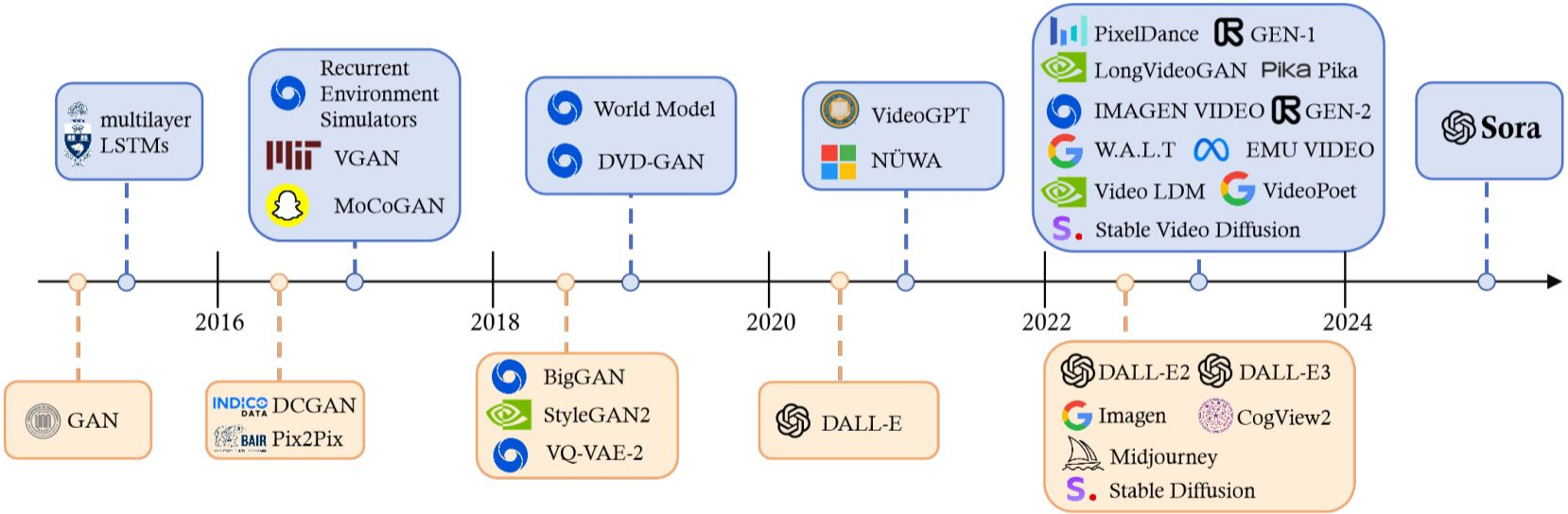
Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan,, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, Lichao Sun
Microsoft Research

History of Generative AI in Vision Domain.



What we expect in Large Vision Models (LVMs)?

1. Scaling Laws:
    a. Diffusion transformers scale effectively as video models. A comparison of video samples with fixed seeds and inputs demonstrates that sample quality improves markedly as training compute increases
2. Emergent Abilities:
    a. Emergent abilities in LLMs and LVMs are sophisticated behaviors or functions that manifest at certain scales, arising from comprehensive training across varied datasets and extensive parameter counts.

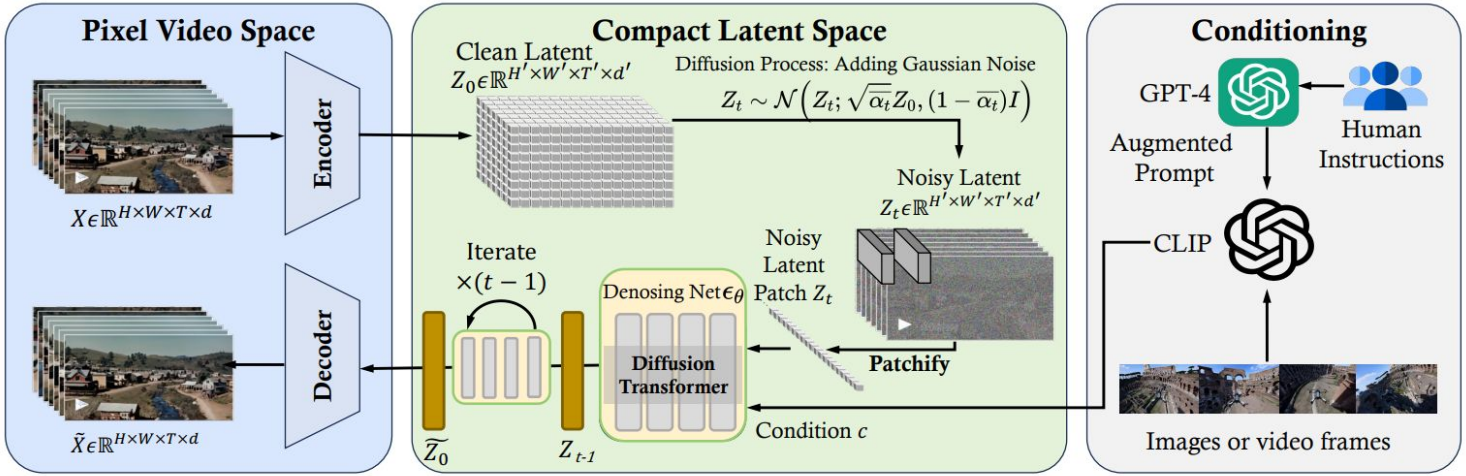
A comparison of video samples with fixed seeds and inputs (Base, 4x compute, 32x compute)

Overview

Sora is a diffusion transformer with flexible sampling dimensions.

Comprises three main components:

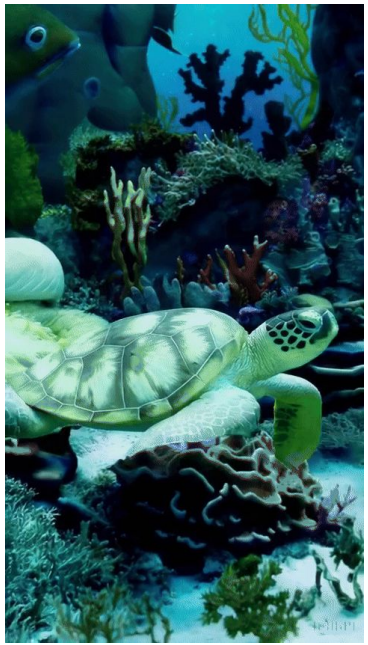
- Time-space compressor
- Vision Transformer (ViT)
- CLIP-like conditioning mechanism



Variable Durations, Resolutions, Aspect Ratios

Native Size Training & Generation

- Preserves Original Dimensions: No resizing, cropping, or aspect ratio adjustments
- Supports Diverse Formats: Handles widescreen 1920x1080p, vertical 1080x1920p, etc.



Variable Durations, Resolutions, Aspect Ratios

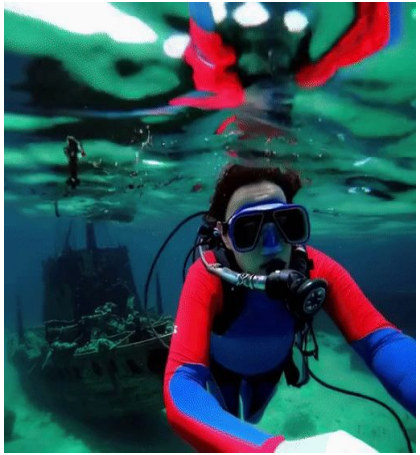
Traditional Methods:

- Often resize, crop, or adjust aspect ratios
- Typically use short clips with square frames (Fixed low resolutions)

Sora's Advantages:

- Sora's videos exhibit better framing
- Ensures subjects are fully captured

Trained on videos that are cropped to squares.



Trained on native sizes.

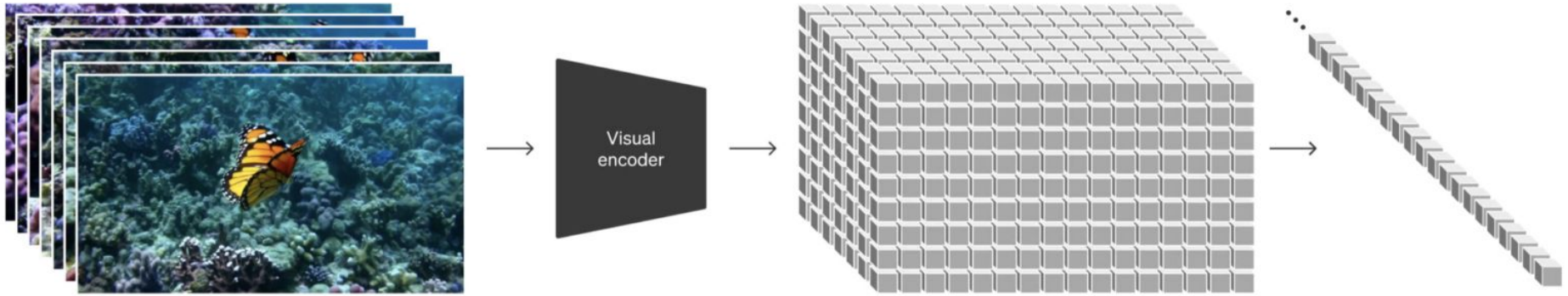
Unified Visual Representation

Transforming Visual Data:

- Handles images and videos of varying durations, resolutions, and aspect ratios
- Facilitates large-scale training of generative models

Sora's Method:

- Compresses videos into a lower-dimensional latent space (spacetime patches)



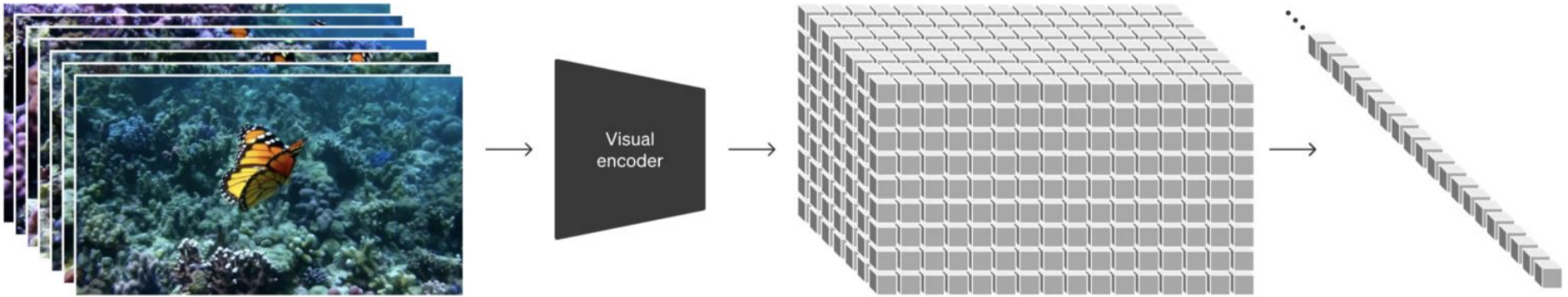
Video Compression Network

Objective:

- Reduce dimensionality of raw video input
- Output a compressed latent representation both temporally and spatially

Technical Foundation:

- Built upon VAE or Vector Quantised-VAE (VQ-VAE)
- Challenge: Mapping visual data of any size to a unified, fixed-sized latent space without resizing and cropping



Approach 1 (Spatial-patch Compression)

Methodology:

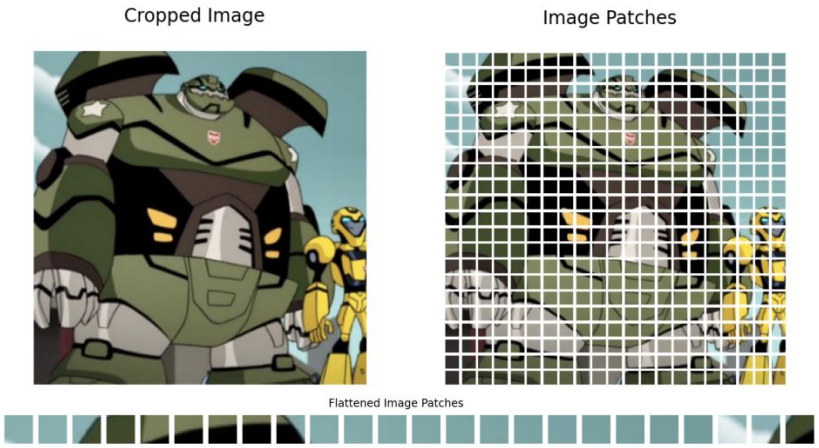
- Transform Video Frames into Fixed-Size Patches (Similar to ViT and MAE)
- Encodes frames through individual patches
- Organize Spatial Tokens in Temporal Sequence

Temporal Dimension Variability:

- Challenge: Varying video durations
- Solutions:
  - Sample specific number of frames (use padding or temporal interpolation for short videos)
  - Define a universally extended input length

Temporal Information Aggregation:

- Importance: Captures dynamic changes over time
- Mechanism: Additional mechanism required



An image is worth 16x16 words (A. Dosovitskiy), Masked autoencoders are scalable vision learners (K. He), Preserve your own correlation(S. Ge)

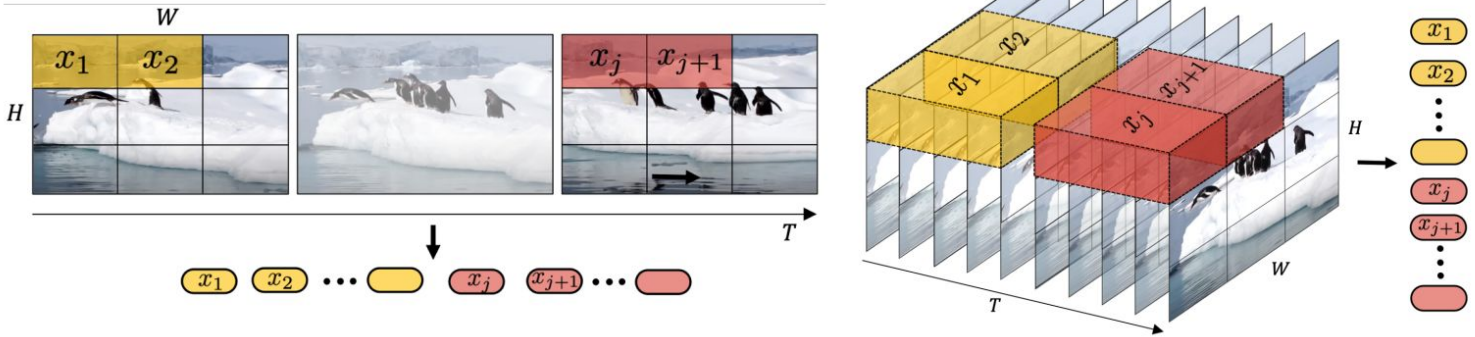
Approach 2 (Spatial-temporal-patch Compression)

Encapsulates Both Spatial and Temporal Dimensions:

- Captures movement and changes across frames
- Represents video's dynamic aspects

Methodology:

- Utilization of 3D Convolution:
- Simple and effective integration method



Tokenlearner: Adaptive space-time tokenization for videos (M. Ryoo)

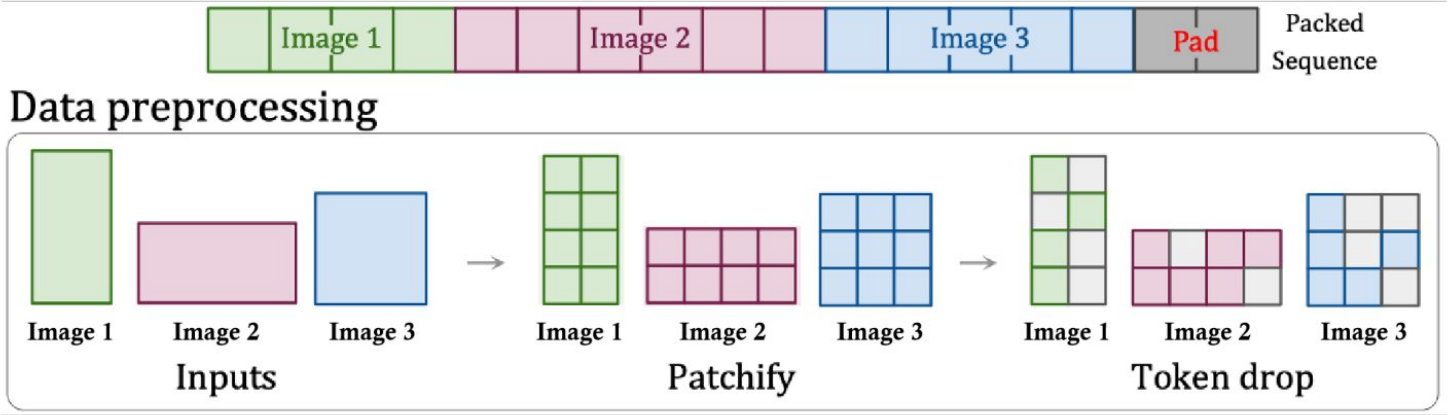
Spacetime Latent Patches

Handling Latent Space Variability:

- Challenge: Variability in number of latent patches from different video types
- Solution: Patch n' Pack (PNP) which packs multiple patches from different images in a single sequence

Patchification and Token Embedding:

- Completed in the compression network
- Potential for second-round patchification for transformer tokens as in Diffusion Transformer

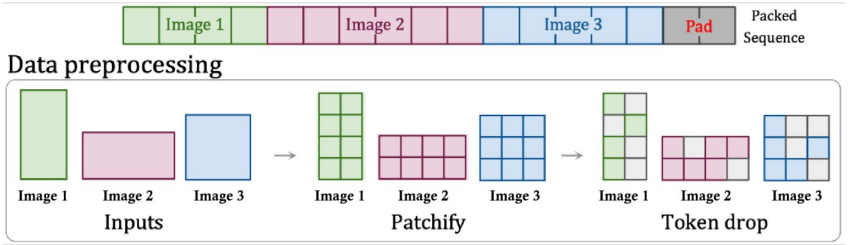


Patch n'pack (M. Dehghani), Scalable diffusion models with transformers (W. Peebles)

Dropping Tokens

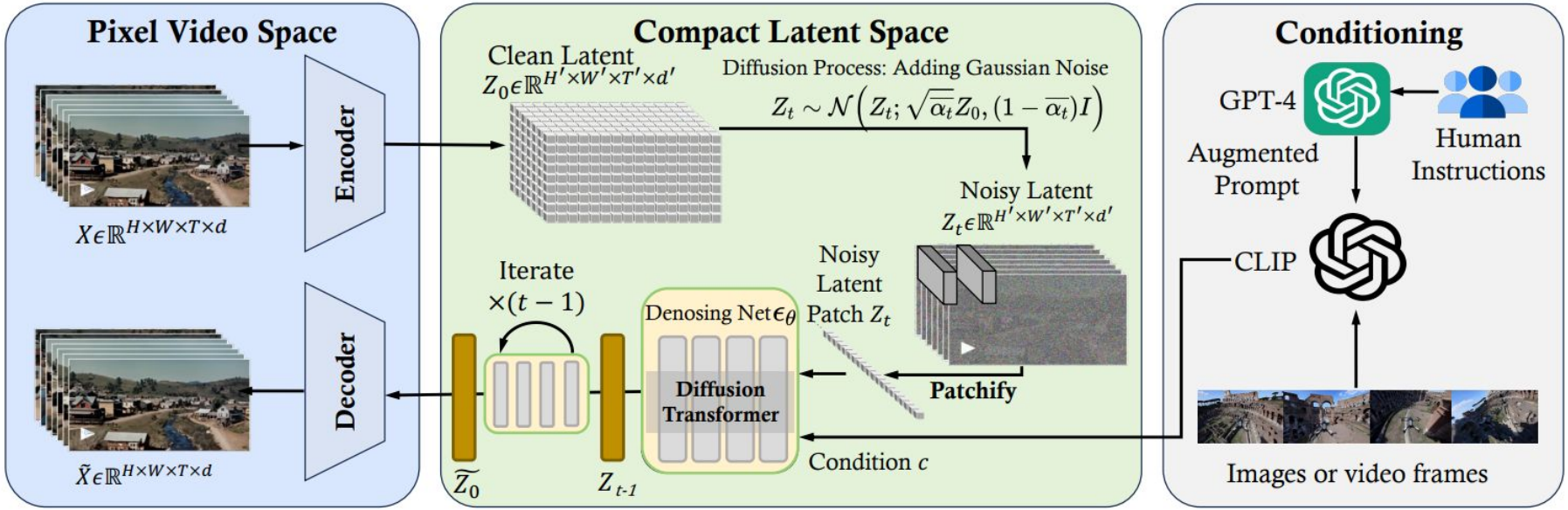
Concerns and Solutions:

- **Packing Tokens Compactly:**
  - Simple greedy approach: Adds examples to the first sequence with enough remaining space
  - Sequences filled with padding tokens for fixed lengths needed for batched operations
  - Tuning sequence length and limiting padding
- **Dropping Tokens:**
  - Methods: Drop similar tokens or use dropping rate schedulers
  - Note: Dropping tokens may ignore fine-grained details (3D Consistency)
  - Likely Approach: Use a super long context window to pack all tokens from videos despite computational cost (multi-head attention has quadratic cost in sequence length)



Is space-time attention all you need for video understanding? (G. Bertasius)

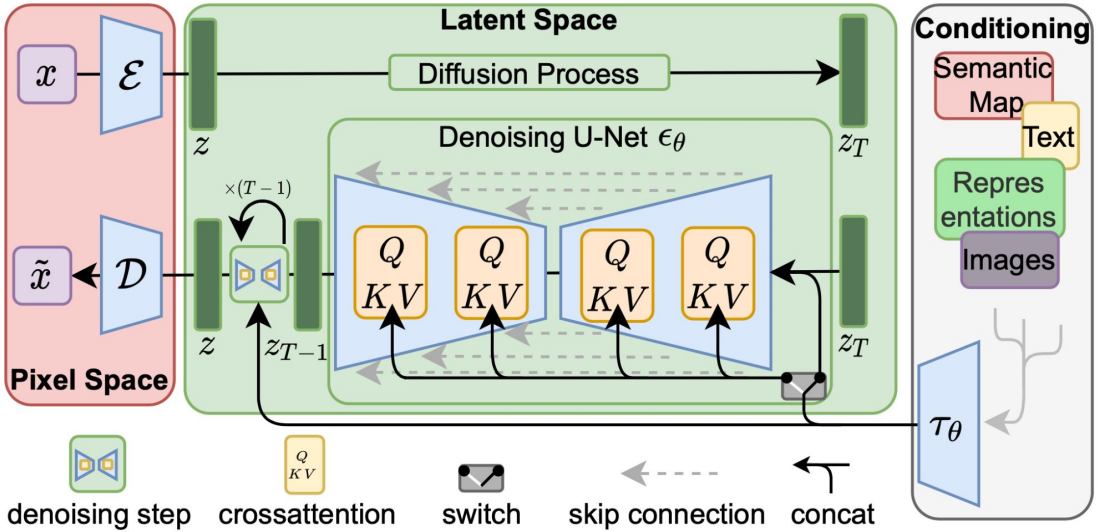
Methodology



U-Net-Based Diffusion Models

Traditional Diffusion Models:

- Utilize convolutional U-Nets with downsampling and upsampling blocks [51, 52, 53]
- Recent studies show U-Net architecture is not crucial for good performance

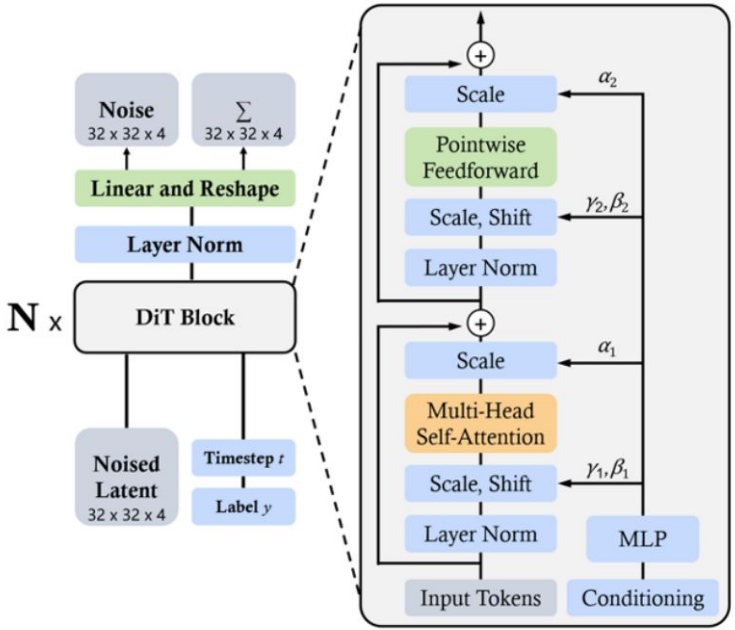


Denoising diffusion probabilistic models (J. Ho)

Transformer-Based Diffusion Models

DiT (Diffusion Transformer)

- Employs vision transformers (ViT)
- Uses multi-head self-attention, pointwise feed-forward network, adaptive layer norm (AdaLN)
- Stabilizes training with zero-initializing MLP layer
- Empirically validated scalability and flexibility



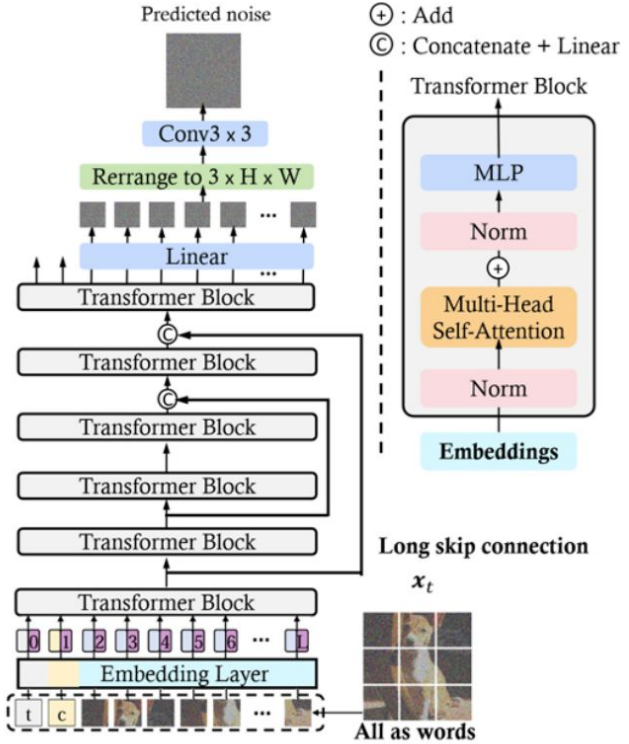
Latent Diffusion Transformer DiT Block with adaLN-Zero

Scalable diffusion models with transformers (W. Peebles)

Transformer-Based Diffusion Models

U-ViT (U-Net Vision Transformer)

- Treats all inputs (time, condition, noisy image patches) as tokens
- Uses long skip connections between shallow and deep transformer layers
- Achieves record-breaking FID scores

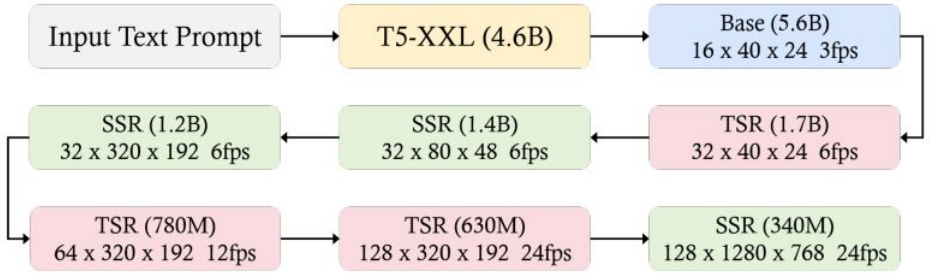


All are worth words: A vit backbone for diffusion models (F. Bao)

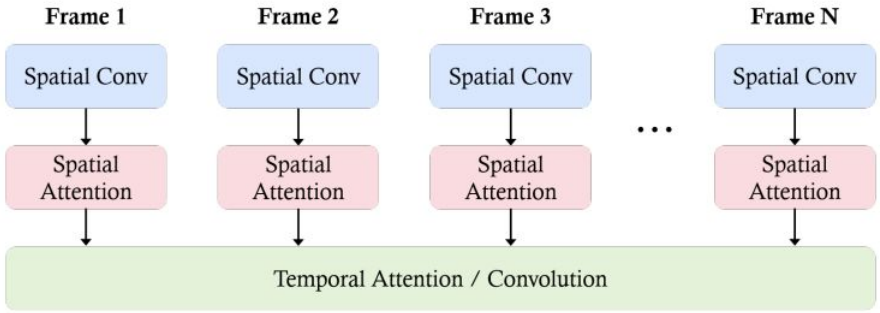
Video Diffusion

Imagen Video

- Text Encoding: Frozen T5 text encoder generates contextual embeddings
- Base Model: Generates low-resolution video from embeddings
- Cascaded Diffusion Models: Refines video resolution
- Architecture: Utilizes 3D U-Net in spatial-temporal separable fashion
- Joint Training: On images and videos, treating each image as a frame



(a) **Cascaded diffusion models.** The cascaded sampling pipeline with a base diffusion model and six up-sampling models that operate spatially and temporally. The text embeddings are injected into all the diffusion models.



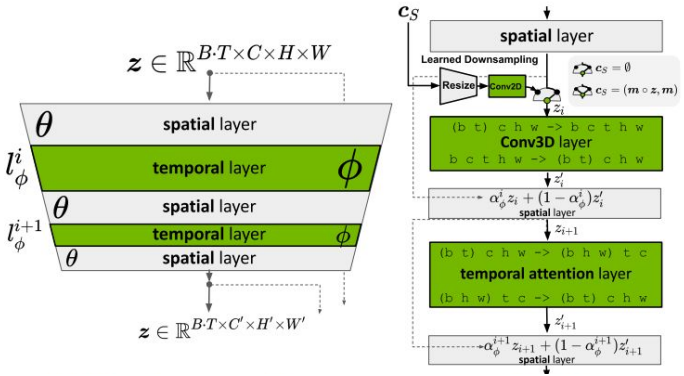
(b) **Video U-Net space-time separable block.** Spatial operations are performed independently over frames with shared parameters, whereas the temporal operation mixes activations over frames. Temporal attention is only used in the base model for memory efficiency.

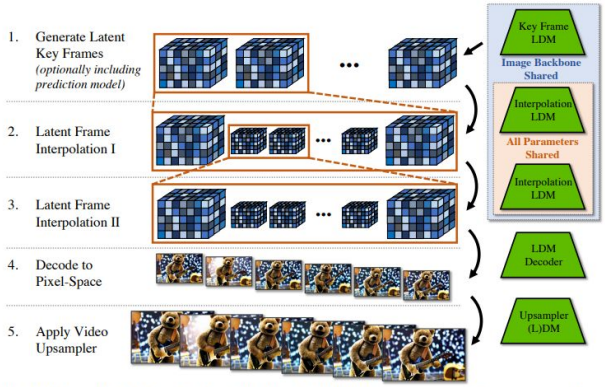
Imagen video: High definition video generation with diffusion model (J. Ho)

Video Diffusion

Video Latent Diffusion Model (Video LDM)

- Post-hoc Temporal Layers:
  - Added among existing spatial layers in U-Net backbone and VAE decoder
  - Temporal layers trained on encoded video data; spatial layers remain fixed
- Key Frame Generation and Interpolation
  - Divides synthesis process into key frame generation and interpolation for high temporal resolution
- Cascaded LDMs
  - Use DM to scale up outputs by 4x





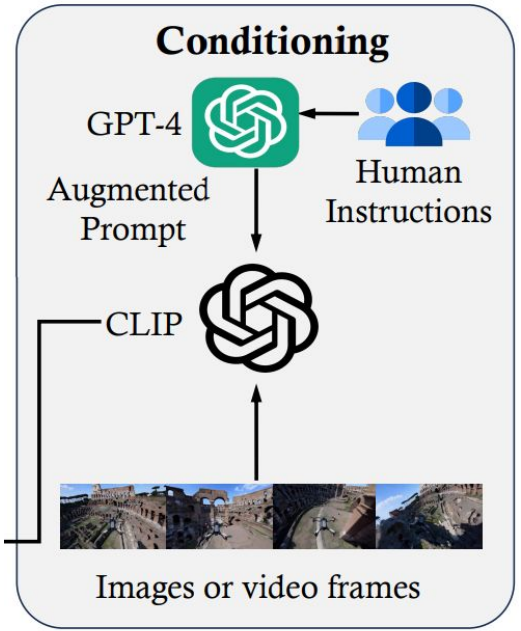
Align your Latents:High-Resolution Video Synthesis with Latent Diffusion Models (A. Blattmann)

Language Instruction Following

- User Engagement:
  - Generative AI models interact via natural language instructions (text prompts)
- Model Instruction Tuning:
  - Enhances AI models' ability to follow prompts accurately, generating outputs that resemble human responses
- DALL·E 3:
  - Caption Improvement: Addresses data quality issues by re-captioning with detailed descriptions
  - Method: Trains a vision-language model to generate precise captions; fine-tunes text-to-image models using these captions

Language Instruction Following

- Video Caption Improvement:
  - Trains a video captioner to produce detailed video descriptions
  - Generates high-quality (video, descriptive caption) pairs for fine-tuning Sora
- VideoCoCa Architecture:
  - Utilizes CoCa's image encoder on sampled video frames
  - Processes frame tokens with generative and contrastive poolers
- Alternatives:
  - mPLUG-2, GIT, FrozenBiLM
- Prompt Extension:
  - To align user prompts with training data formats, Sora uses GPT-4V to expand user inputs into detailed descriptive prompts



Video-text modeling with zero-shot transfer from contrastive captioners (S. Yan), GIT (J. Wang), mPLUG-2(H. Xu), FrozenBiLM (A. Yang)

Prompt Engineering

Importance: Essential for guiding text-to-video models like Sora to generate high-quality videos aligned with user expectations.

Types of Prompts:

- Text Prompts: Crafting detailed descriptions to direct content creation.
- Image Prompts: Acting as visual anchors, guiding animation of elements.
- Video Prompts: Used for tasks such as video extension, editing, and connection.

Text Prompt

Prompt:

- New York City submerged like Atlantis. Fish, whales, sea turtles and sharks swim through the streets of New York.

Image Prompt

Video Prompt

Each of the three videos starts different from the others, yet all three videos lead to the same ending.

Video Prompt

Editing using SDEdit

This technique enables Sora to transform the styles and environments of input videos zero-shot.

Change the setting to be in a lush jungle

SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations (C. Meng)

Limitations

Physical Realism:

- Sometimes fails to handle physical principles accurately
- Issues with incorrect motion simulation and unrealistic object interactions

Limitations

Spatial and Temporal Complexities:

- Misunderstands instructions related to object placement and event timing
- Results in inaccurate scene compositions and disrupted temporal flow

Thank you!

- Thank you for your attention!
- I appreciate your time and interest.
- If you have any questions, please feel free to ask.
- Contact information: alimohammadiamirhossein@gmail.com

Extra Slides

Emergent Abilities

Emergent abilities in LLMs are sophisticated behaviors or functions that manifest at certain scales, often linked to the size of the model's parameters.

These abilities are not explicitly programmed or anticipated by developers but emerge from the model's comprehensive training across varied datasets and extensive parameter count.

Emergent abilities surpass mere pattern recognition or rote memorization, enabling the model to form connections and draw inferences.

They cannot be straightforwardly predicted by extrapolating from the performance of smaller-scale models.

LLMs like ChatGPT and GPT-4 exhibit emergent abilities in natural language processing tasks.

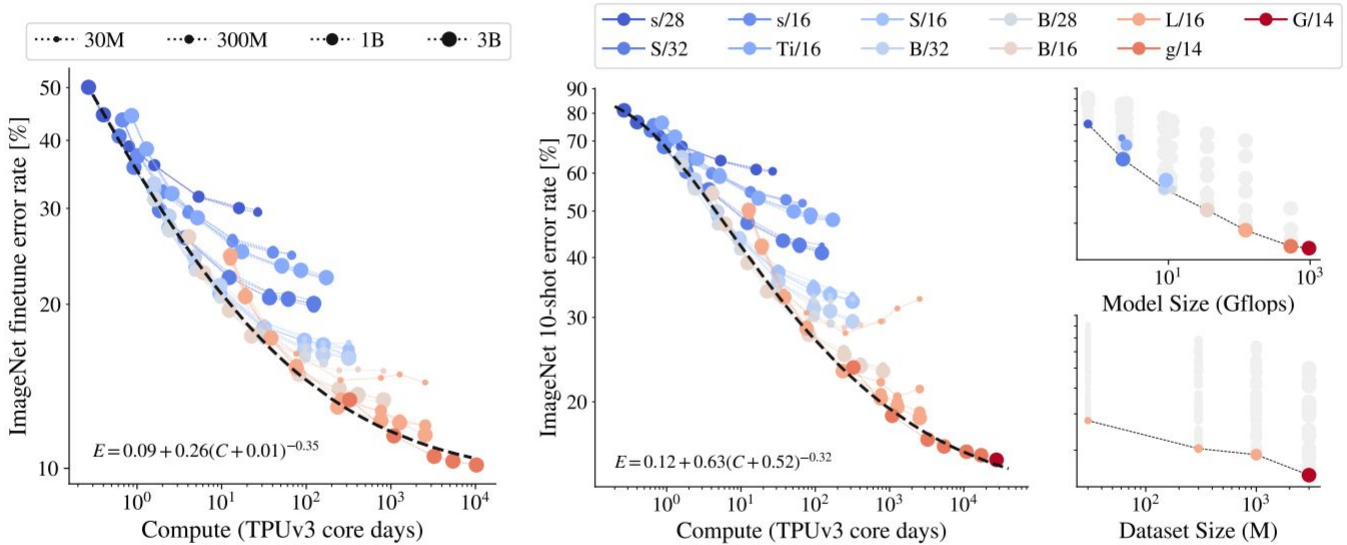
However, until the advent of Sora, vision models demonstrating comparable capabilities were scarce.

Sora is the first vision model to exhibit confirmed emergent abilities, marking a significant milestone in the field of computer vision.

Scaling Laws for Vision Models

Inspired by Language Models: Just as language models (LLMs) exhibit scaling laws, recent research suggests that vision models also follow similar patterns.

Performance-Compute Frontier: Zhai et al. demonstrated that Vision Transformer (ViT) models adhere to a power law relationship between performance and computational resources, with saturating returns.



Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. "Scaling Vision Transformers." arXiv preprint arXiv:2106.04560 (2022).

Scaling Laws for Vision Models

Efficient Training Recipe: Google Research proposed an efficient training method for a 22B-parameter ViT, achieving impressive performance by training only thin layers on top of frozen embeddings.

Alignment with Scaling Principles: Sora, a large vision model (LVM), embraces these scaling principles, revealing emergent abilities in text-to-video generation.

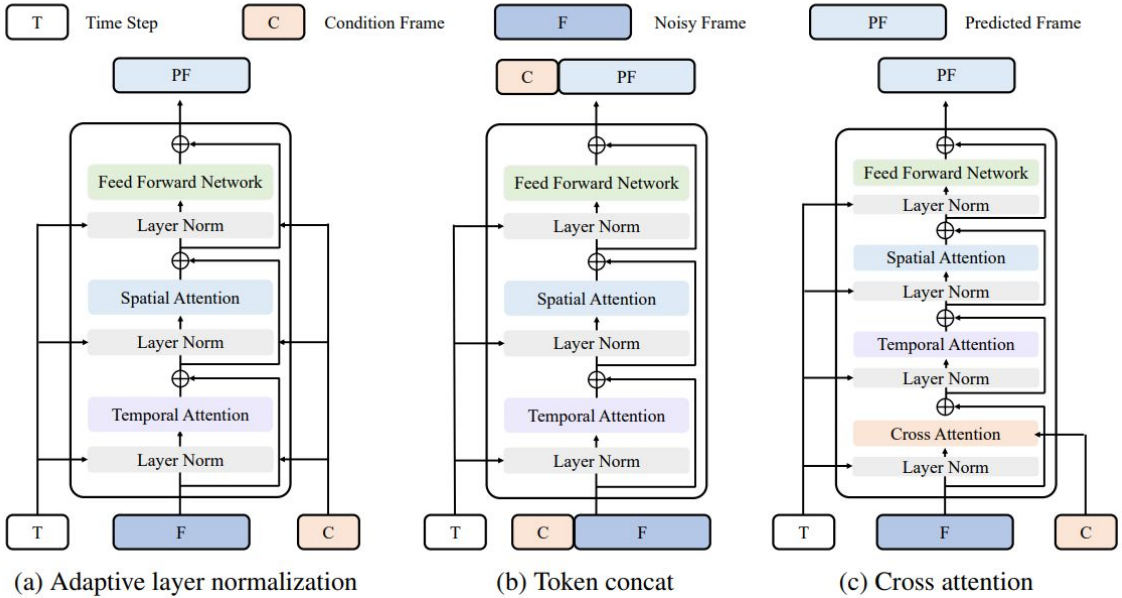
Potential Advancements: The alignment of vision models with scaling laws suggests the potential for similar advancements to those seen in language models, marking significant progress in computer vision research.

Model	IN	ReaL	INv2	ObjectNet	IN-R	IN-A
224px linear probe (frozen)						
B/32	80.18	86.00	69.56	46.03	75.03	31.2
B/16	84.20	88.79	75.07	56.01	82.50	52.67
ALIGN (360px)	85.5	-	-	-	-	-
L/16	86.66	90.05	78.57	63.84	89.92	67.96
g/14	88.51	90.50	81.10	68.84	92.33	77.51
G/14	88.98	90.60	81.32	69.55	91.74	78.79
e/14	89.26	90.74	82.51	71.54	**94.33**	81.56
22B	**89.51**	**90.94**	**83.15**	**74.30**	94.27	**83.80**
High-res fine-tuning						
L/16	88.5	90.4	80.4	-	-	-
FixNoisy-L2	88.5	90.9	80.8	-	-	-
ALIGN-L2	88.64	-	-	-	-	-
MaxViT-XL	89.53	-	-	-	-	-
G/14	90.45	90.81	83.33	70.53	-	-
e/14	90.9	91.1	84.3	72.0	-	-

Mostafa Dehghani et al. "Scaling Vision Transformers to 22 Billion Parameters." arXiv:2302.05442 (2023).

VDT: An Empirical Study on Video Diffusion with Transformers

Methods	FVD ↓	SSIM ↑	PSNR ↑
Ada. LN	270.8	0.6247	16.8
Cross-Attention	134.9	0.8523	28.6
Token Concat	129.1	0.8718	30.2



Limitations

Limitations in Human-Computer Interaction:

- Difficulty in making precise modifications to generated content
- Struggles with understanding complex language instructions