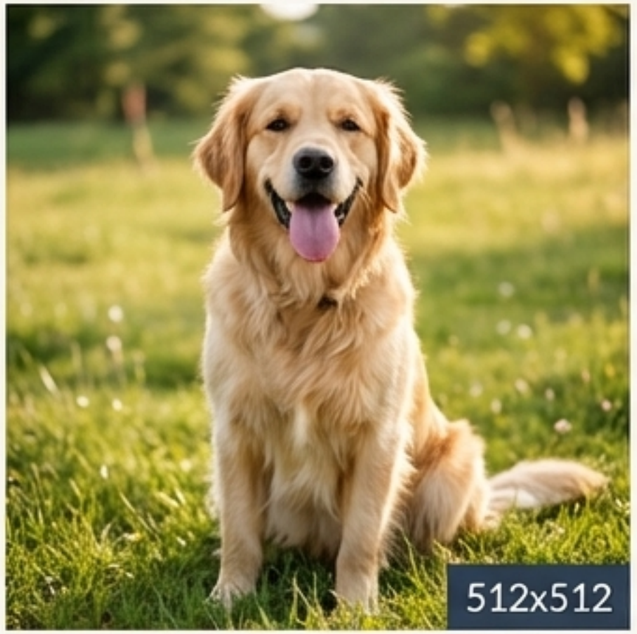
O espazo de saída **non é discreto** (un conxunto finito de palabras), senón **continuo** (un espectro case infinito de cores e texturas).

Operar directamente sobre os píxeles sería computacionalmente prohibitivo. Necesítase unha estratexia máis intelixente.

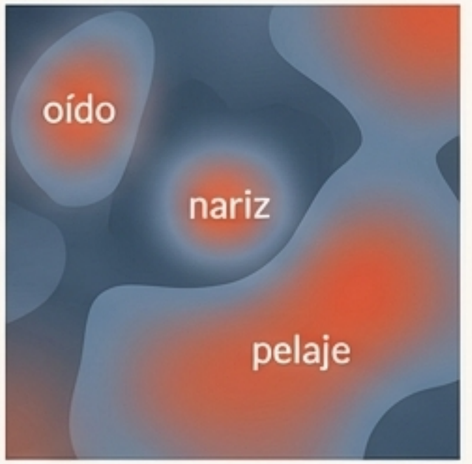
A Solución: Traballar con Ideas, non con Píxeles

Benvido ao Espazo Latente

A Gran Idea: Os Modelos de Difusión Latente (LDMs) non traballan directamente cos píxeles. Primeiro, comprimen a imaxe nun **espazo latente** de baixa dimensión.



Codificador VAE



Representación Latente

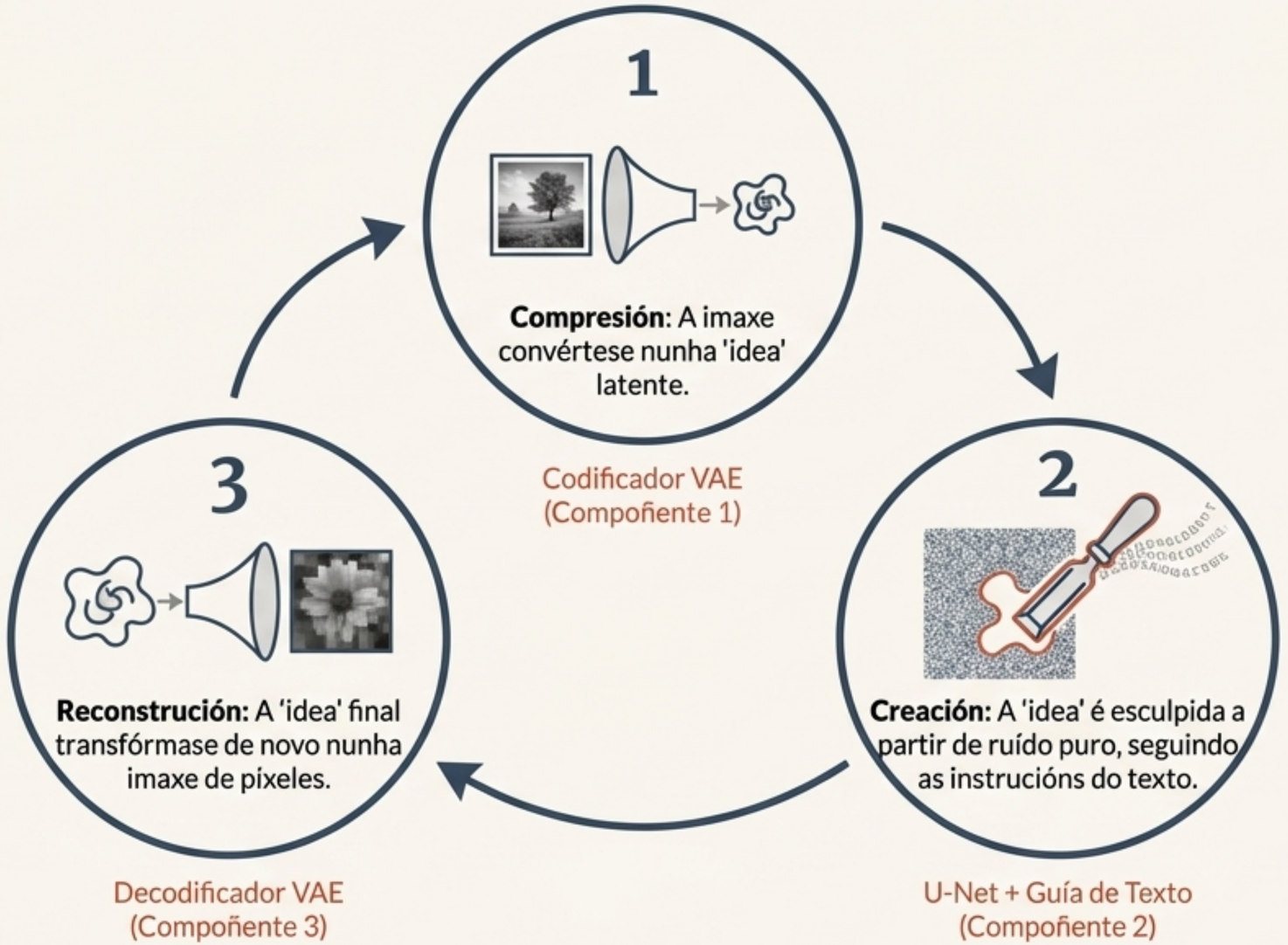
Cifras Impactantes
Imaxe de Píxeles: `512 x 512 x 3 = 786.432` dimensións.
Mapa Latente: `64 x 64 x 4 = 16.384` dimensións.

Unha redución de máis do 98% na complexidade computacional.

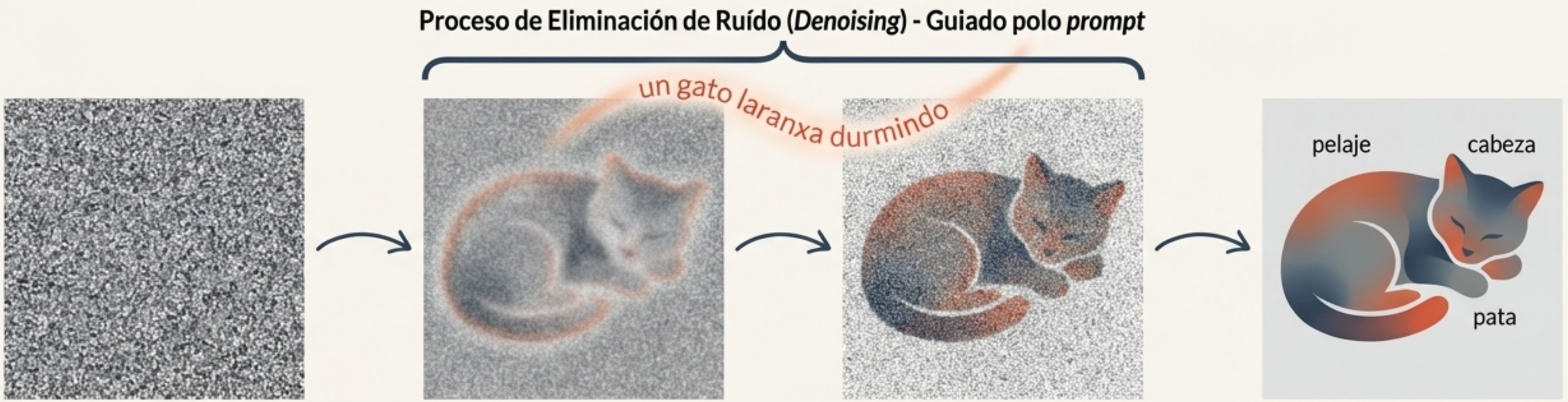
Que é realmente?: O espazo latente non é unha imaxe máis pequena. É un **mapa conceptual** que codifica as características esenciais e semánticas: 'can', 'peludo', 'feliz', 'exterior', antes de preocuparse polos píxeles individuais.

A Arquitectura dun Soño: Tres Pasos para Crear unha Imaxe

A arquitectura dun LDM consta de tres módulos clave que traballan en harmonía. Vexamos como funciona o corazón creativo do proceso: a creación a partir do ruído.



O Proceso Creativo: Esculpindo unha Imaxe a partir do Caos



Comezo: Ruído Gaussiano puro

Final: Representación Latente "limpa"

O Corazón Xerativo

O núcleo do modelo (unha arquitectura chamada **U-Net**) aprendeu a facer unha soa cousa de forma experta: predicir e eliminar o ruído dun mapa latente.

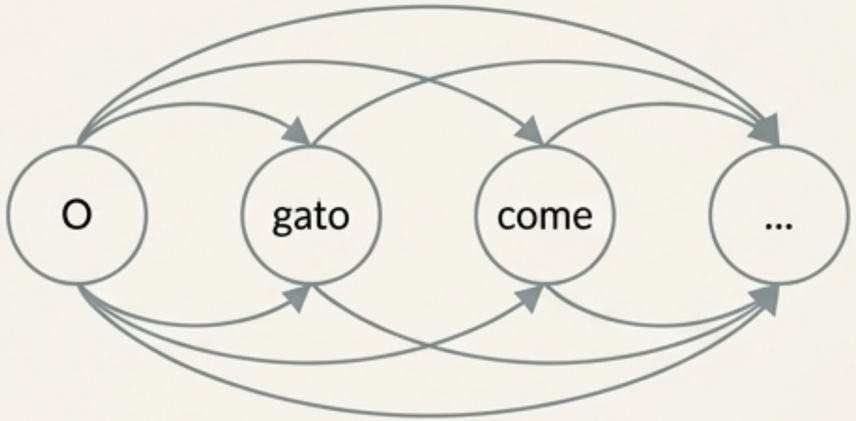
A Metáfora do Escultor

O proceso é como un escultor que comeza cun bloque de mármore ruidoso. En cada paso, o modelo, guiado polas instrucións do *prompt*, elimina unha capa de 'ruído', revelando gradualmente a imaxe oculta no seu interior.

O Papel do *Prompt*: Pero, como sabe o escultor que esculpir? Como o *prompt* de texto guía este proceso?

A Ponte entre Palabras e Píxeles: O Poder da Atención

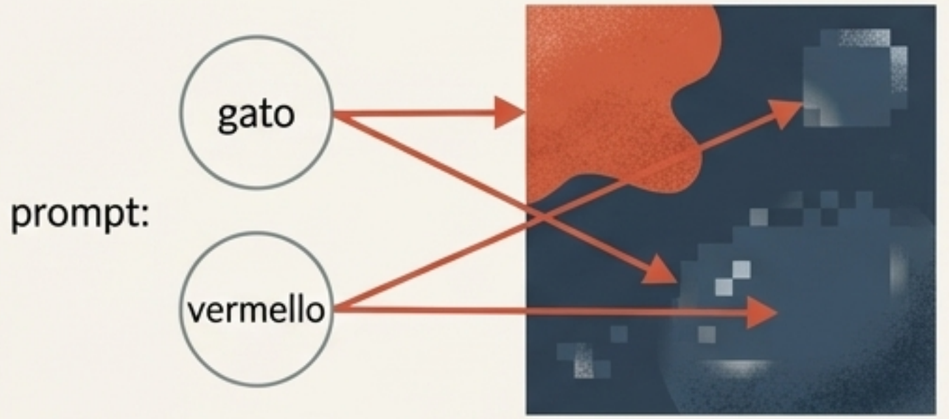
Auto-Atención (*Self-Attention*)



Coherencia Lingüística

Conecta un token con todos os demais da *súa propia secuencia* para construír contexto e razoamento interno (p. ex., 'el' refírese a 'o can').

Atención Cruzada (*Cross-Attention*)



Fusión Modal e Aliñamento

Inxecta a guía semántica do *prompt* no espazo visual. Conecta o *embedding* de texto co vector latente da imaxe, actuando como un 'tradutor' ou 'director de orquestra'.

A Auto-Atención mira cara a dentro para manter a coherencia. A Atención Cruzada mira cara a fóra para seguir instrucións. Esta é a innovación clave que permite a xeración de texto a imaxe.

A Atención Cruzada en Acción

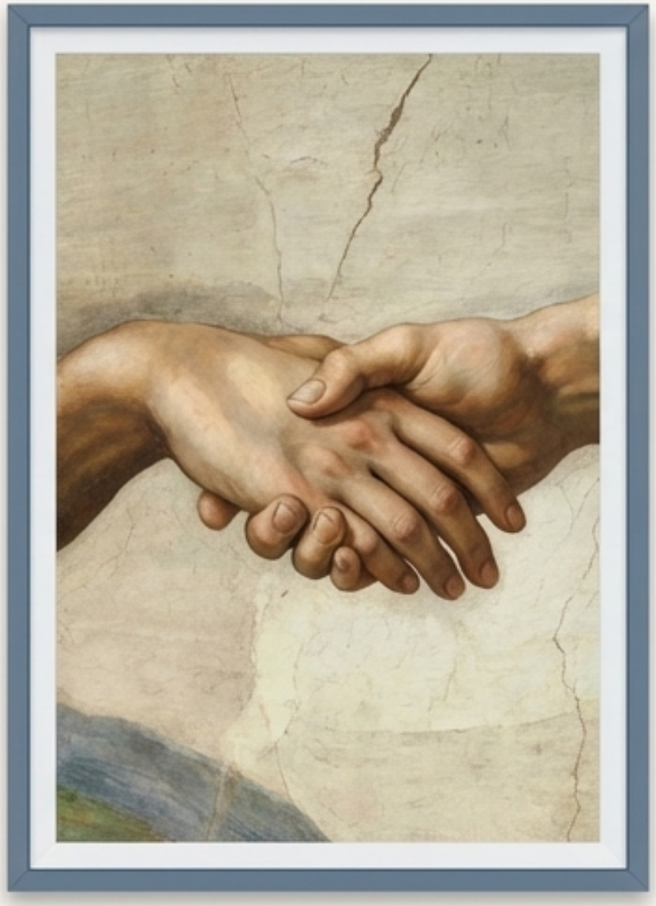
Unha foto dun astronauta montando un porco rosa, cun pardal que leva un sombreiro de mago ao seu lado

- A Atención Cruzada permite que cada rexión espacial do mapa latente preste atención á palabra máis relevante do prompt.
- O modelo non só entende os conceptos, senón que os "sitúa" correctamente porque o mecanismo de atención aliña as características textuais coas espaciais.

Mellora Continua

Modelos como Stable Diffusion 3 utilizan arquitecturas máis avanzadas (Diffusion Transformers) que melloran drasticamente esta capacidade de atención. O resultado? Mellor seguimento de instrucións e incluso a capacidade de xerar texto lexible con precisión.

O Fantasma na Máquina: Por que as IAs teñen problemas coas mans?



O problema non é a falta de imaxes de mans nos datos de adestramento. É unha consecuencia directa da compresión no espazo latente.

A "Alucinación Xeométrica"

O codificador VAE prioriza as características semánticas xerais ("man", "figura humana") sobre os detalles xeométricos de alta frecuencia e precisión (o número exacto de articulacións e dedos).

A IA non "conta" os dedos; reconstrúe o que é "plausible" a partir dun plano conceptual que perdeu parte da súa precisión estrutural.

Este é un fallo no razoamento espacial, non no coñecemento visual. O modelo entende o concepto de "man", pero non a súa estrutura precisa.

Un Contraste de Fallos: Estrutura vs. Verdade

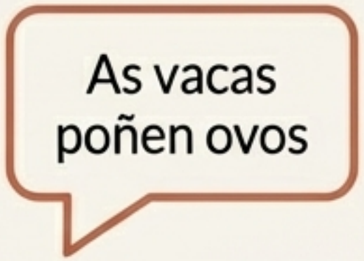
LDM (Imaxes)

Tipo de Erro: Alucinación Xeométrica

Causa: Perda de precisión estrutural no espazo latente comprimido.

Resultado: A imaxe é visualmente plausible pero estruturalmente incorrecta. Falla na **coherencia xeométrica**.

LLM (Texto)



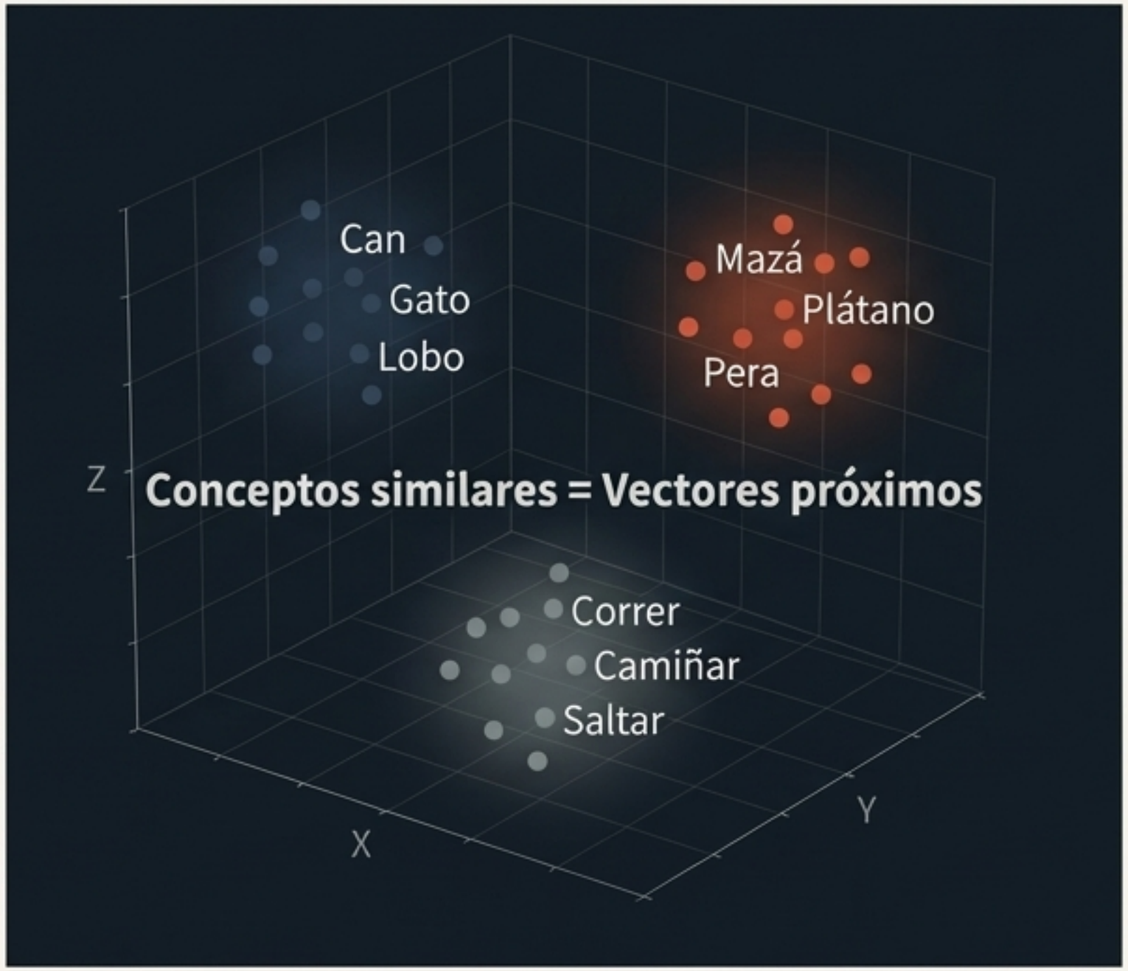
Tipo de Erro: Alucinación Semántica

Causa: Prioriza o patrón estatístico sobre a veracidade factual.

Resultado: A frase é gramaticalmente perfecta pero factualmente incorrecta. Falla na **coherencia semántica (a verdade)**.

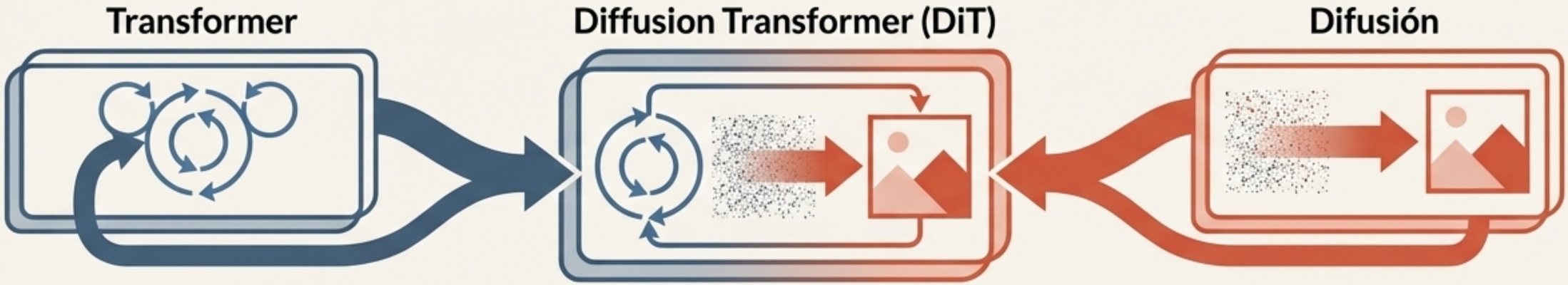
As debilidades de cada modelo revelan a súa natureza fundamental: os LDMs loitan coa ***estrutura física precisa***, mentres que os LLMs loitan coa ***verdade abstracta***.

A Linguaxe Universal da IA: O Espazo de *Embeddings*



- A pesar das súas arquitecturas diferentes, ambos os modelos operan sobre o mesmo principio: **a representación vectorial.**
- Tanto o **Espazo Latente** dunha imaxe como os **Embeddings de Tokens** dun texto son espazos de alta dimensión onde os conceptos se representan como vectores numéricos.
- A IA non manipula píxeles ou palabras, senón que realiza operacións matemáticas sobre estes vectores. **A distancia entre vectores determina a relación conceptual.**
- Isto é o que permite á IA 'entender' que '**Rei**' - '**Home**' + '**Muller**' ≈ '**Raíña**', ou que a imaxe dun 'gato persa' debe estar preto da dun 'gato siamés'.

O Futuro é Híbrido: Cando o que Ve Aprende do que Fala



A Converxencia

O futuro da IA xerativa reside na fusión destas arquitecturas. Os modelos máis recentes comezan a integrar as fortalezas de ambos os mundos.

Exemplo: Stable Diffusion 3

SD3 incorpora un **Diffusion Transformer (DiT)** no seu núcleo. Isto significa que o razoamento contextual avanzado e a comprensión de secuencias dos LLMs se integran directamente no proceso de eliminación de ruído da imaxe.

O Que Significa Isto?

- Unha adherencia ao *prompt* e unha coherencia semántica aínda maiores.
- O potencial para superar limitacións históricas como a coherencia xeométrica.
- Un paso máis cara a modelos xenerativos verdadeiramente multimodais, capaces de razoar e crear a través de diferentes dominios sen problemas.

A división entre ver e falar está a desaparecer. A próxima xeración de IAs non só creará, senón que tamén comprenderá a un nivel máis profundo.