Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling

Haoyu Wu*, Diankun Wu*, Tianyu He, Junliang Guo, Yang Ye,
Yueqi Duan, Jiang Bian
Microsoft Research, Tsinghua University, * Denotes equal contribution

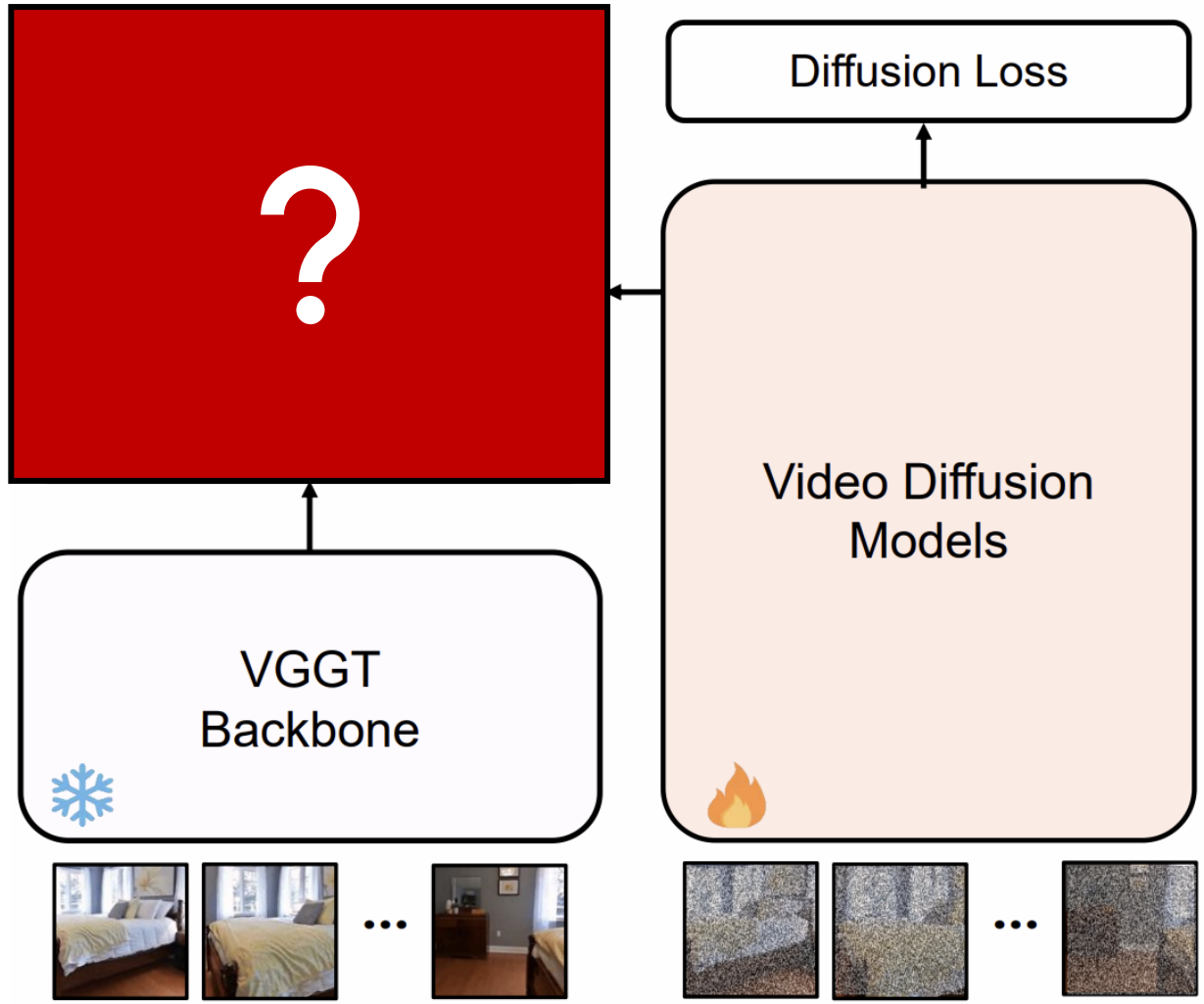
Problem Statement

- Videos are 2D projections of a 3D world.
- Video diffusion models are trained on raw 2D data.
- They fail to learn meaningful 3D geometric structure.
- Lack of 3D awareness leads to poor spatial consistency.

How to Solve This?

Geometry Forcing Solves This By:

- Injecting 3D geometric awareness into video models.



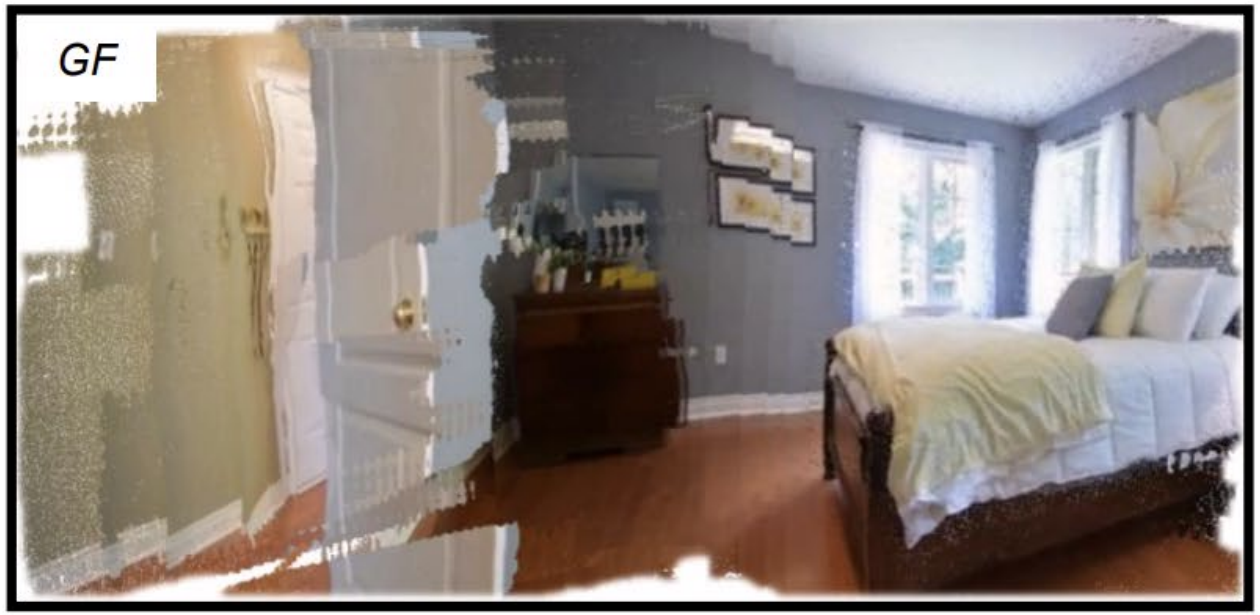
Video Generation Enhancement

Video Generation Enhancement

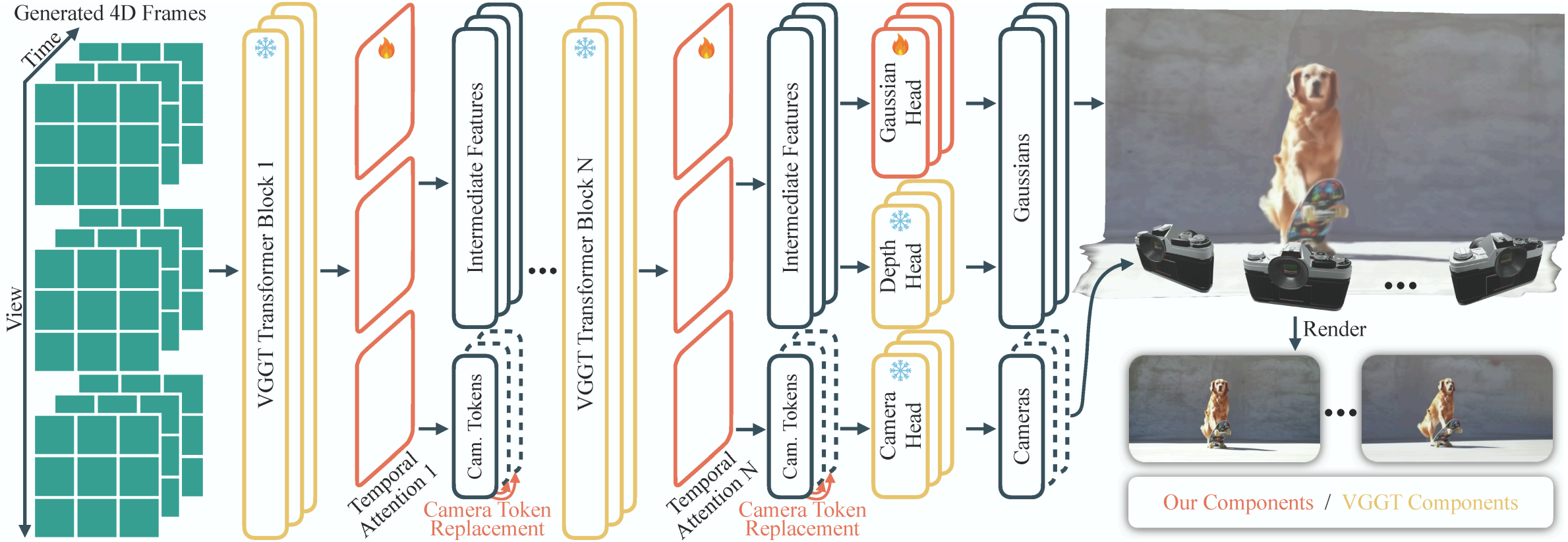
Baseline

Geometry Forcing

3D Reconstruction Enhancement



What We Covered Last Time: 4Real-Video-V2



Recap: VGGT

VGG Transformer

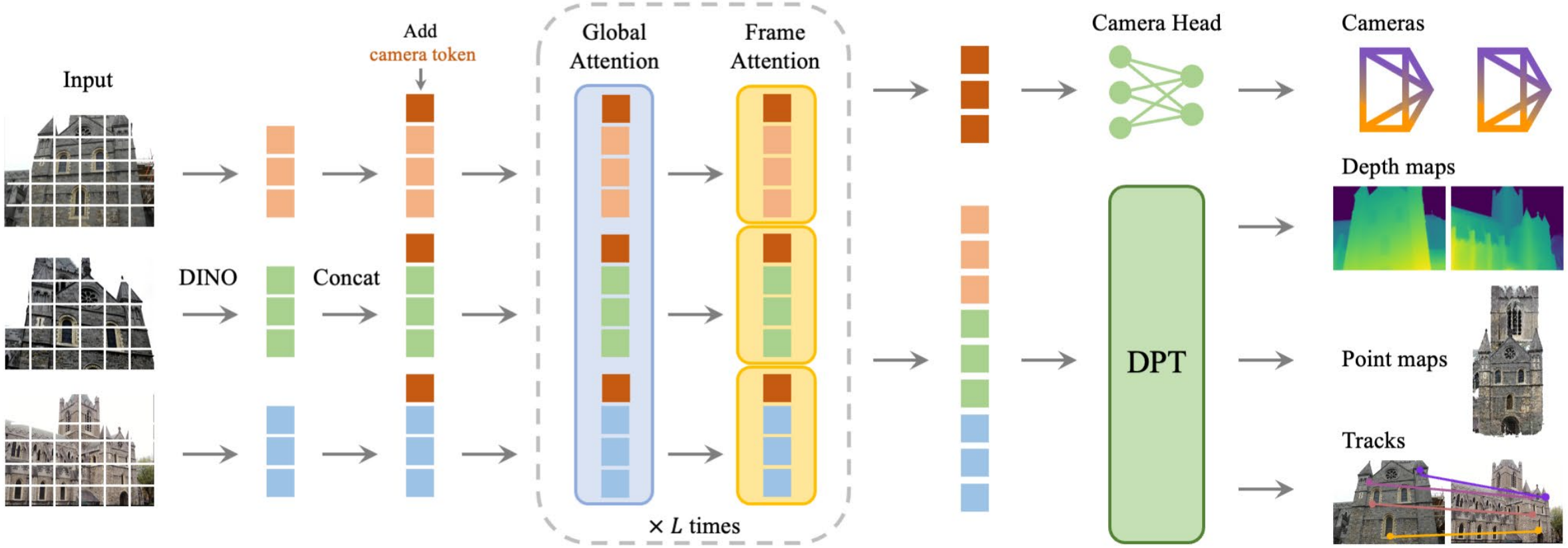
Images

VGGT

Reconstruction

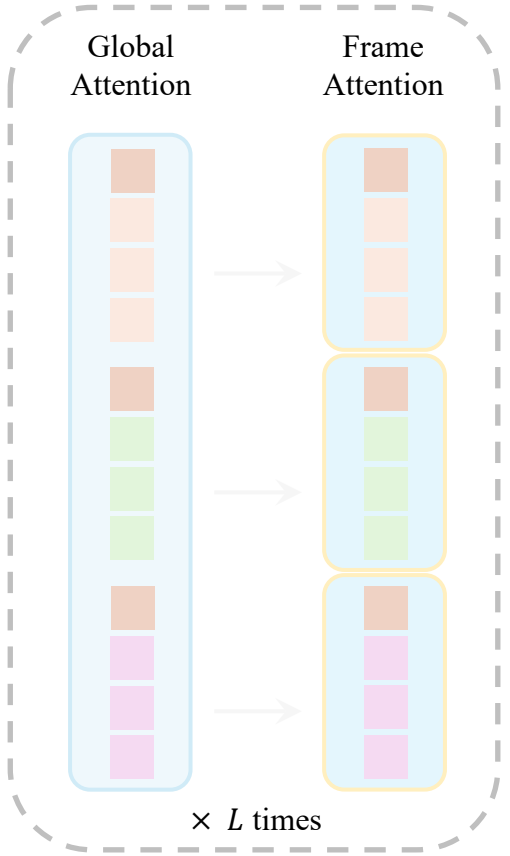
Cameras, Depths, Points, and Correspondences

VGG Transformer



Why Alternating-Attention?

- Global Attention
  - Ensures scene-level coherence
- Frame-wise Attention
  - Eliminates frame index embedding
    - For permutation equivariance
    - For flexible input length



How to Learn from a Foundation Model w/o Extra Inference Cost?

REPRESENTATION ALIGNMENT FOR GENERATION: TRAINING DIFFUSION TRANSFORMERS IS EASIER THAN YOU THINK

Sihyun Yu[1] **Sangkyung Kwak**[1,3] **Huiwon Jang**[1]
Jongheon Jeong[2] **Jonathan Huang**[3] **Jinwoo Shin**[1*] **Saining Xie**[4*]
[1]KAIST [2]Korea University [3]Scaled Foundations [4]New York University

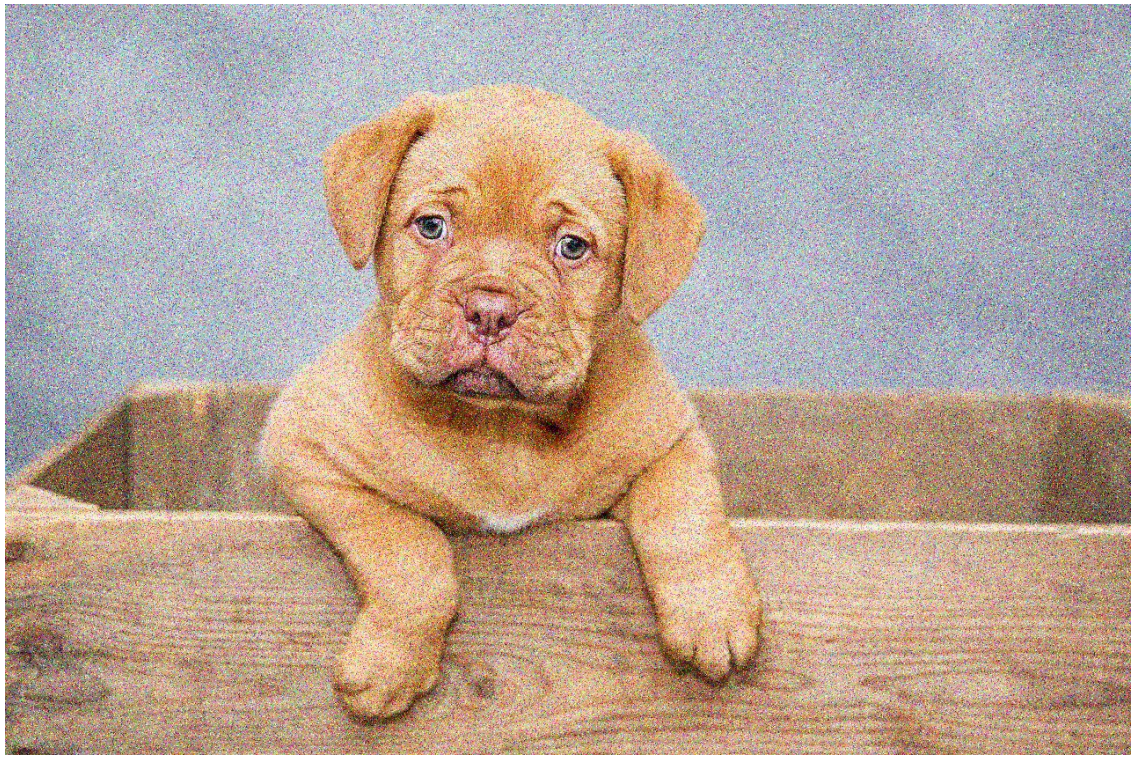
Diffusion Process

Diffusion Process

+ Noise

Diffusion Process

Denoising



Classifying

Denoising

Feature

Classifying

Denoising

Feature

Linear Layer

Classifying

Denoising

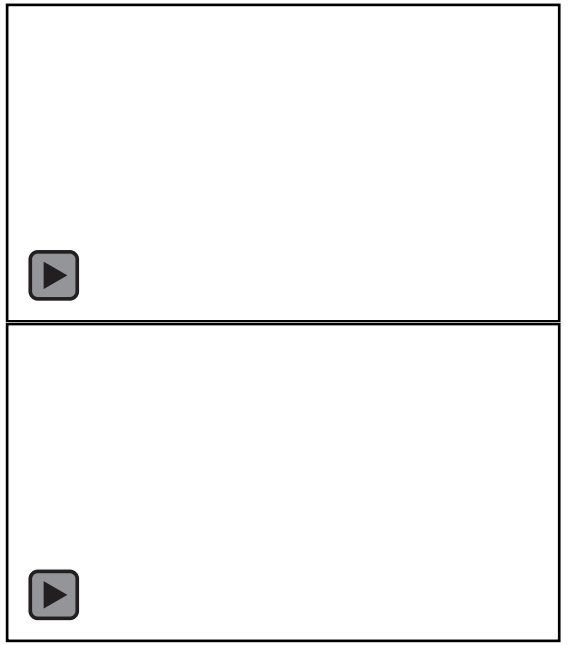
Feature

Linear Layer

Dog X

DINOv2: Learning Robust Visual Features without Supervision

Figure 8: **Examples of out-of-distribution examples** with frozen DINOv2-g features and a linear probe.



Visualization of the three first principal components

Dino Training

Dino Training

Dino Training

Dino Training

Student

Teacher

SOFTMAX

p_t

SOFTMAX

p_s

Dino Training

Student

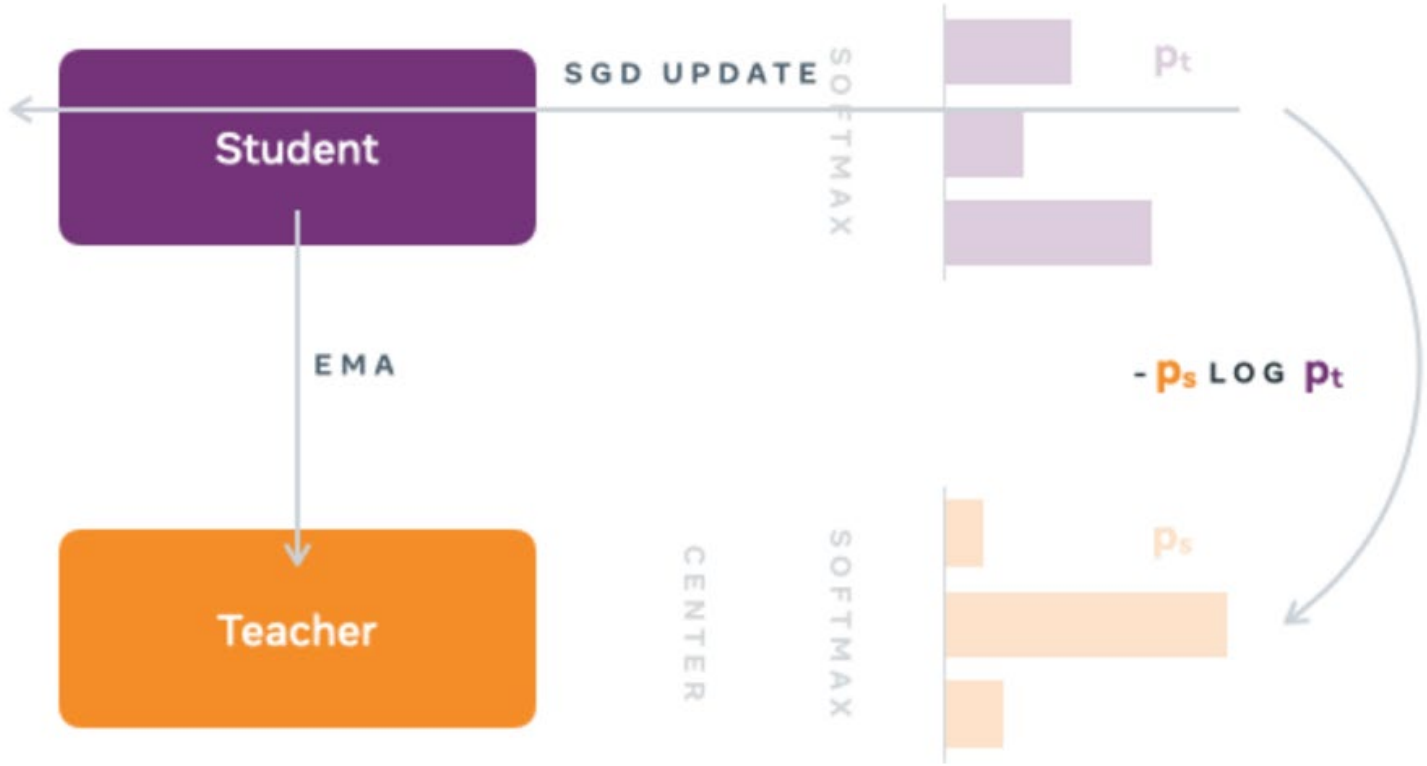
Teacher

SOFTMAX

CENTER

SOFTMAX

Dino Training



REPA (REPresentation Alignment)

Encoder DINOv2

Feature

Maximize Similarity

Feature

Denoising

REPA (REPresentation Alignment)

Encoder
DINOv2

→ Feature

Denoising

REPA (REPresentation Alignment)

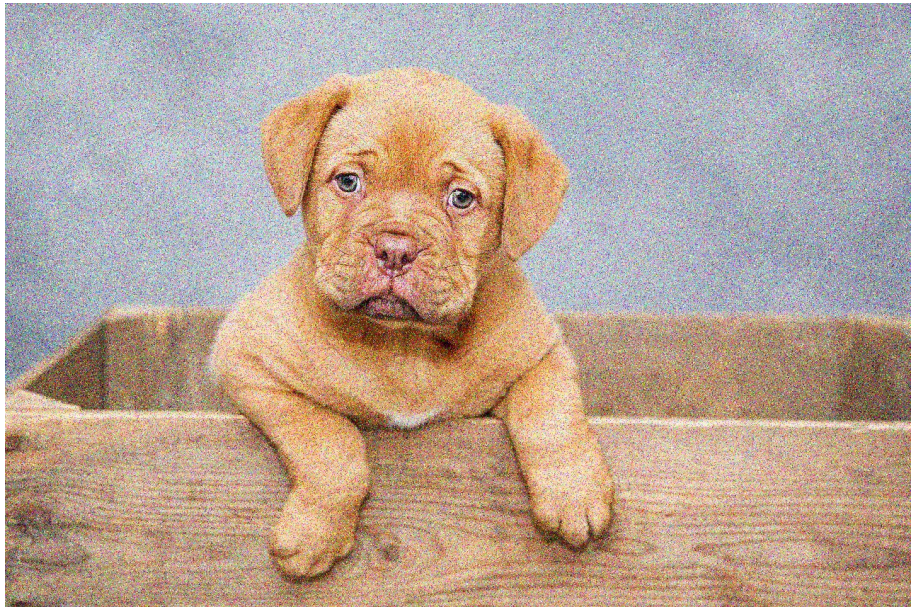
Encoder
DINOv2

→ Feature

REPA (REPresentation Alignment)

Encoder
DINOv2

→ Feature

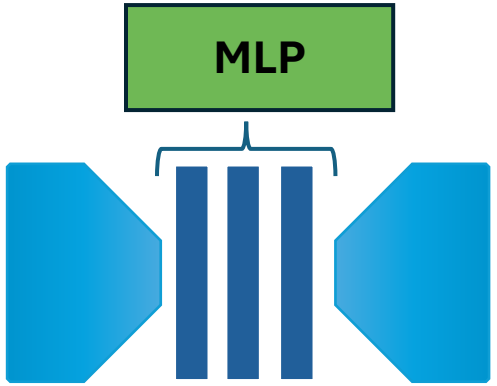


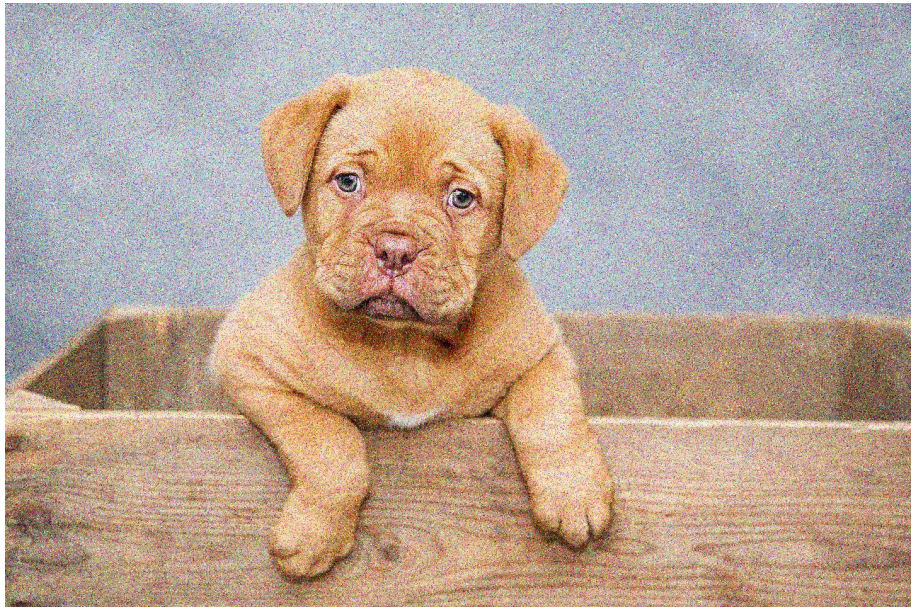
REPA (REPresentation Alignment)

Encoder
DINOv2

→ Feature

MLP





REPA (REPresentation Alignment)

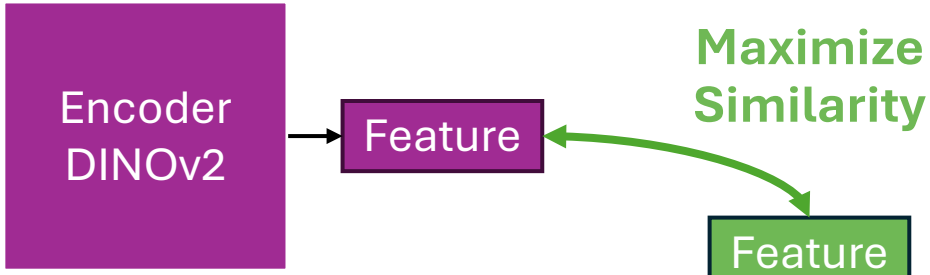
Encoder DINOv2

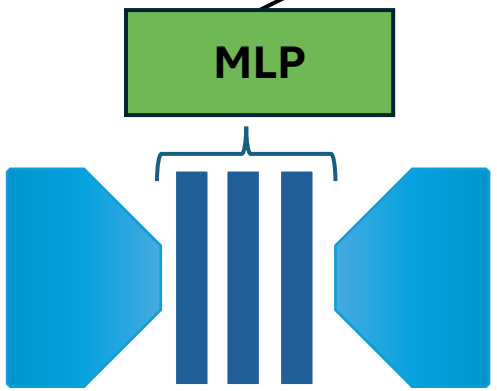
Feature

Feature

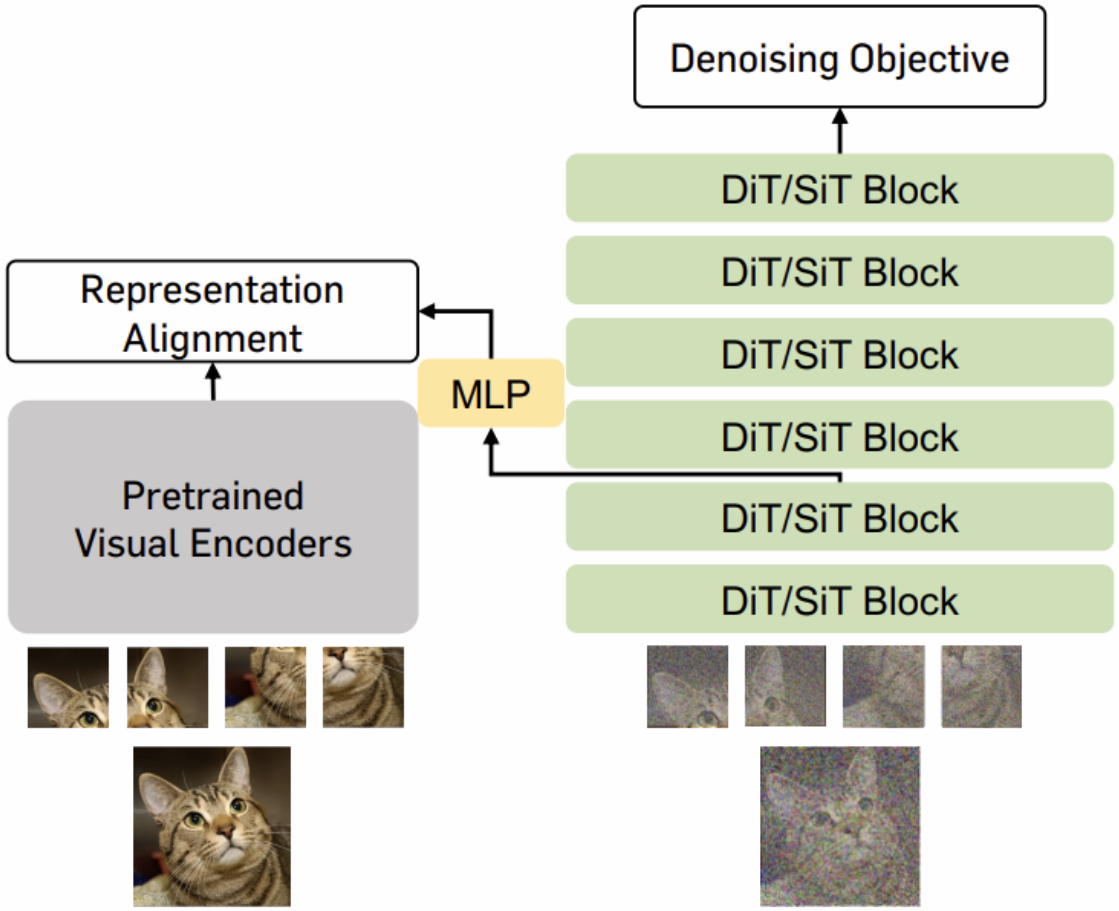
MLP

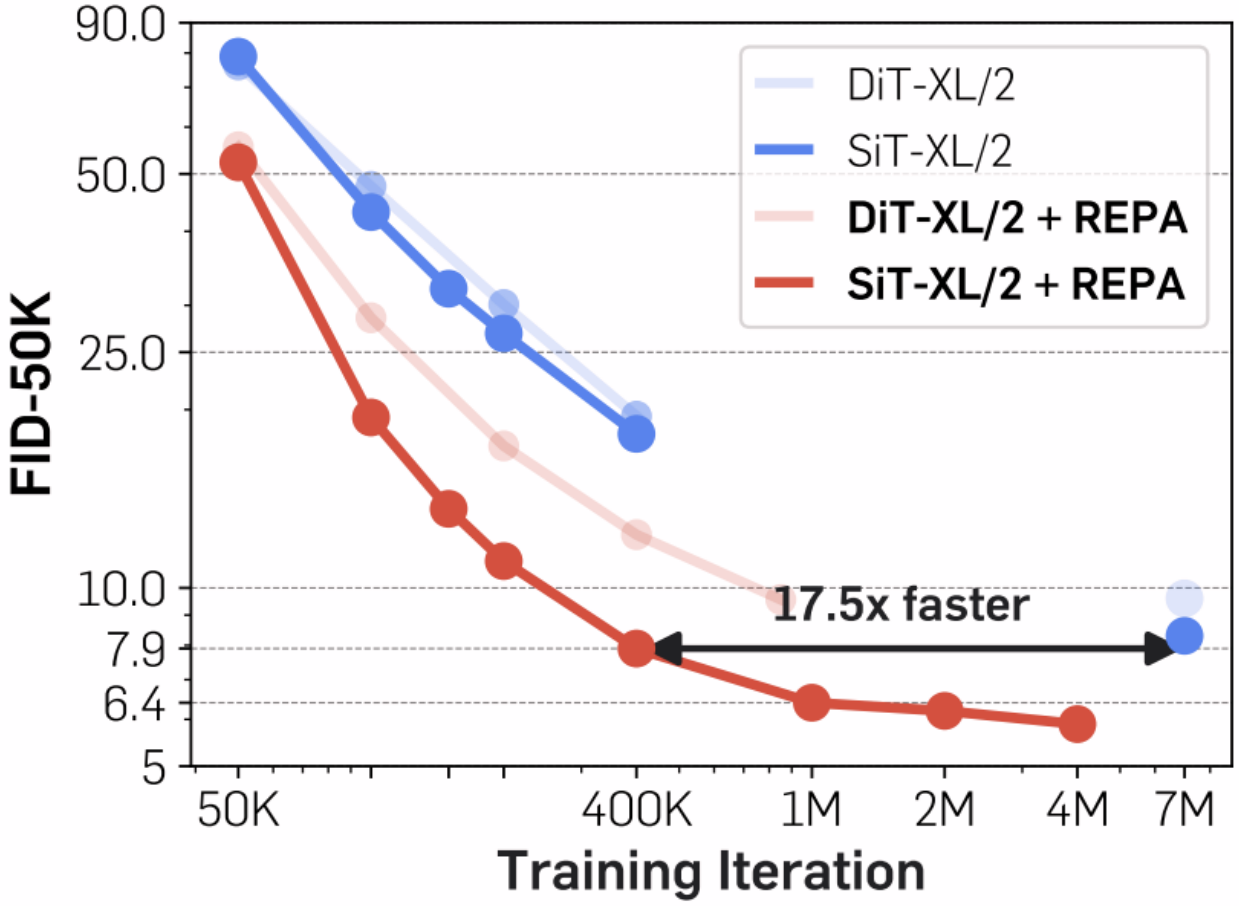
REPA (REPresentation Alignment)



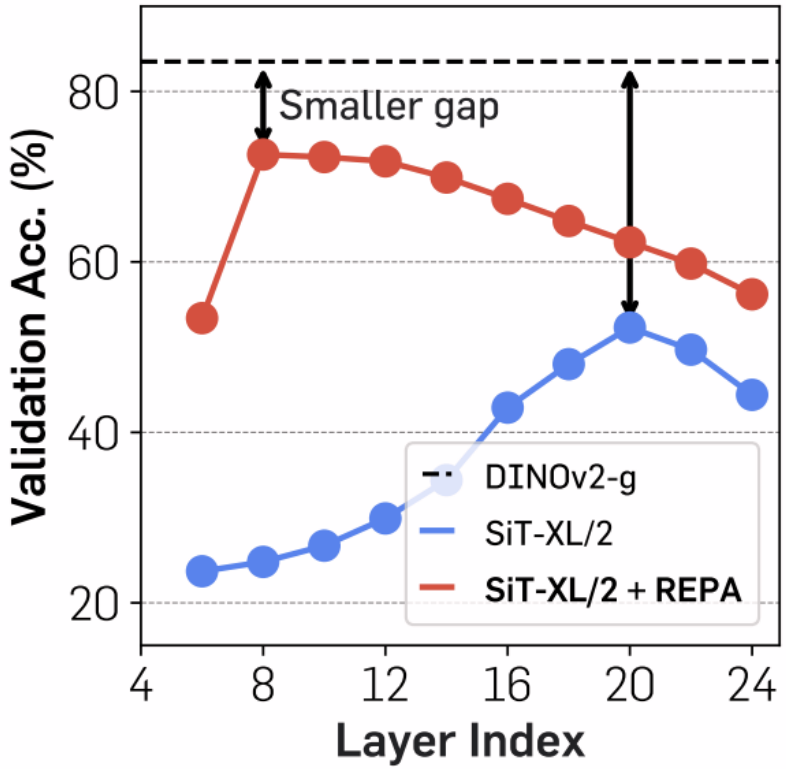


REPA (REPresentation Alignment)

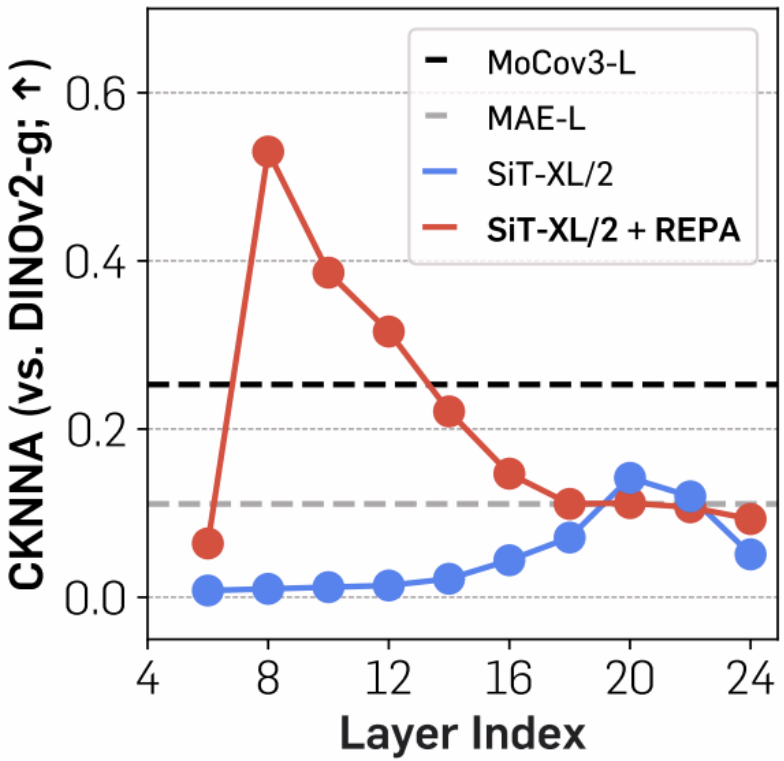




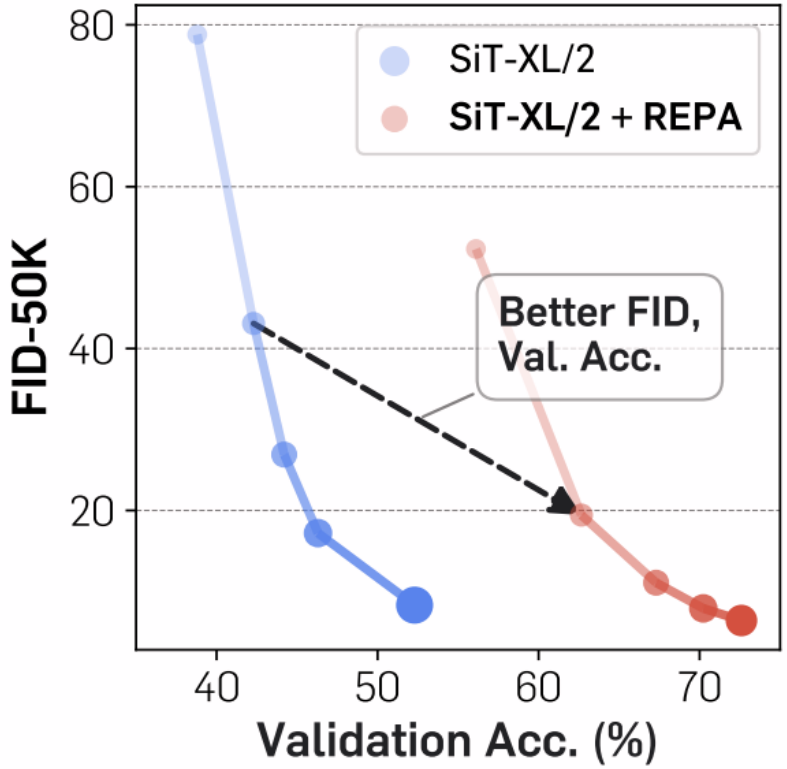
REPA (REPresentation Alignment)



(a) Semantic gap: Linear probing

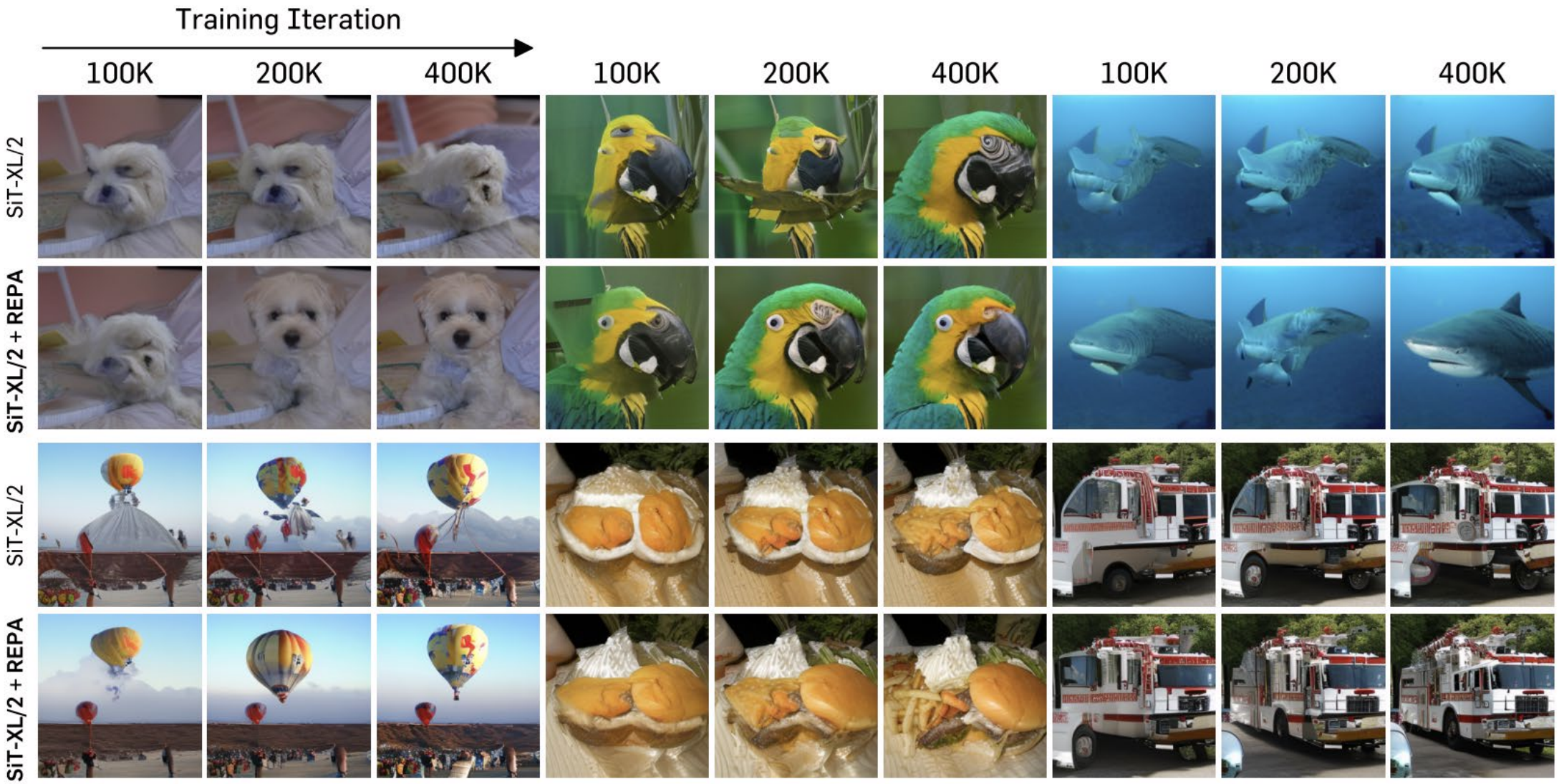


(b) Alignment to DINOv2-g



(c) Acc. and FID progression

REPA improves Visual Scaling



FID comparisons with vanilla DiTs and SiTs

Model	#Params	Iter.	FID↓
DiT-L/2	458M	400K	23.3
+ REPA (ours)	458M	**400K**	**15.6**
DiT-XL/2	675M	400K	19.5
+ REPA (ours)	675M	**400K**	**12.3**
DiT-XL/2	675M	7M	9.6
+ REPA (ours)	675M	**850K**	**9.6**
SiT-B/2	130M	400K	33.0
+ REPA (ours)	130M	**400K**	**24.4**
SiT-L/2	458M	400K	18.8
+ REPA (ours)	458M	**400K**	**9.7**
+ REPA (ours)	458M	**700K**	**8.4**
SiT-XL/2	675M	400K	17.2
+ REPA (ours)	675M	**150K**	**13.6**
SiT-XL/2	675M	7M	8.3
+ REPA (ours)	675M	**400K**	**7.9**
+ REPA (ours)	675M	**1M**	**6.4**
+ REPA (ours)	675M	**4M**	**5.9**

How to Extend REPA?

How to Extend REPA?

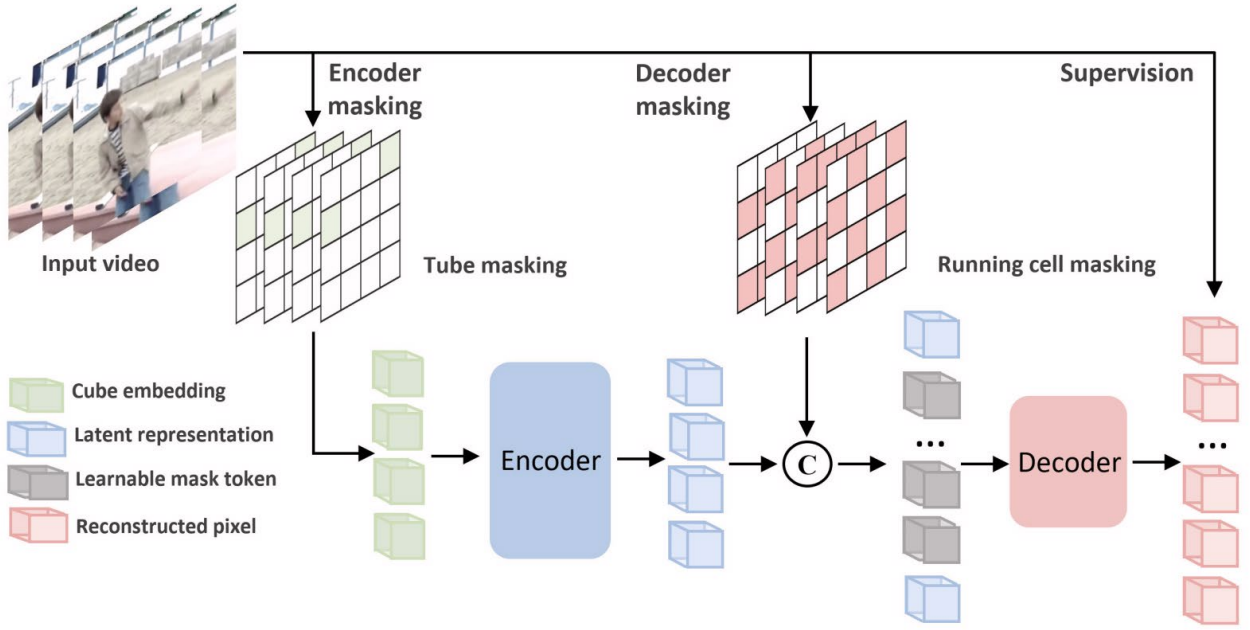
- Base REPA
  - DinoV2 + DiT
- Semantic Representation
  - DINOv2 - CLIP – ImageBind
- Geometric / 3D Representation
  - VGGT - Trellis – PointNeXt - Hunyuan3D
- Temporal / Video Representation
  - VideoMAE-v2 – Timesformer - I3D

How to Extend REPA?

- Base REPA
  - DinoV2 + DiT
- Semantic Representation
  - DINOv2 - CLIP – ImageBind
- Geometric / 3D Representation
  - VGGT - Trellis – PointNeXt - Hunyuan3D
- Temporal / Video Representation
  - **VideoMAE-v2** – Timesformer - I3D

VideoMAE-v2

- VideoMAE-v2 is a self-supervised video autoencoder.
- Trained to reconstruct videos with up to 90% spatiotemporal masking.
- Learns features that capture motion, temporal dynamics, and scene changes.



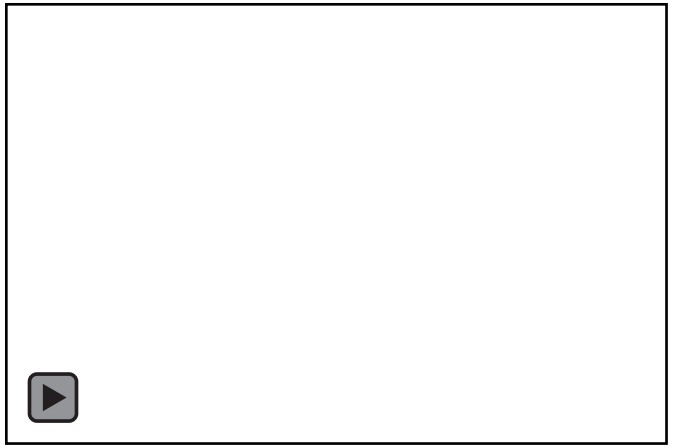
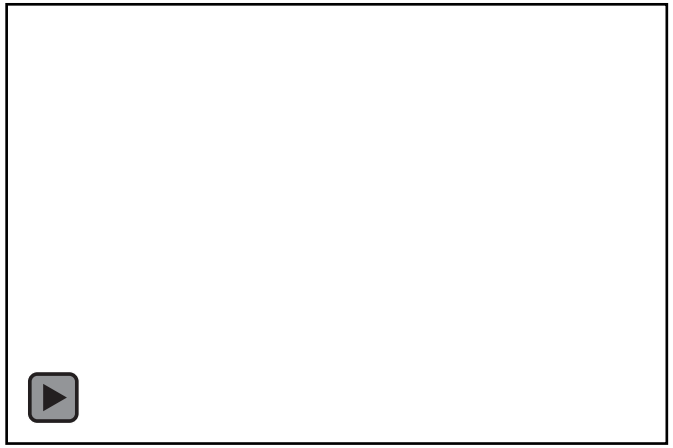
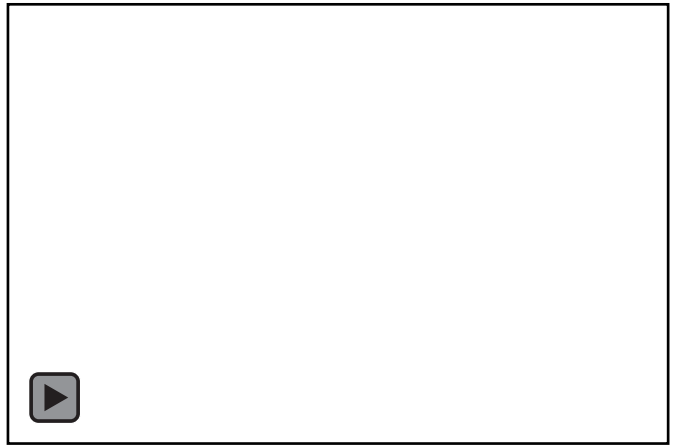
VideoREPA

- VideoREPA aligns CogVideoX’s internal features with VideoMAE-v2.
- This alignment injects temporal and motion understanding.

VideoREPA

- VideoREPA aligns CogVideoX's internal features with VideoMAE-v2.
- This alignment injects temporal and motion understanding.

Glass shatters on the floor.

CogVideoX	CogVideoX+REPA	VideoREPA
		

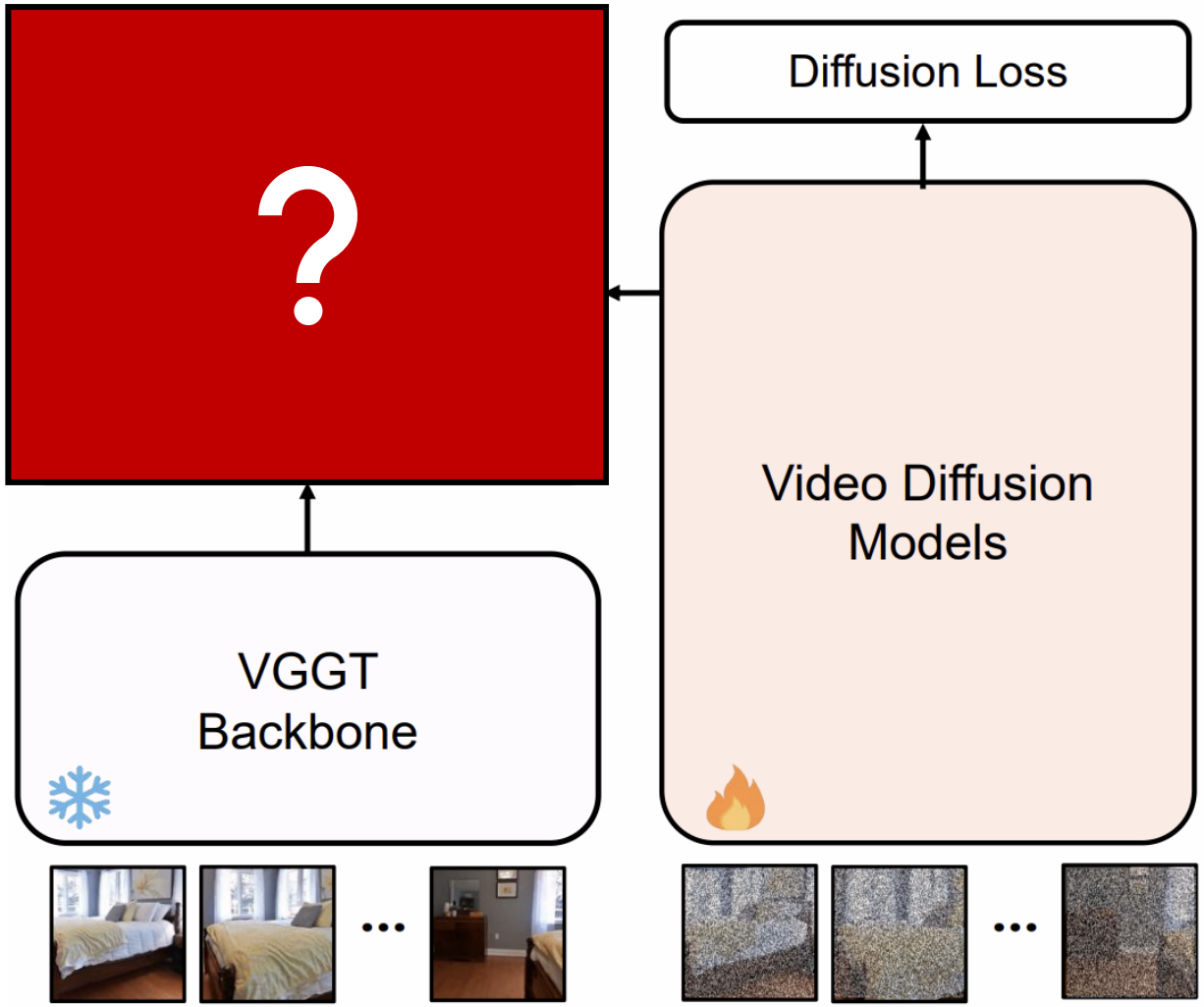
VideoREPA

- VideoREPA aligns CogVideoX's internal features with VideoMAE-v2.
- This alignment injects temporal and motion understanding.

Back to the Original Paper

Geometry Forcing:

- Injecting 3D geometric awareness into video models.

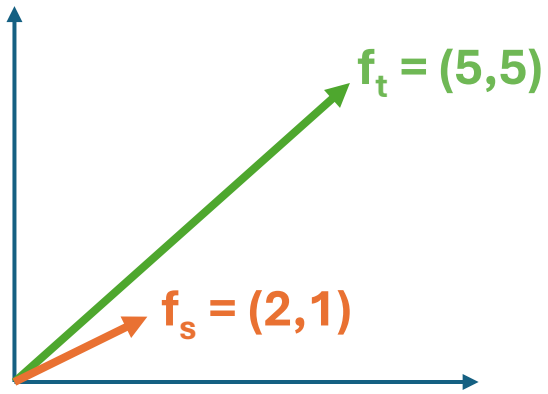


Angular Alignment

- f_t : Teacher feature vector (from VGGT)
- f_s : Student feature vector (from the diffusion model)
- Goal: Align the direction (angle) of f_s with f_t.
- Loss:
  - $L_{Angular} = \cos(\theta) = \frac{f_t . f_s}{\|f_t\| \, \|fs\|}$
- Intuition:
  - If both vectors point in the same direction, they represent the same concept.

Why Angular Alignment Is Not Enough

- Magnitude differences that still carry good information.
- Example: For the vectors f_t=(5,5) and f_s=(2,1):
  - Cosine similarity ≈ 0.95, they point in a similar direction
  - L2 distance = 5.0, their magnitudes differ significantly

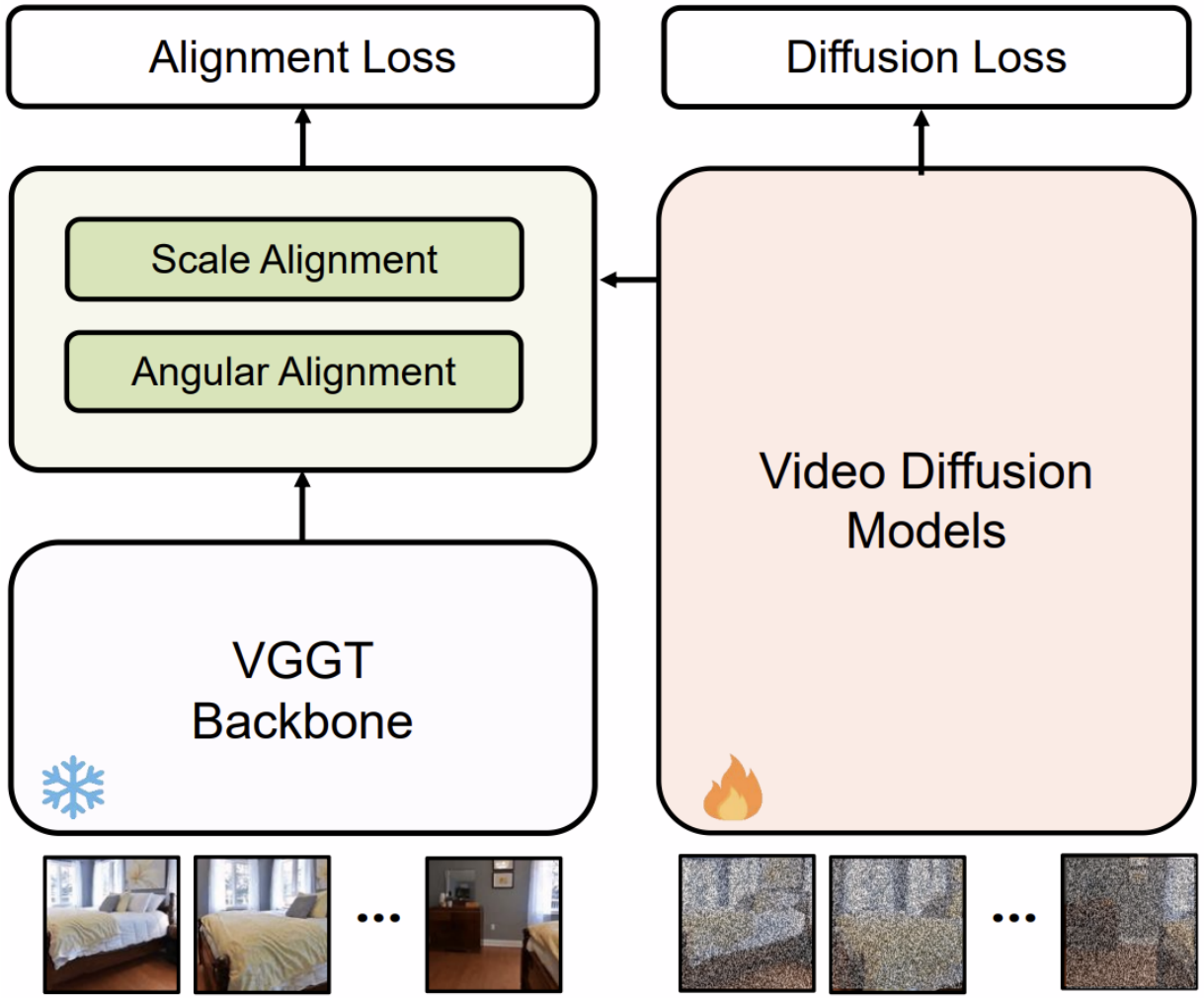


Scale Alignment

- Norms vary across diffusion blocks → L_2 loss can explode or vanish.
- Solution: Per-Layer Affine Rescaling:
  - $f'_s = \frac{f_s}{||fs||_2}$ and $f''_s = g_\phi(f'_s)$
- Loss:
  - $L_{Scale} = ||f''_s - f_t||_2$

Geometry Forcing:

- Injecting 3D geometric awareness into video models.



Which Representation Should be Aligned?

Target Representation	FVD-256
Baseline	364
DINOv2 Only	297
VGGT Only	243
VGGT + DINOv2	**237**

Ablation study on Alignment Loss

Alignment Loss	FVD-256
Baseline	364.0
Angular	253.0
Angular + Scale	**243.0**
MSE	1648.0

Visualization: 360 Rotation

Initial View

Revisit View

Visualization: 360 Rotation

Initial View

Revisit View

Comparison with baseline

Ground Truth

Baseline

Geometry Forcing

Comparison with baseline

Ground Truth

Baseline

Geometry Forcing

Comparison with baseline

Ground Truth

Baseline

Geometry Forcing

Comparison with baseline

Ground Truth

Baseline

Geometry Forcing

Takeaway:

Why Representation Alignment Matters?

- Geometry-based supervision (VGGT) is crucial for realistic 3D video generation.
- Combining geometric and semantic features (VGGT + DINOv2) yields even better results.
- This approach adds richer understanding without extra inference cost.

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