Specializing
Video Diffusion Models

Kashyap Chitta
PhD Student, Autonomous Vision Group

Tübingen AI Center
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Video Latent Diffusion

Where are we?

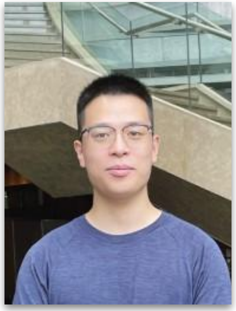
Building Vista

Can we specialize SVD for driving?

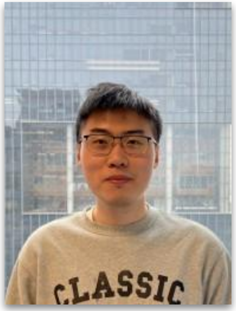
Practical Tips

What matters most during training?

Team



Shenyuan Gao

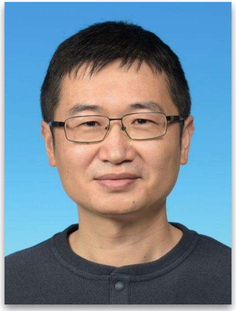


Jiazhi Yang

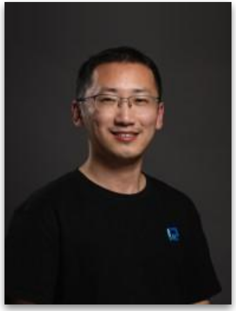
Li Chen

Kashyap Chitta

Yihang Qiu



Jun Zhang



Hongyang Li

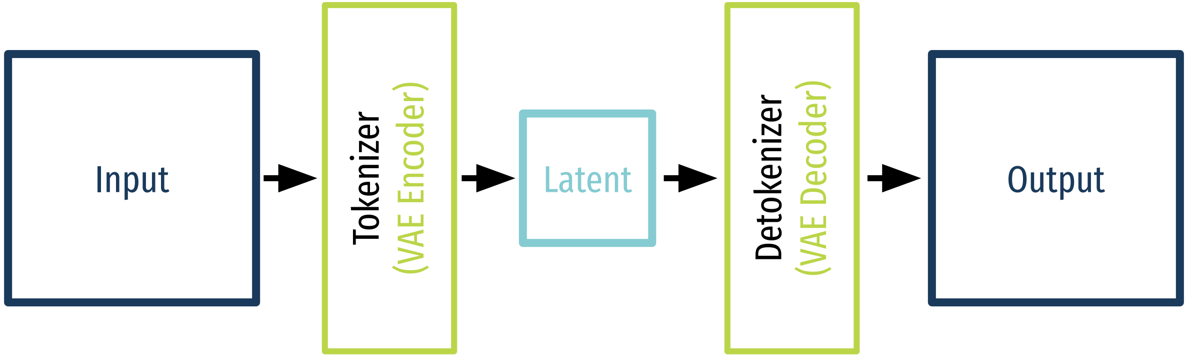
Andreas Geiger

Video Latent Diffusion

Where are we?

The tried and tested LDM recipe

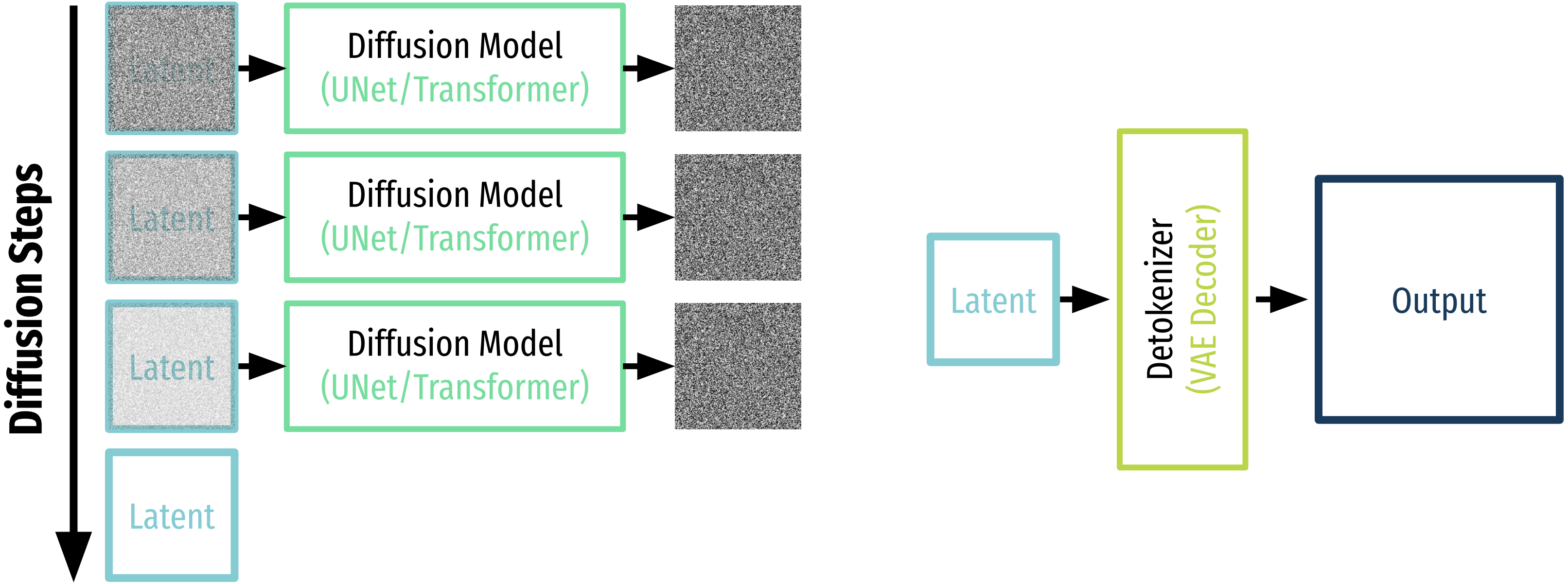
Step 1: Autoencoder with fixed-size latent



Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models." CVPR, 2022.

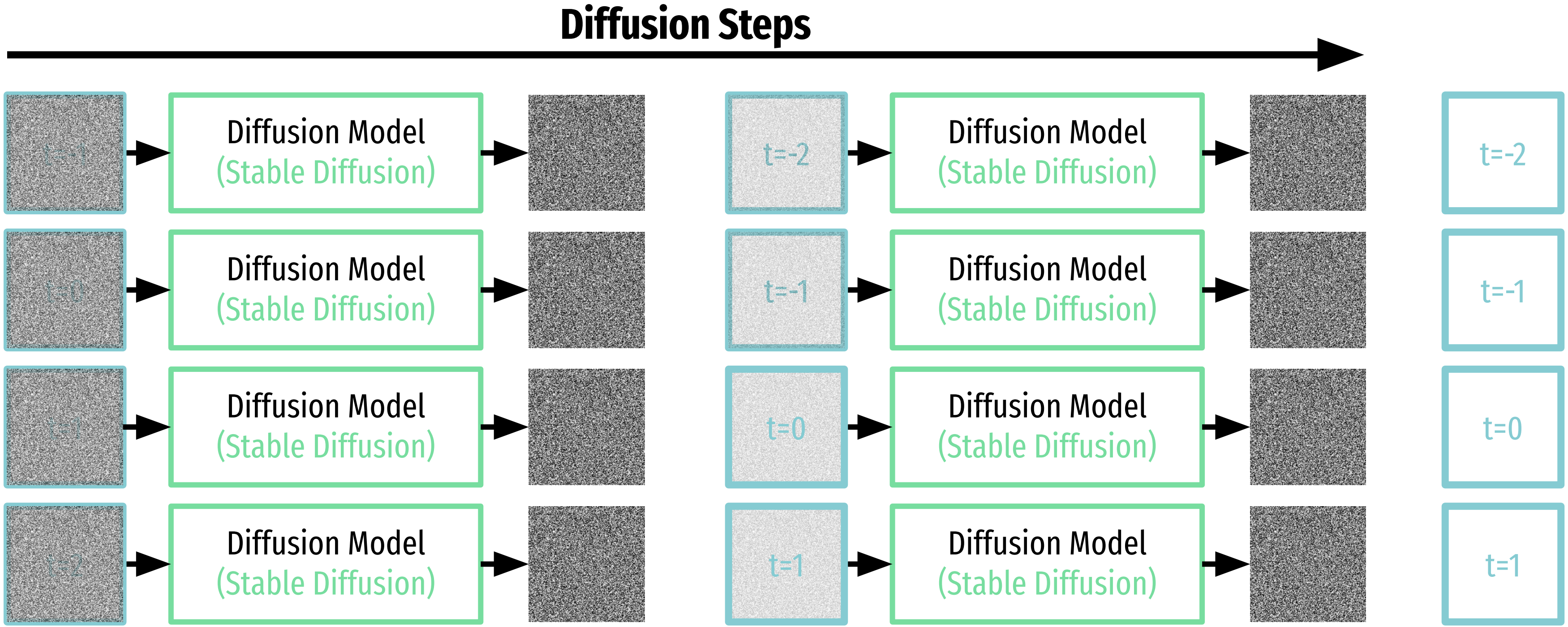
The tried and tested LDM recipe

Step 2: Latent denoiser



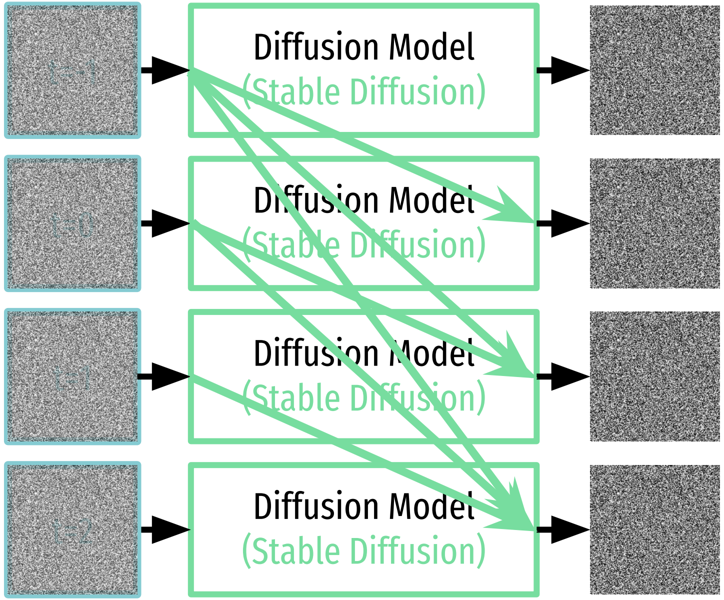
Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models." CVPR, 2022.

Video LDM by ‘aligning your latents’



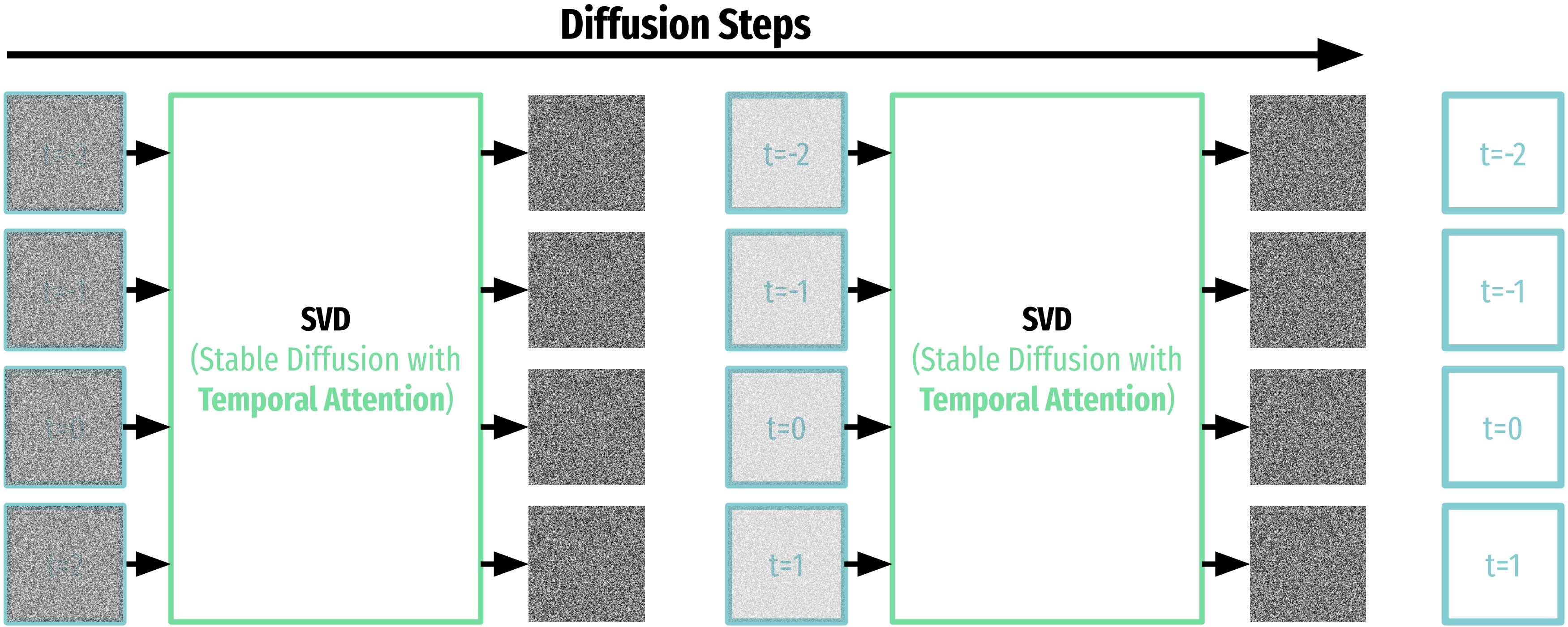
Blattmann et al. "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models." CVPR, 2023.

Video LDM by ‘aligning your latents’



Blattmann et al. "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models." CVPR, 2023.

Stable Video Diffusion: temporal attention blocks



Blattmann et al. "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets." ArXiv, 2023.

Building Vista

Can we specialize SVD for driving?

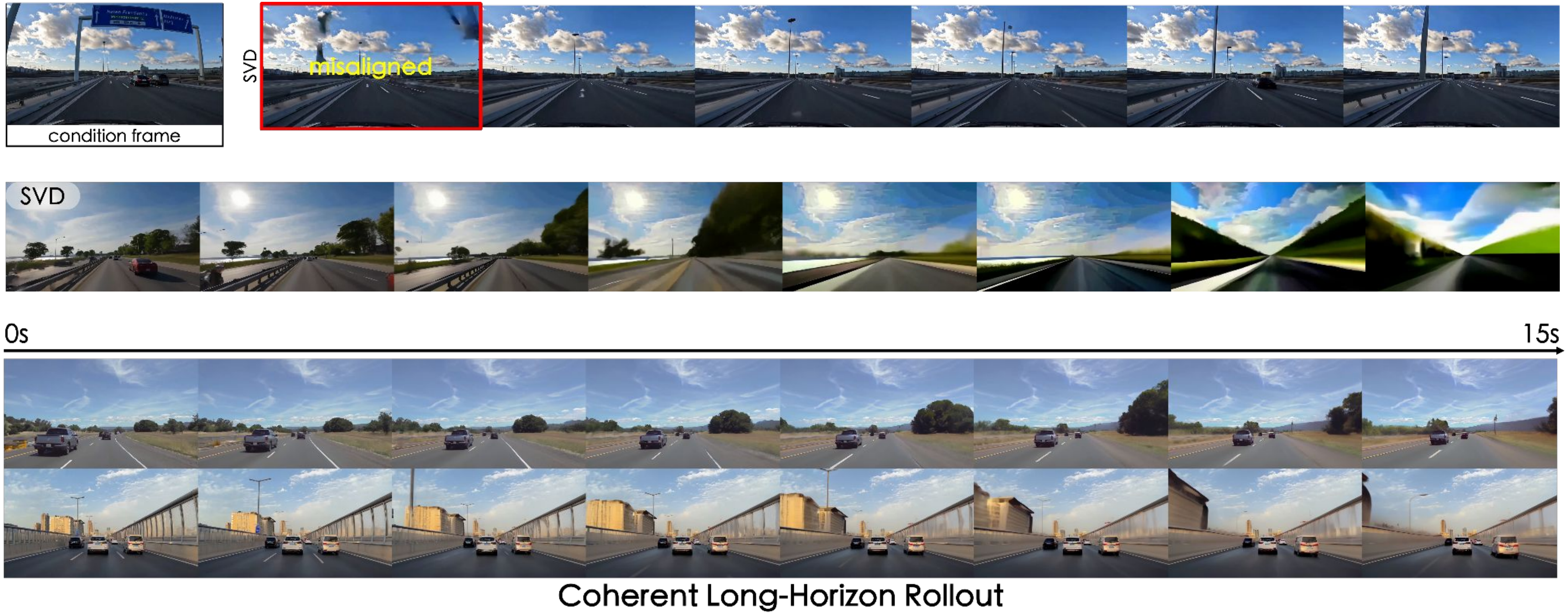
Is SVD enough? Not for long rollouts

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Is SVD enough? Not for long rollouts

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Is SVD enough? Not for long rollouts



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

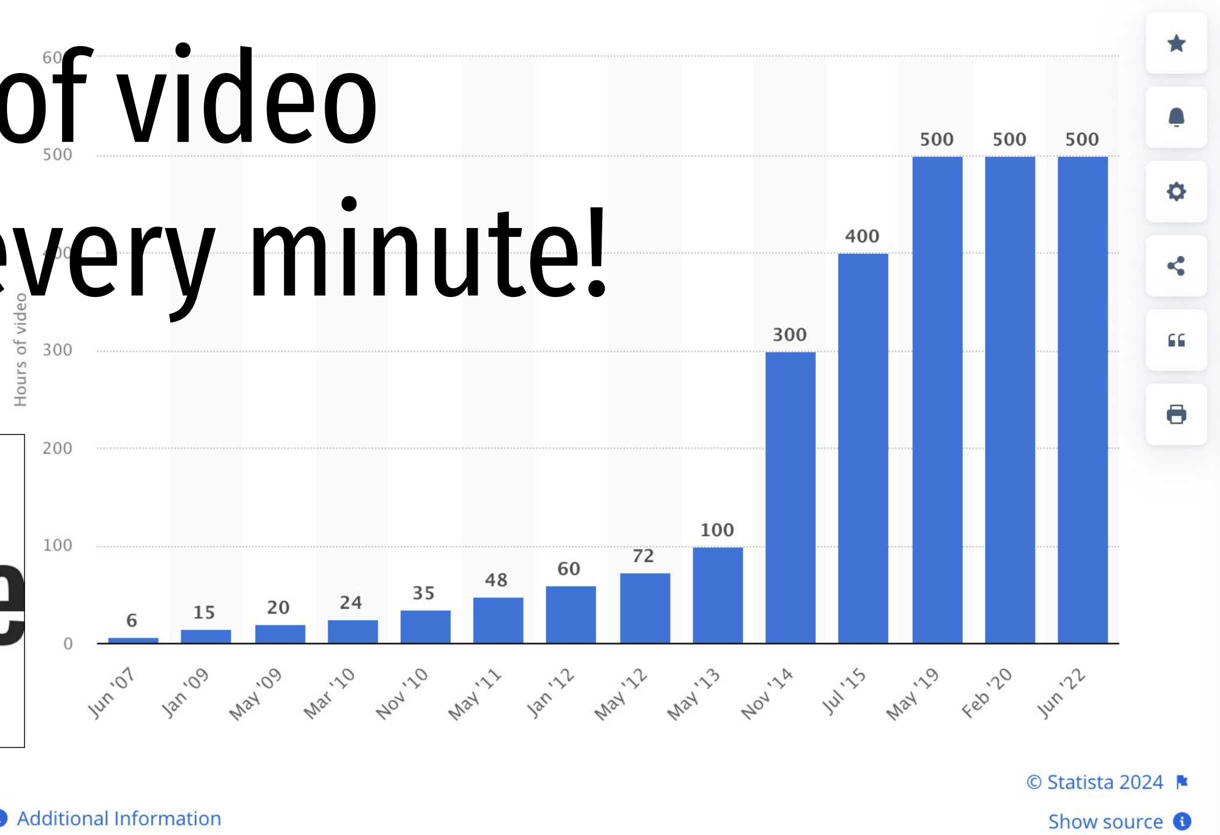
First problem: datasets lacking diversity and scale





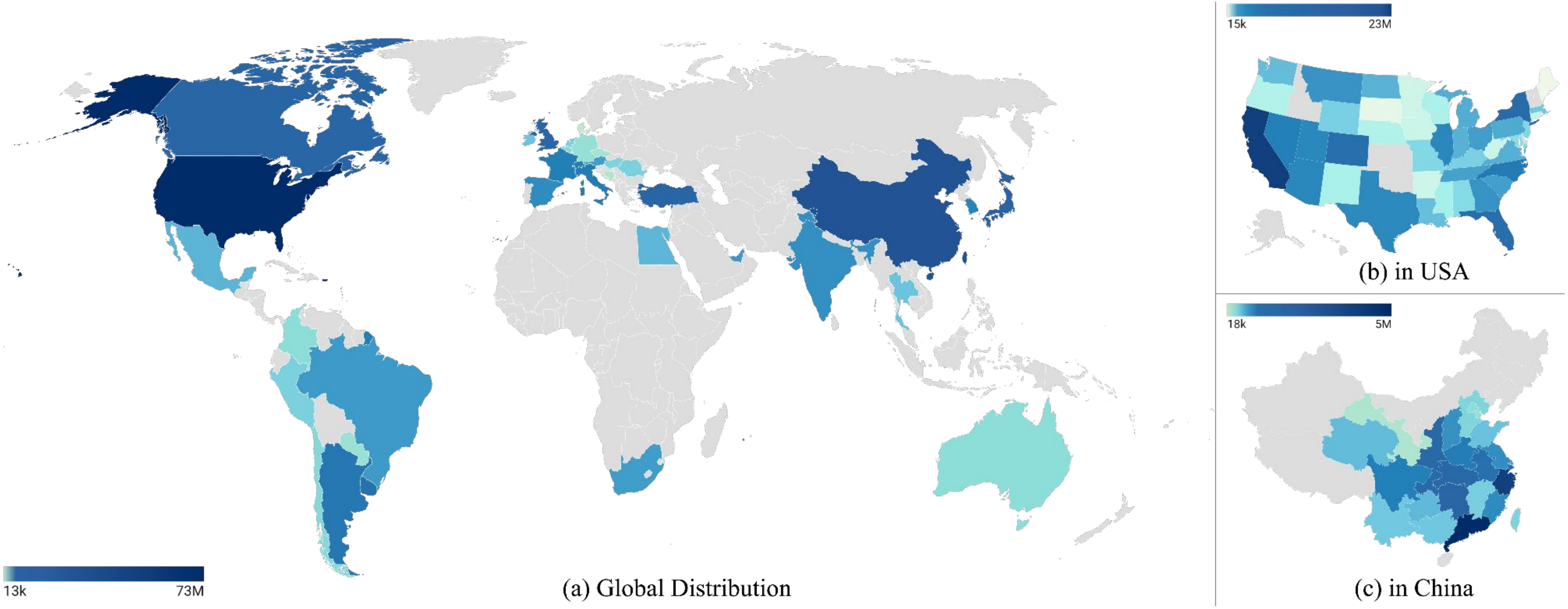

Karnchanachari et al. "Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving." ICRA, 2024.

500 hours of video uploaded every minute!




OpenDV-2k
Wind Walk Travel Videos

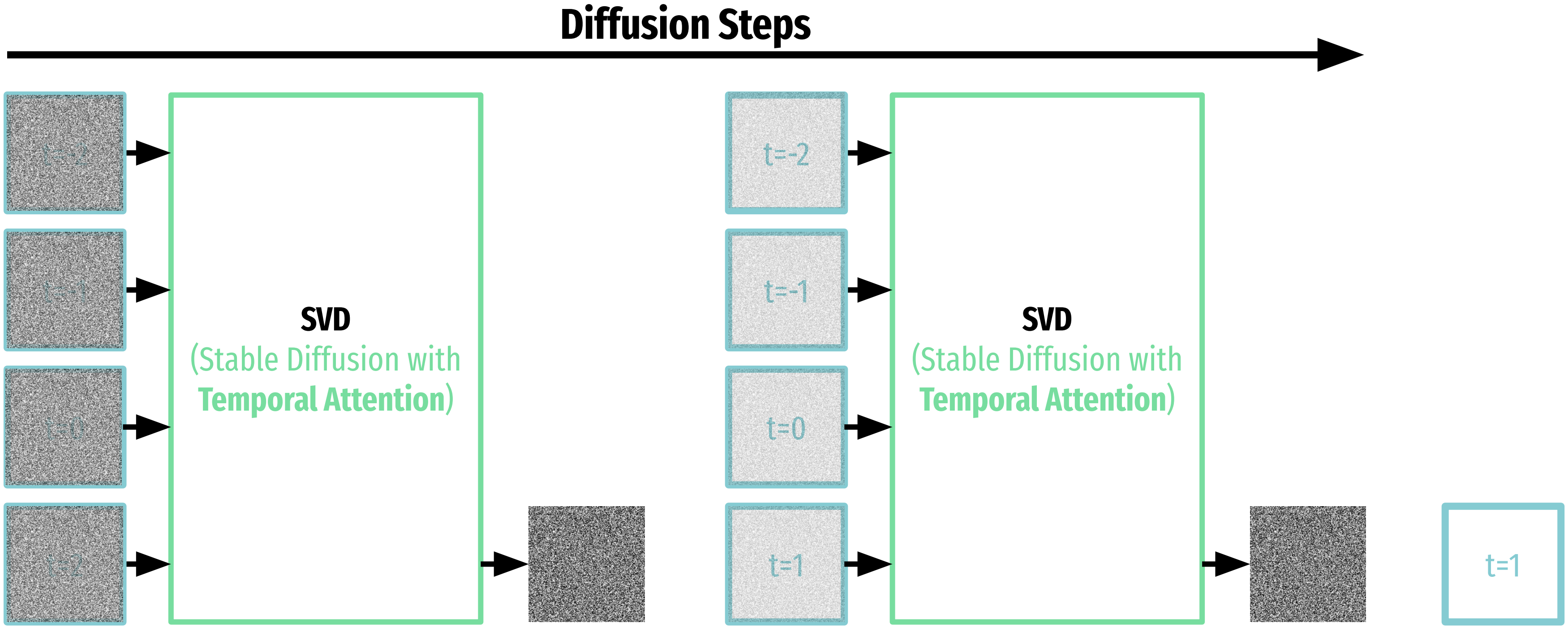
2000+ hours, 65M+ frames, 40+ countries, 700+ cities



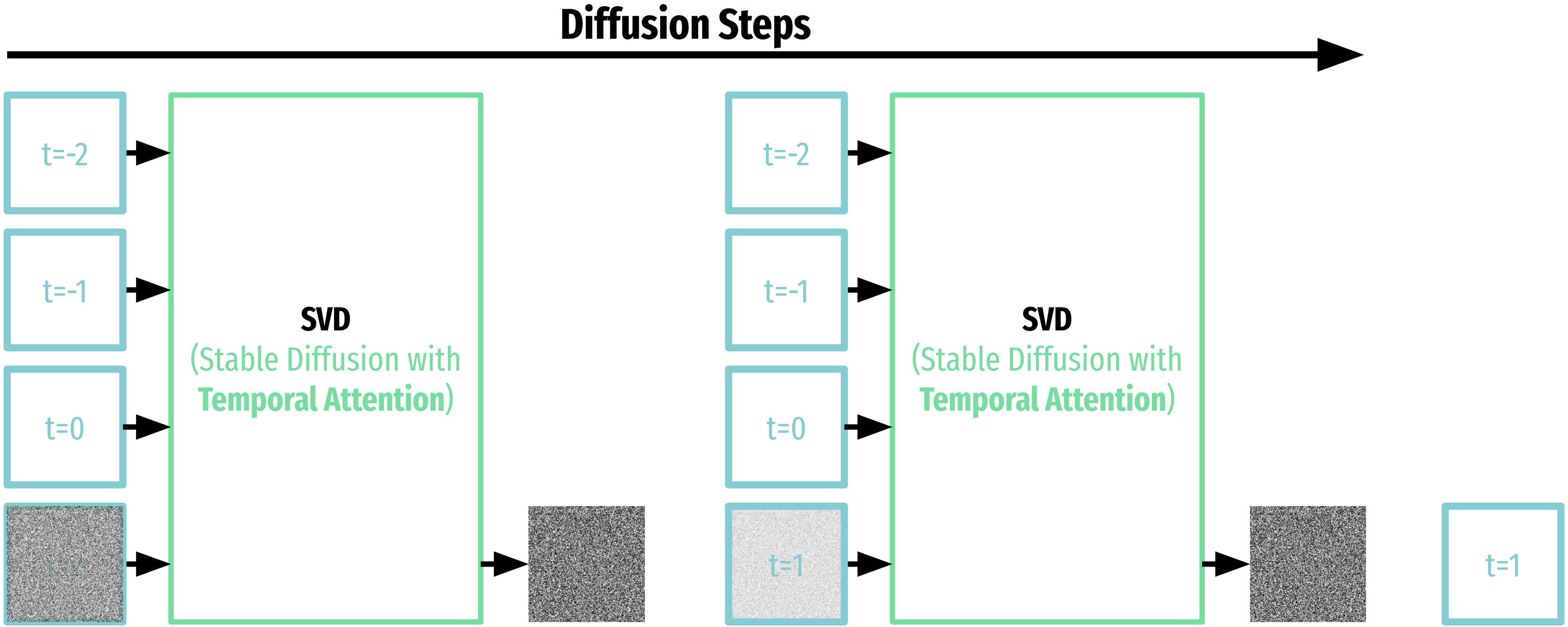
(a) Global Distribution (b) in USA (c) in China

Yang et al. “Generalized Predictive Model for Autonomous Driving.” CVPR, 2024

Adapting SVD for long rollouts: latent replacement



Adapting SVD for long rollouts: latent replacement

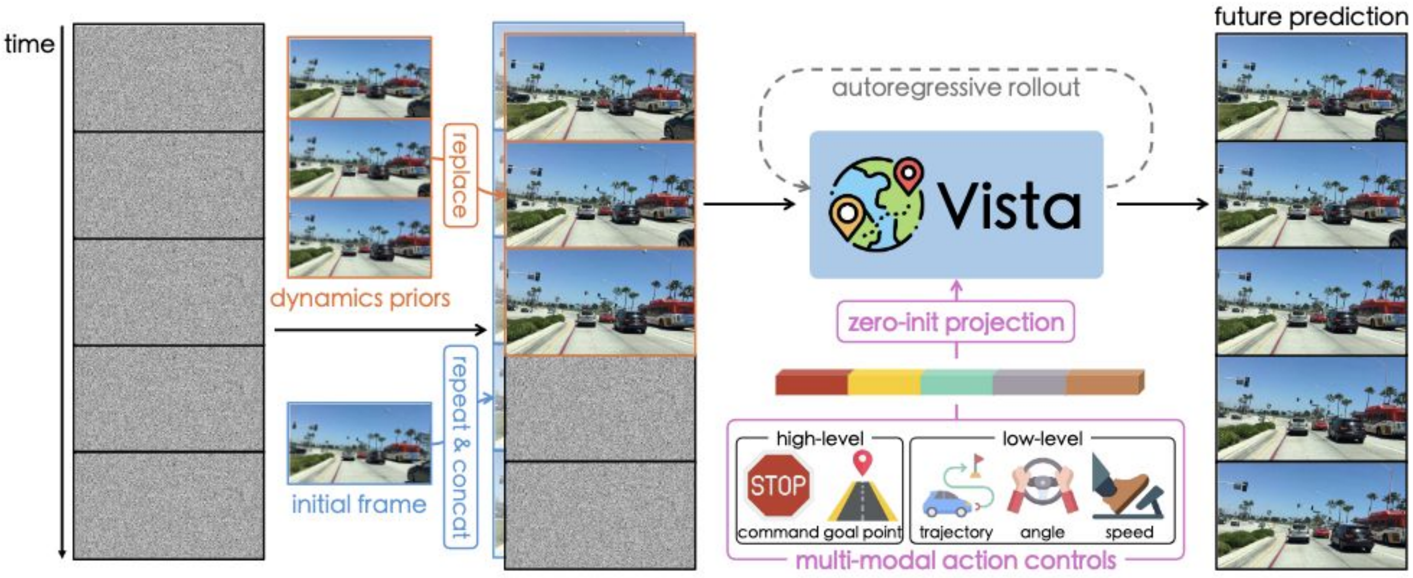


Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Adapting SVD for long rollouts: latent replacement

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Adapting SVD for versatile controllability: zero-init projections



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Adapting SVD for versatile controllability: zero-init projections

Turn left | Go straight | Turn right | Stop

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

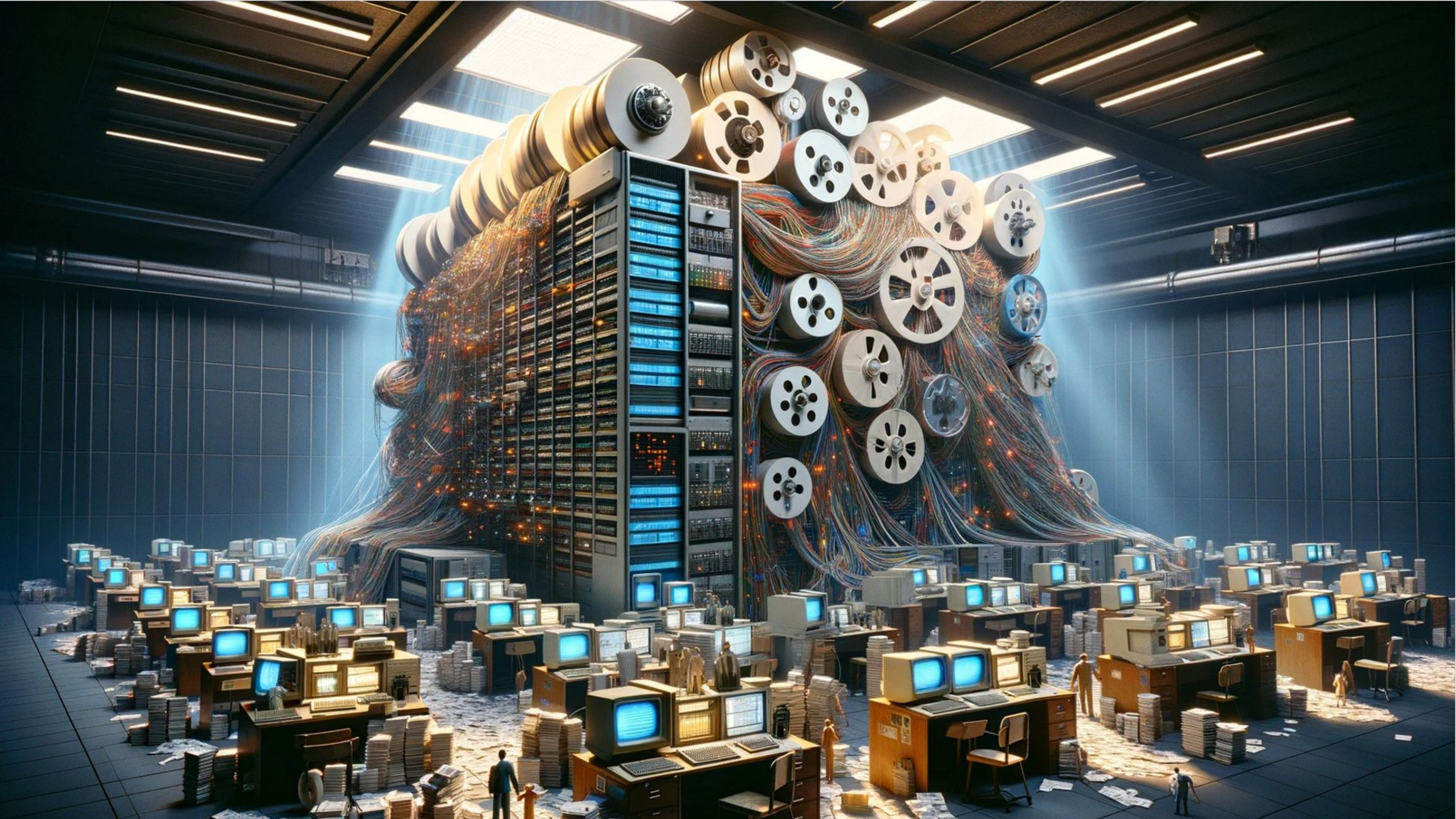
Adapting SVD for versatile controllability: zero-init projections

Turn left | Go straight | Turn right | Stop

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Practical Tips

What matters most during training?



ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION

Diederik P. Kingma*
University of Amsterdam, OpenAI
dpkingma@openai.com

Jimmy Lei Ba*
University of Toronto
jimmy@psi.utoronto.ca

7.2 TEMPORAL AVERAGING

Since the last iterate is noisy due to stochastic approximation, better generalization performance is often achieved by averaging. Previously in Moulines & Bach (2011), Polyak-Ruppert averaging (Polyak & Juditsky, 1992; Ruppert, 1988) has been shown to improve the convergence of standard SGD, where $\bar{\theta}_t = \frac{1}{t}\sum_{k=1}^{n}\theta_k$. Alternatively, an exponential moving average over the parameters can be used, giving higher weight to more recent parameter values. This can be trivially implemented by adding one line to the inner loop of algorithms 1 and 2: $\bar{\theta}_t \leftarrow \beta_2 \cdot \bar{\theta}_{t-1} + (1-\beta_2)\theta_t$, with $\bar{\theta}_0 = 0$.

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

EMA has a huge memory overhead but is essential

Phase 1: 100% OpenDV-YouTube

- Resource-intensive (128 x A100, 8 days)
- All 1.7B UNet params

Without EMA:

- Batch size 1 per 80GB A100 possible
- **But validation FID worsens over training!**

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

EMA has a huge memory overhead but is essential

Phase 1: 100% OpenDV-YouTube

- Resource-intensive (128 x A100, 8 days)
- All 1.7B UNet params

With EMA:

- Batch size 1 per 80GB A100 not possible!
- EMA requires 11GB additional memory per GPU
- Training possible, but slower: **need gradient accumulation**

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Offset noise improves temporal consistency

RESEARCH

Diffusion with Offset Noise

Fine-tuning against a modified noise, enables Stable Diffusion to generate very dark or light images easily.

By Nicholas Guttenberg | January 30, 2023

In code terms, the current training loop uses noise that looks like:

```
noise = torch.randn_like(latents)
```

But instead, I could use something like this:

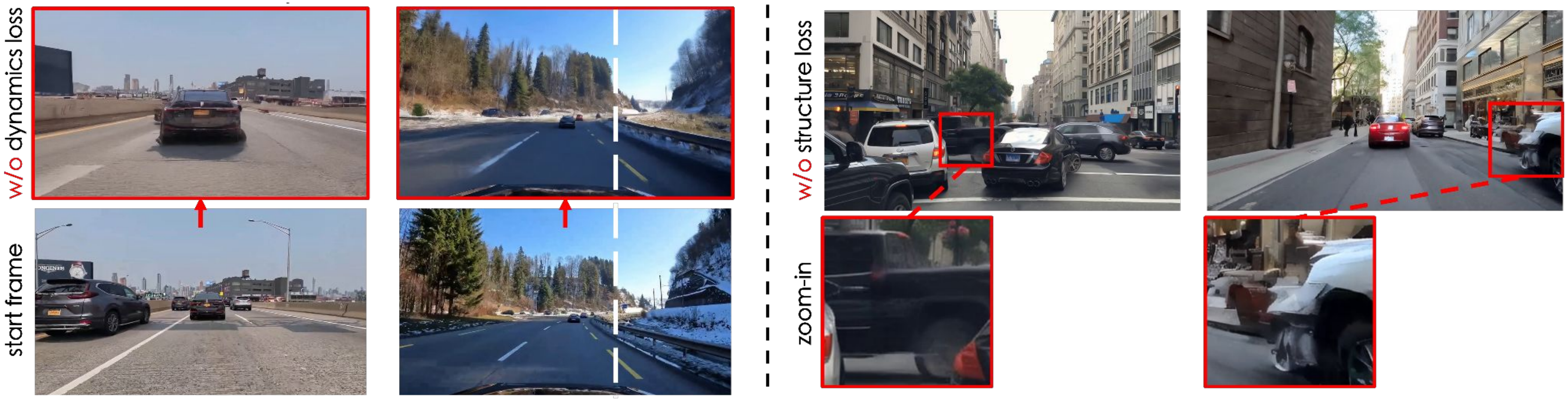
```
noise = torch.randn_like(latents) + 0.1 * torch.randn(latents.shape[0],
latents.shape[1], 1, 1)
```

https://www.crosslabs.org/blog/diffusion-with-offset-noise

Offset noise improves temporal consistency

https://www.crosslabs.org/blog/diffusion-with-offset-noise

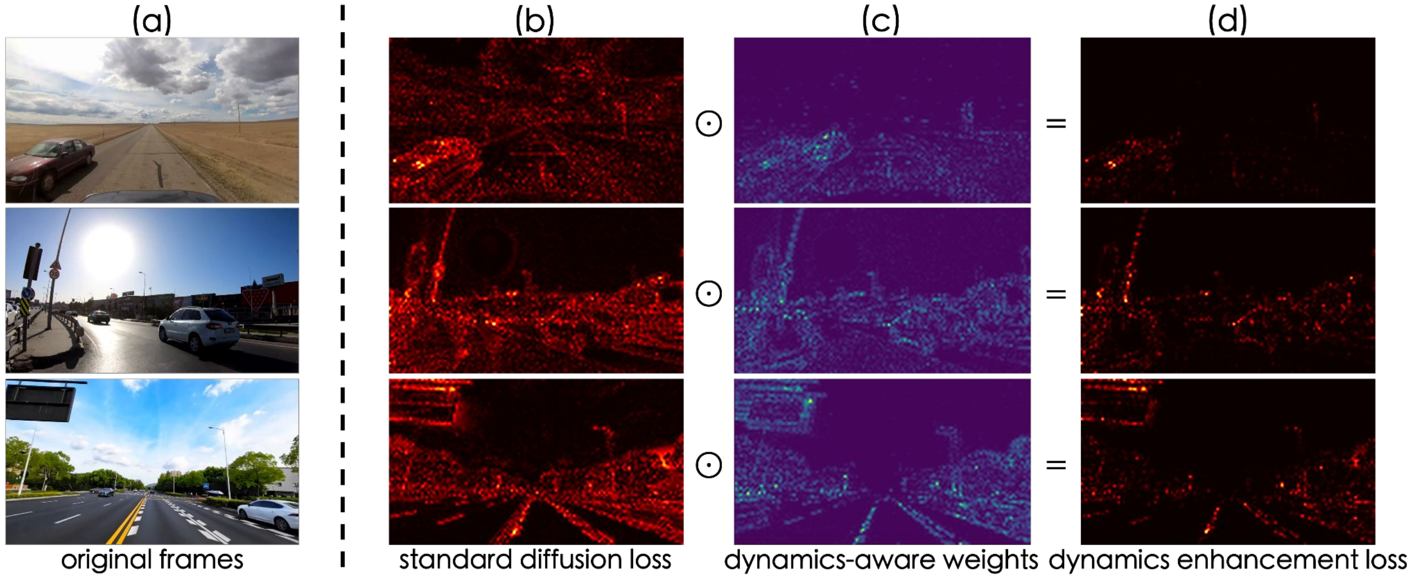
Domain-specific loss weights may be necessary



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Domain-specific loss weights may be necessary

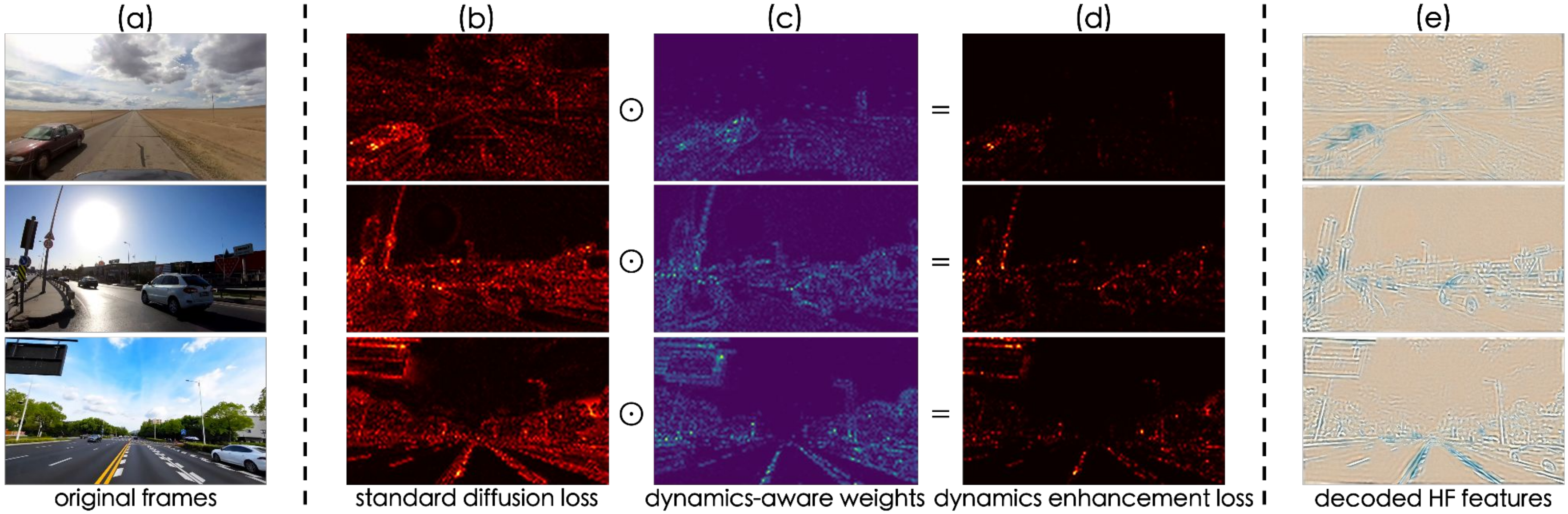
$$w_i = \left\|(D_\theta(\hat{n}_i;\sigma) - D_\theta(\hat{n}_{i-1};\sigma)) - (z_i - z_{i-1})\right\|^2$$



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

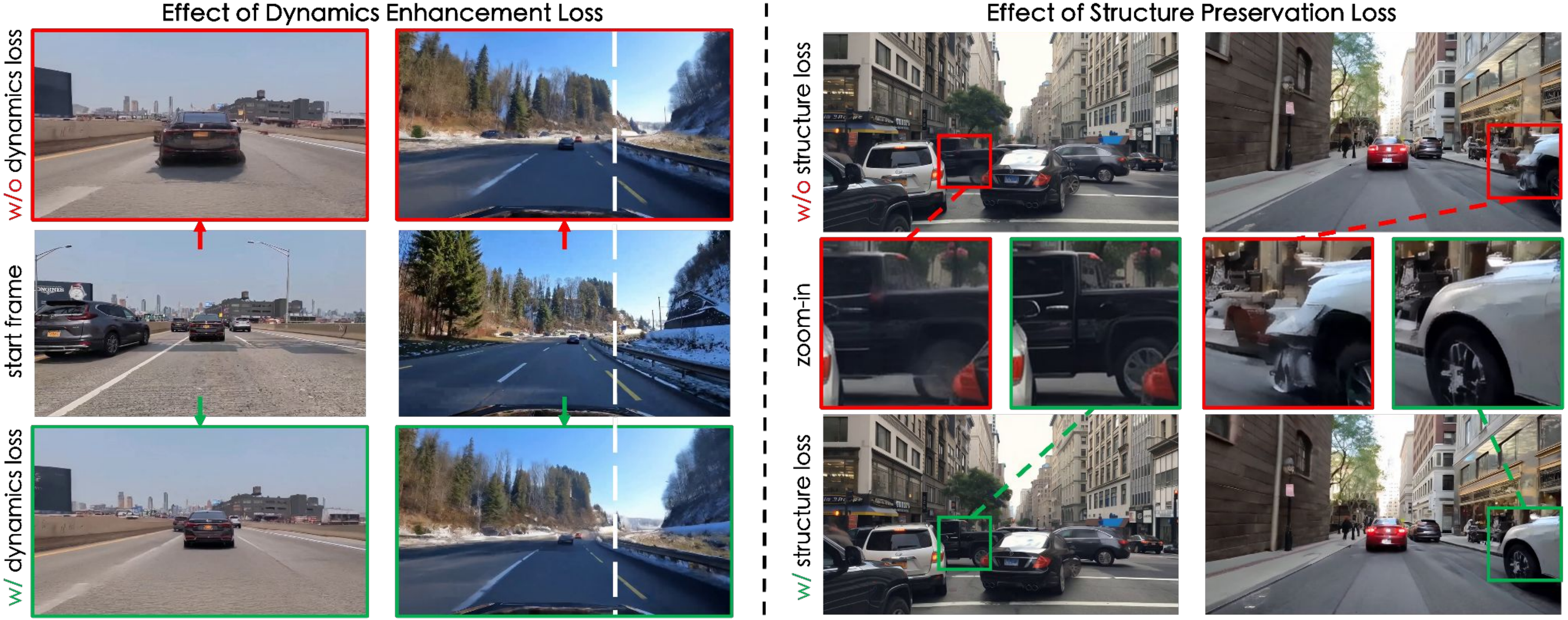
Domain-specific loss weights may be necessary

$$z_i' = \mathcal{F}(z_i) = \mathtt{IFFT}\big(\mathcal{H} \odot \mathtt{FFT}(z_i)\big)$$



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Domain-specific loss weights may be necessary



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Iters/sec is the most important factor to scale

Phase 1: 100% OpenDV-YouTube

- Resource-intensive (128 x A100, 8 days)
- All 1.7B UNet params

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

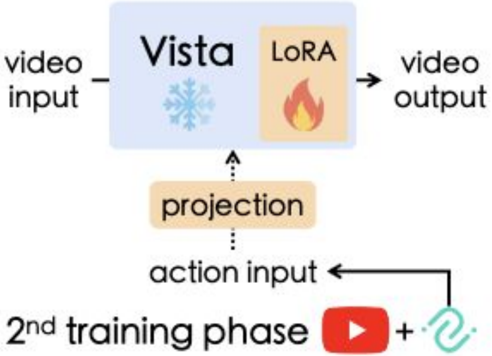
Iters/sec is the most important factor to scale

Phase 1: 100% OpenDV-YouTube

- Resource-intensive (128 x A100, 8 days)
- All 1.7B UNet params

Phase-2: 50% OpenDV-YouTube, **50% nuScenes**

- **Low-res stage:** 320 x 576 (8 x A100, 8 days)
  - 3.5x batch size, and **more iters/sec!**
  - **But doesn't speed up convergence!**
  - LoRA + action projection params
- **High-res stage:** 576 x 1024, (8 x A100, 2 days)



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Summary

#1 EMA has a huge memory overhead but is essential

#2 Offset noise improves temporal consistency

#3 Domain-specific loss weights may be necessary

#4 Iters/sec is the most important factor to scale

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

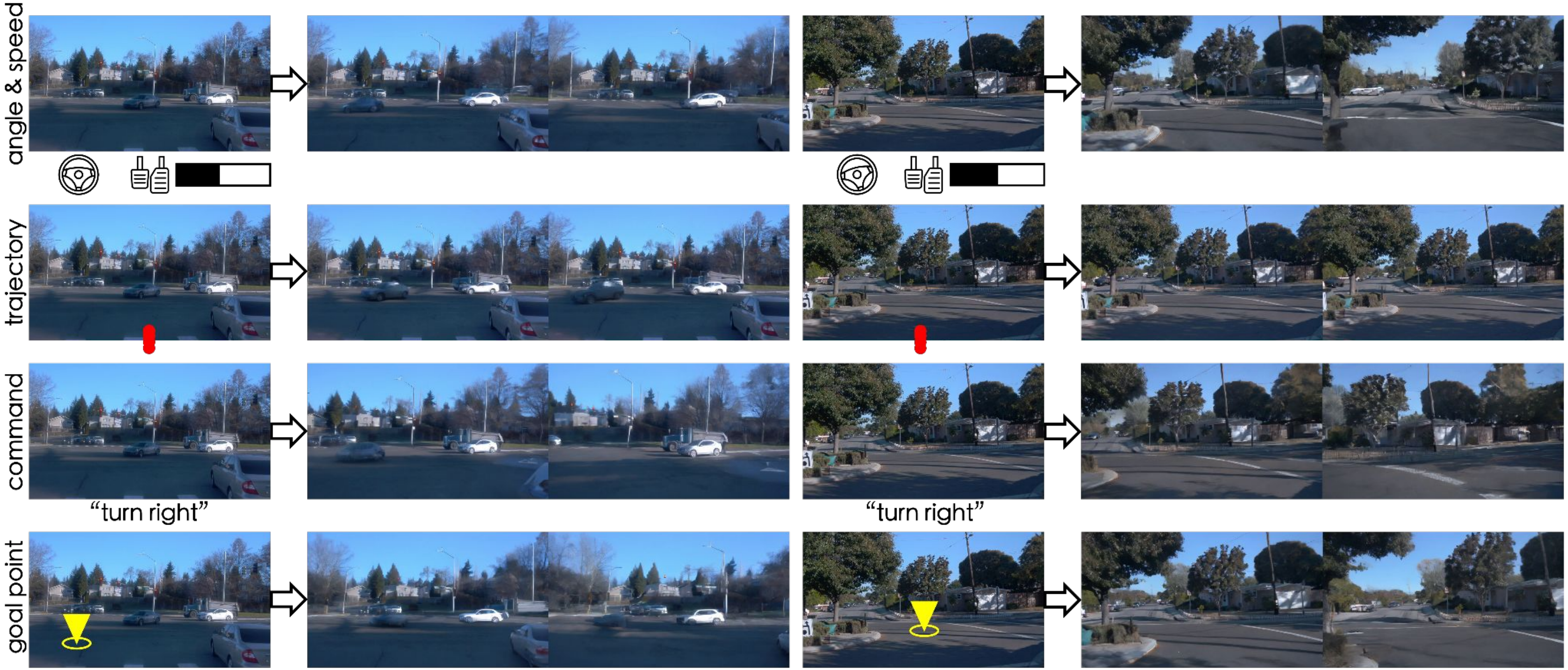
Vista has **open code and weights!**

vista-demo.github.io

Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

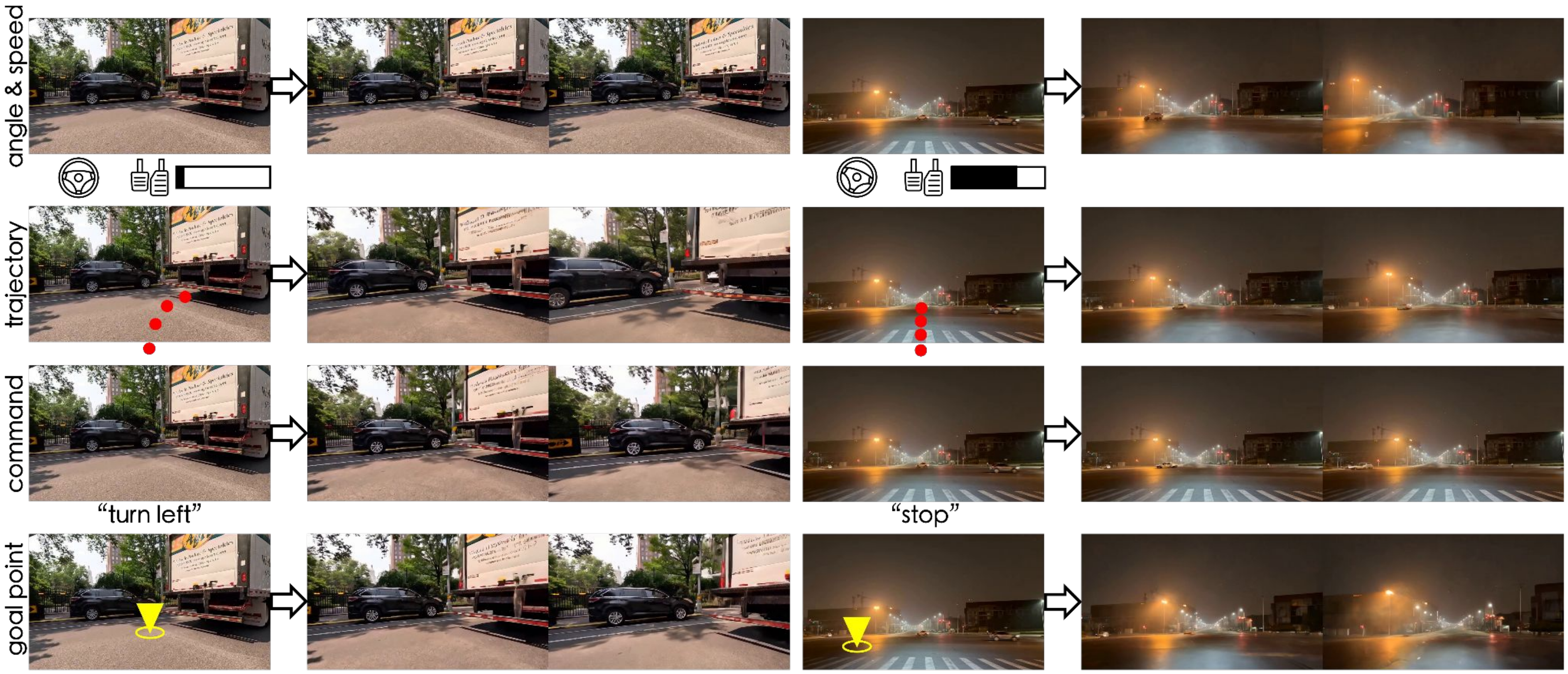
Extra Slides

Adapting SVD for versatile controllability: zero-init projections



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.

Adapting SVD for versatile controllability: zero-init projections



Gao et al. "Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability." NeurIPS, 2024.