World Modeling

A Short Introduction, from Representation Learning to Video Generation

Disclaimer: Trop de Franglais dans ces slides...

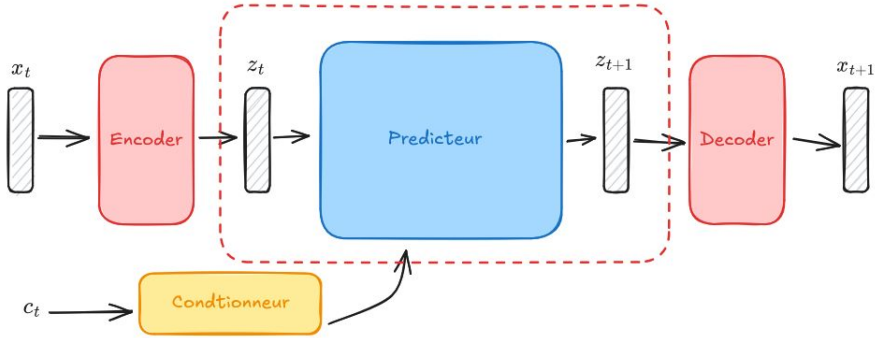
World Modeling

Mot très fancy pour un truc assez simple

Encoder = Représenter une observation de l'environnement (1/plusieurs modalités)

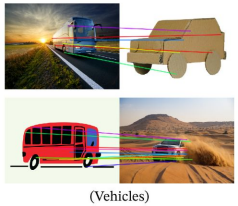
Conditionneur = Représenter un ensemble des actions qui permettent d'avoir une influence sur le futur d'une modalité différente

Prédicteur = Prédire le prochain état

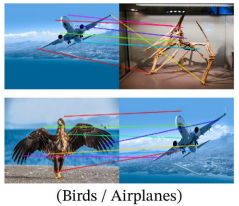


Representation @Dino v1,2,3

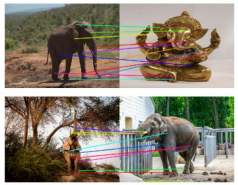
From Image to vectors



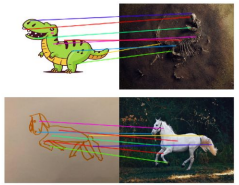
(Vehicles)



(Birds / Airplanes)



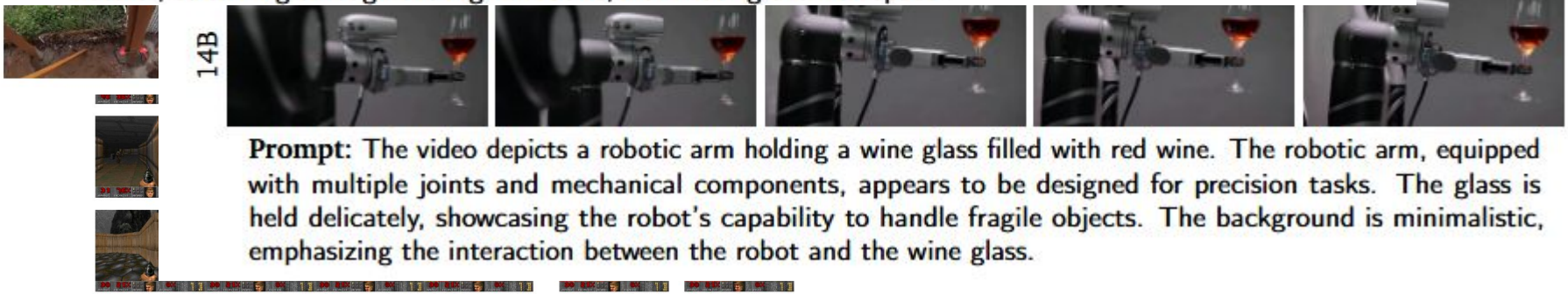
(Elephants)



(Drawings / Animals)

(a) (b) (c) (d)

Prompt: The video depicts the interior of a large industrial facility, likely a factory or warehouse. The space is expansive with high ceilings and metal framework. Overhead cranes and various machinery are visible, indicating a setting for heavy manufacturing or assembly. The floor is mostly empty, with some scattered debris and marked lines. Safety signs and barriers are present, emphasizing the industrial environment. The lighting is natural, streaming through the high windows, illuminating the workspace.



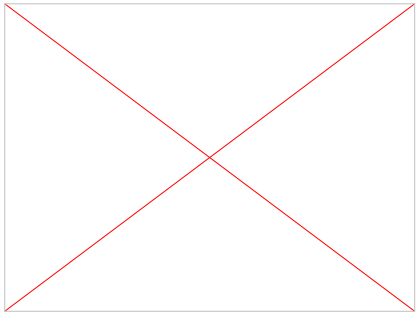
Prompt: The video depicts a robotic arm holding a wine glass filled with red wine. The robotic arm, equipped with multiple joints and mechanical components, appears to be designed for precision tasks. The glass is held delicately, showcasing the robot's capability to handle fragile objects. The background is minimalistic, emphasizing the interaction between the robot and the wine glass.

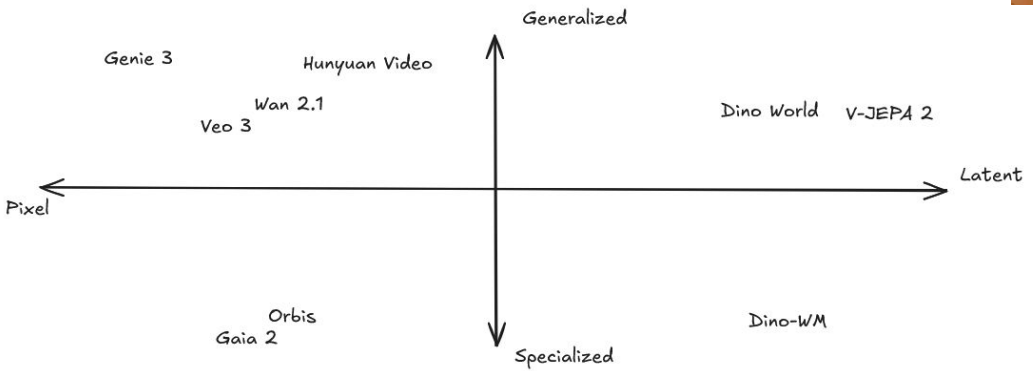
World Modeling

Google : Genie 3

t=14s

Meta : V-JEPA 2



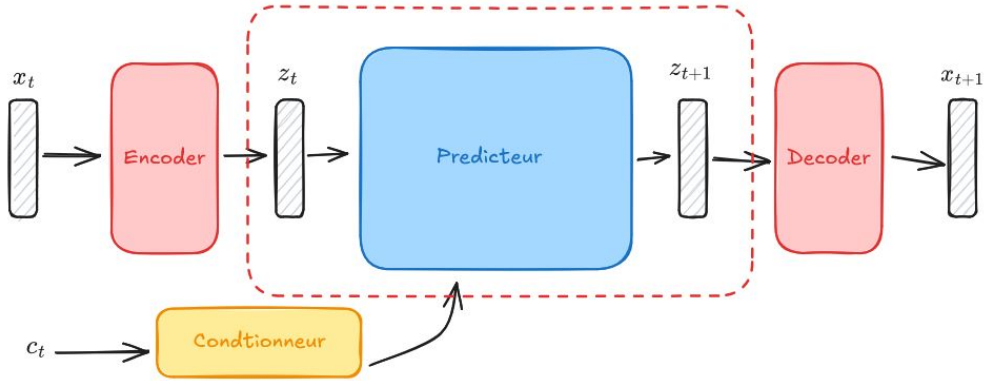


Attention Name-Dropping

En pratique, comment on fait ?

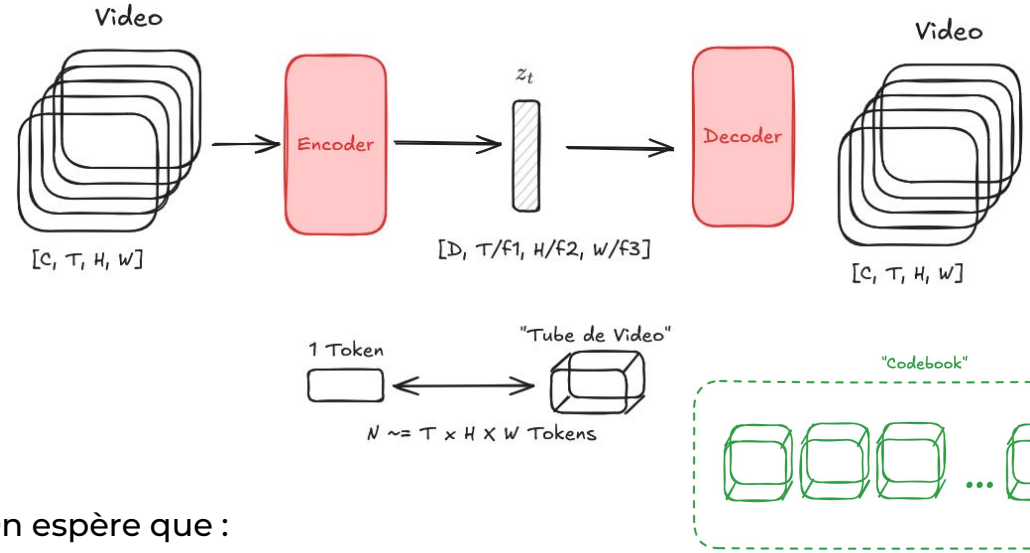
1. Tokenisation = Encoder Video / Image / Text dans un latent space (optimise pour la compression, *Rombach et al. 2021*) = **coute moins cher**
2. Choisir le mode opératoire de génération : AutoRegression ou Diffusion
3. Masquer / Dégrader la donnée réelle
4. Prédire & Optimiser

Discret ou Continu ?



Voir Références a la fin

Encoding/Decoding



Souvent :

- VAE/ VQ-VAE
- “Fondational Image Model”

On espère que :

- Entrée = Sortie
- Que f1, f2, f3 sont élevés (sinon ça coute cher)
- Le vecteur latent est sémantiquement intéressant

(Dans le cas discret, on apprend un codebook en meme temps et en pratique a l’utilisation, on map chaque token au token le plus proche si besoin)

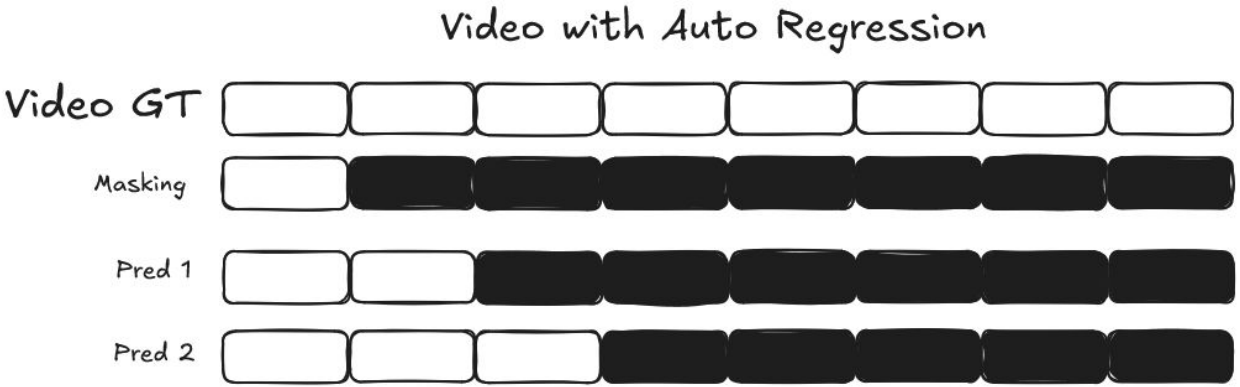
Sémantique = Contient des informations intéressantes facilement exploitables (et on check ça avec une batterie de tests définis en avance)

Voir Références a la fin

Encoding/Decoding

Model 1 | Model 2 | Model 3 | Ground Truth

Prediction



Avantages / Inconvénients

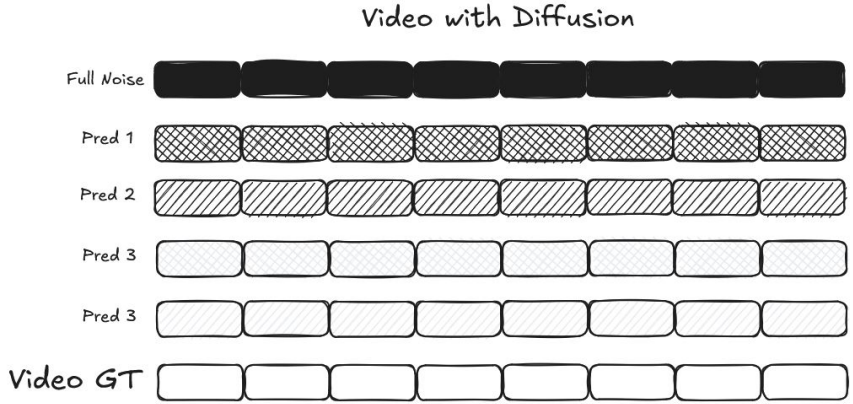
- 1 bloc à la fois
- Cohérence Globale ?
- Drift: on fait une prédiction sur une prédiction

- On a besoin d'avoir des tokens discrets (permet de mitiger le drift)

- Rapide
- C'est un LLM, Mistral sait probablement faire ça et on est en France, donc on est content
- Probablement d'autres choses...

Voir Références a la fin

Prediction



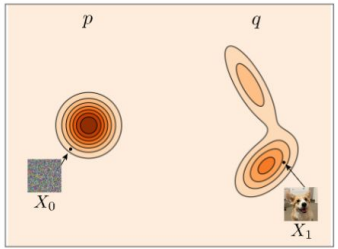
Avantages / Inconvénients

- Cohérence Globale
- C’est un processus dans un espace continu (a priori output de qualité)
- Tout le monde fait ça en image / video donc on s’aligne là-dessus

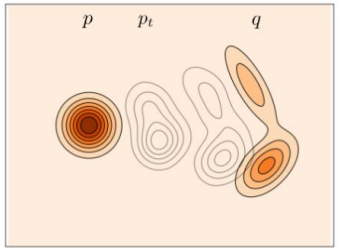
- Vachement Lent
- Plus difficile à entrainer qu’un LLM
- On fera quand même de l’auto régressif pour avoir un long contexte

Voir Références a la fin

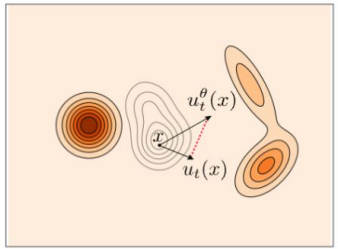
Diffusion / Flow Models



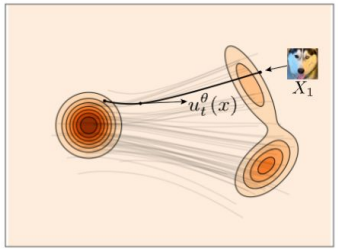
(a) Data.



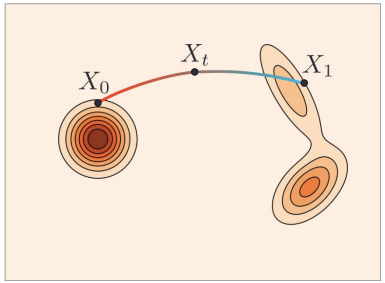
(b) Path design.



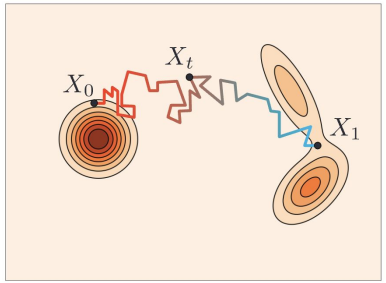
(c) Training.



(d) Sampling.



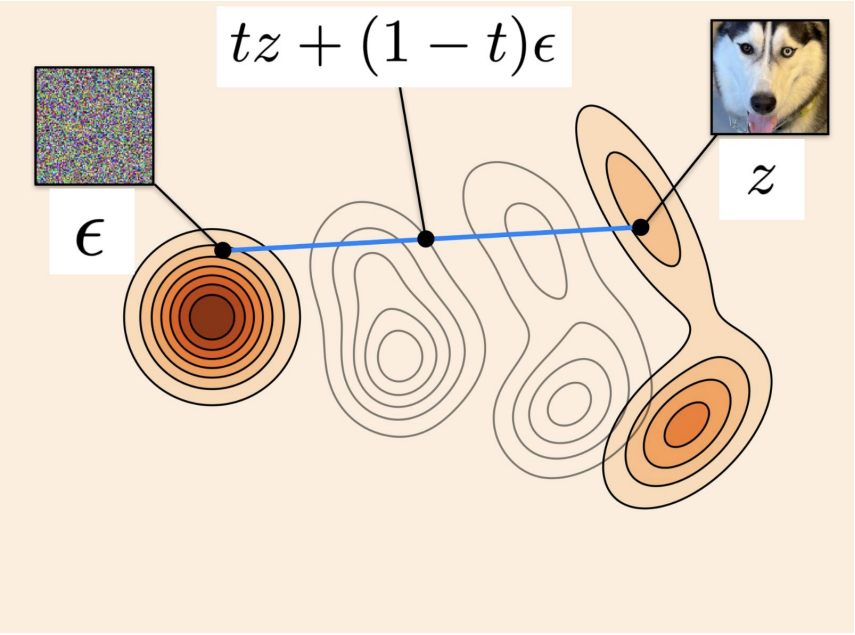
(a) Flow



(b) Diffusion

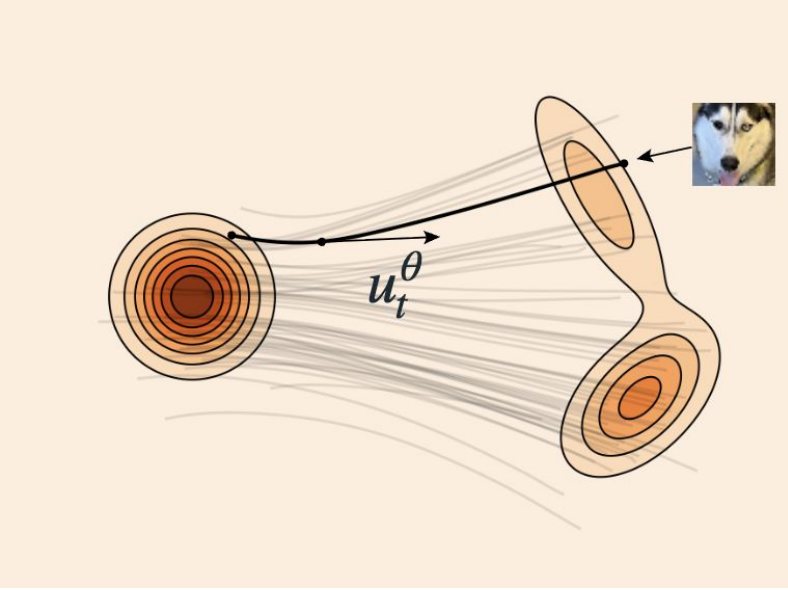
Diffusion / Flow Models

TRAINING:

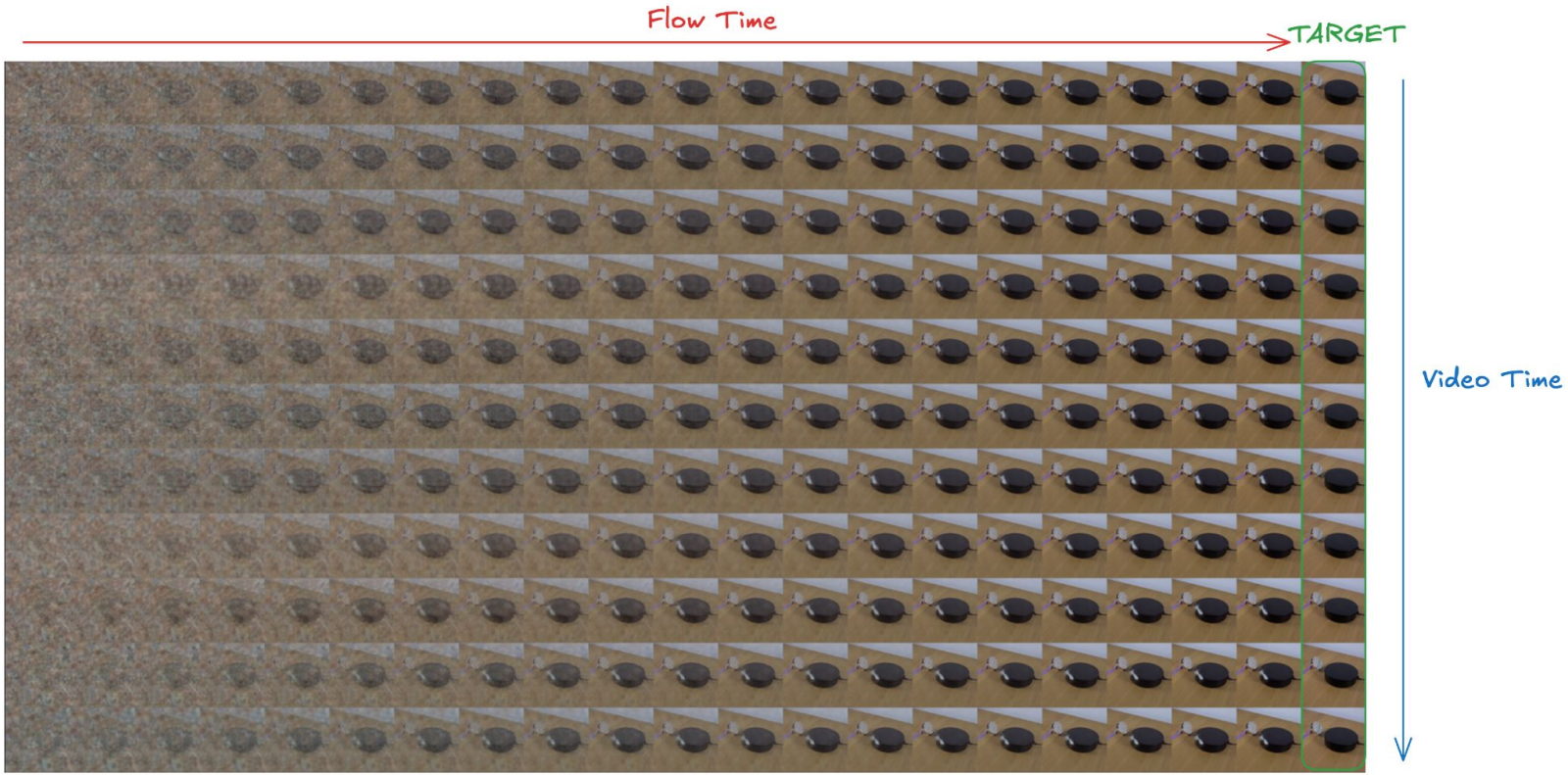


SAMPLING:

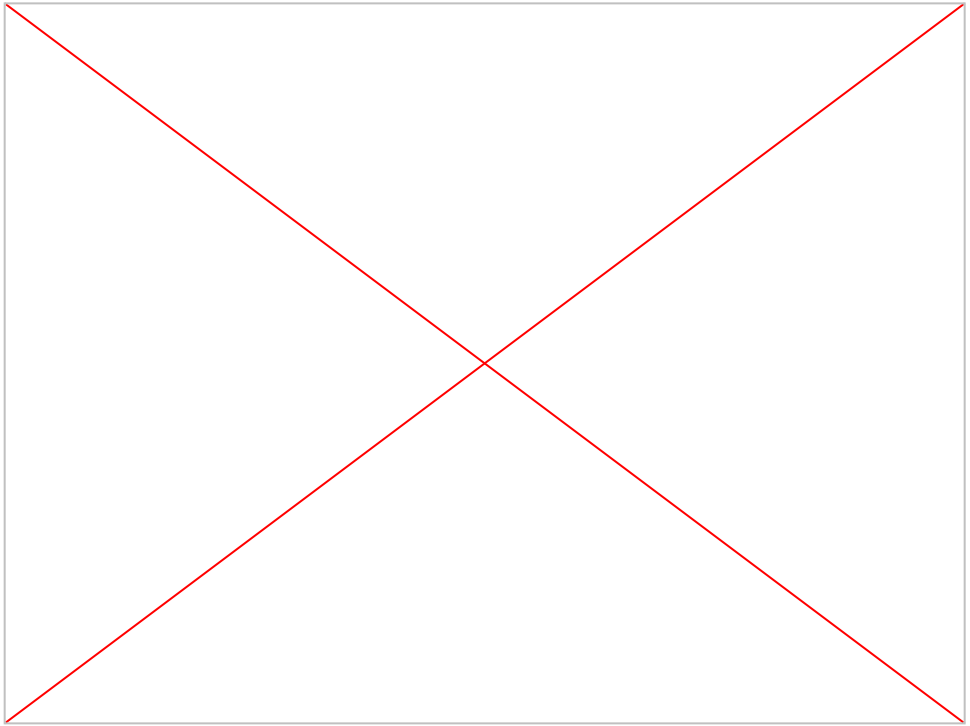
$$x_{t+\Delta t} = x_t + \Delta t \cdot \left.\frac{dx}{dt}\right|_{x_t,t}$$



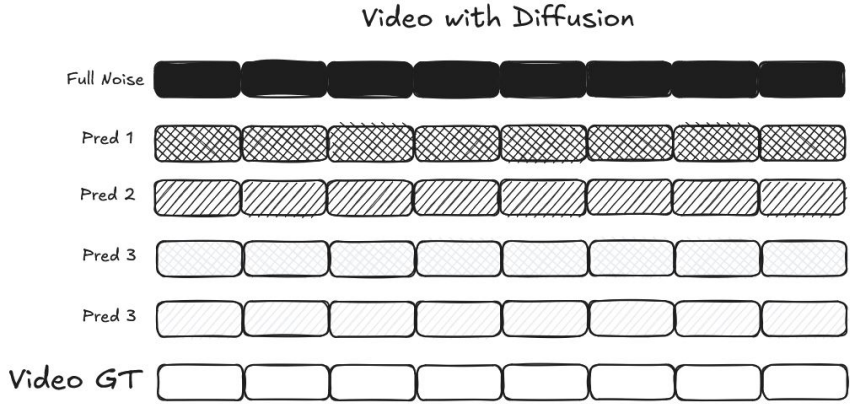
Diffusion / Flow Models for videos



Diffusion models for text generation



Prediction



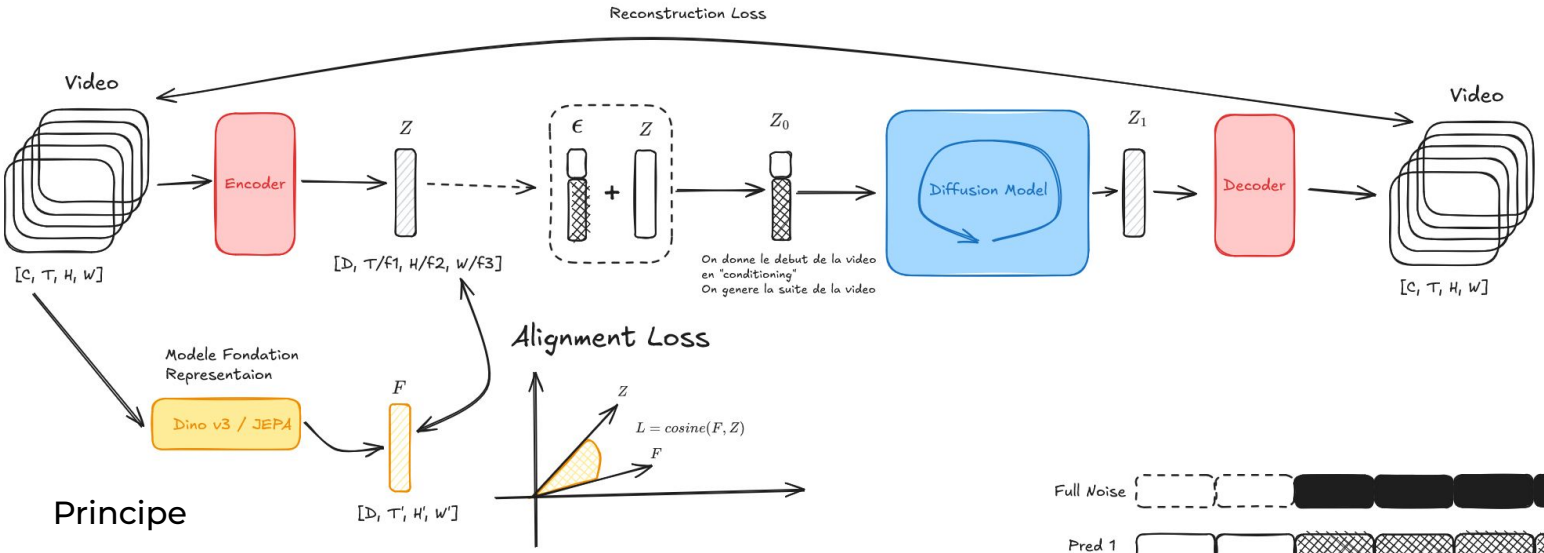
Avantages / Inconvénients

- Cohérence Globale
- C’est un processus dans un espace continu (a priori output de qualité)
- Tout le monde fait ça en image / video donc on s’aligne là-dessus

- Vachement Lent
- Plus difficile à entrainer qu’un LLM
- On fera quand même de l’auto régressif pour avoir un long contexte

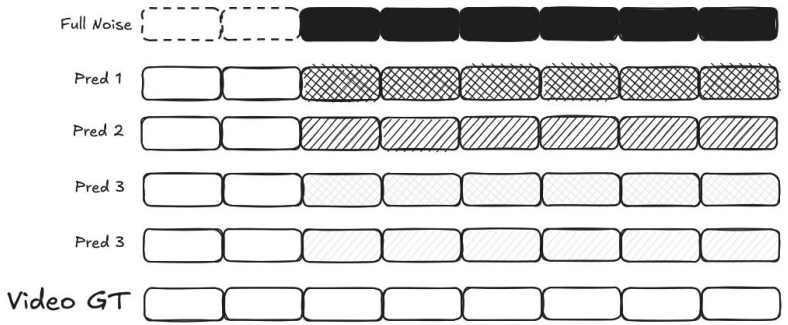
Voir Références a la fin

Mon Stage



Principe

- On veut générer la suite d'une vidéo étant donné X secondes observées avant
- On force notre encoder à avoir une "certaine représentation" (issue d'un autre modèle)
- On optimise quand même pour de la compression élevée
- On regarde si ça va plus vite pour apprendre à générer une vidéo (spoiler OUI)



Quelques References

- CS280 Berkley: https://cs280-berkeley.github.io/lectures/lect11.pdf
- Google Blog: https://diffusionflow.github.io/
- ICLR 2025 - Inria Blog: https://dl.heeere.com/conditional-flow-matching/blog/conditional-flow-matching/
- 3b1b & Welch Lab: https://www.youtube.com/watch?v=iv-5mZ_9CPY
- Lil'log: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Stable Diffusion 1, XL, 3 (Robin Rombach)
Flux, Black Forest Lab (Robin Rombach)
Flow Matching (Yaron Lipman)
Gaia 2 (Wayve) -> en gros mon stage pour les curieux