Video Generation via Latent Space Navigation

Yaohui Wang (王耀晖)

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

State-of-the-art: Deep generative models for *Image* generation

Face image generation [Karras et al., CVPR'20]

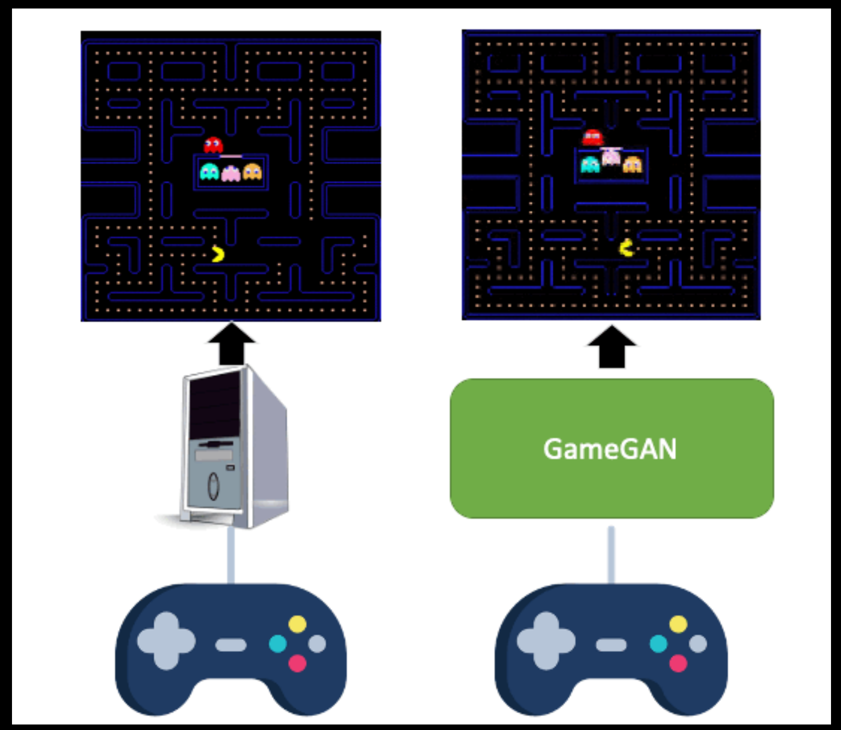
Object generation [Brock et al., arXiv'18]

Interactive image editing [Park et al., NeurIPS'20]

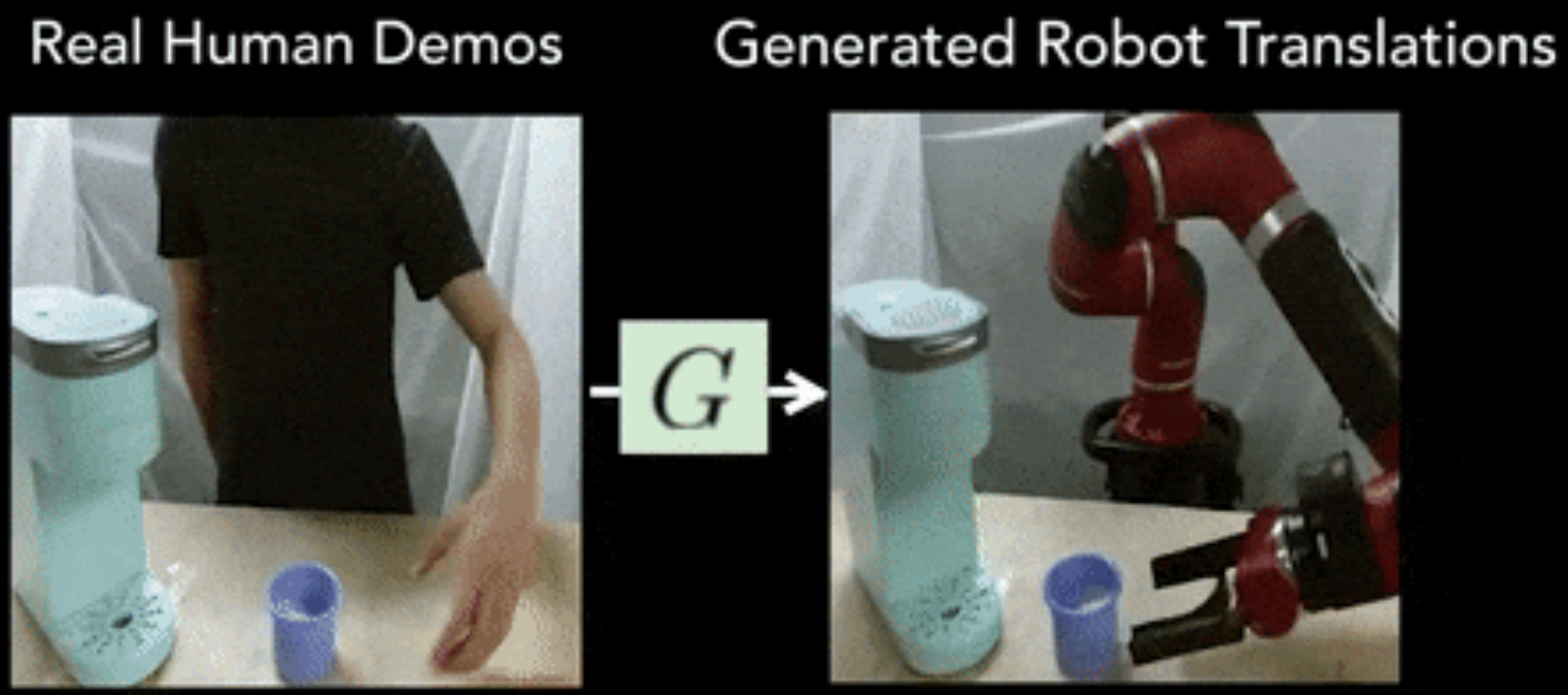
Text-to-image Generation [DALLE2]

Video generation

Autonomous driving [Wang et al., NeurIPS'18]



Video games [Kim et al., CVPR'20]



Robot imitation learning [Smith et al., RSS'20]

3D-aware videos [Menapace et al., CVPR'22]

Challenges in video generation

1. How to design a generator for video generation?
2. How to represent a video in the latent space?
3. General method for video generation tasks?

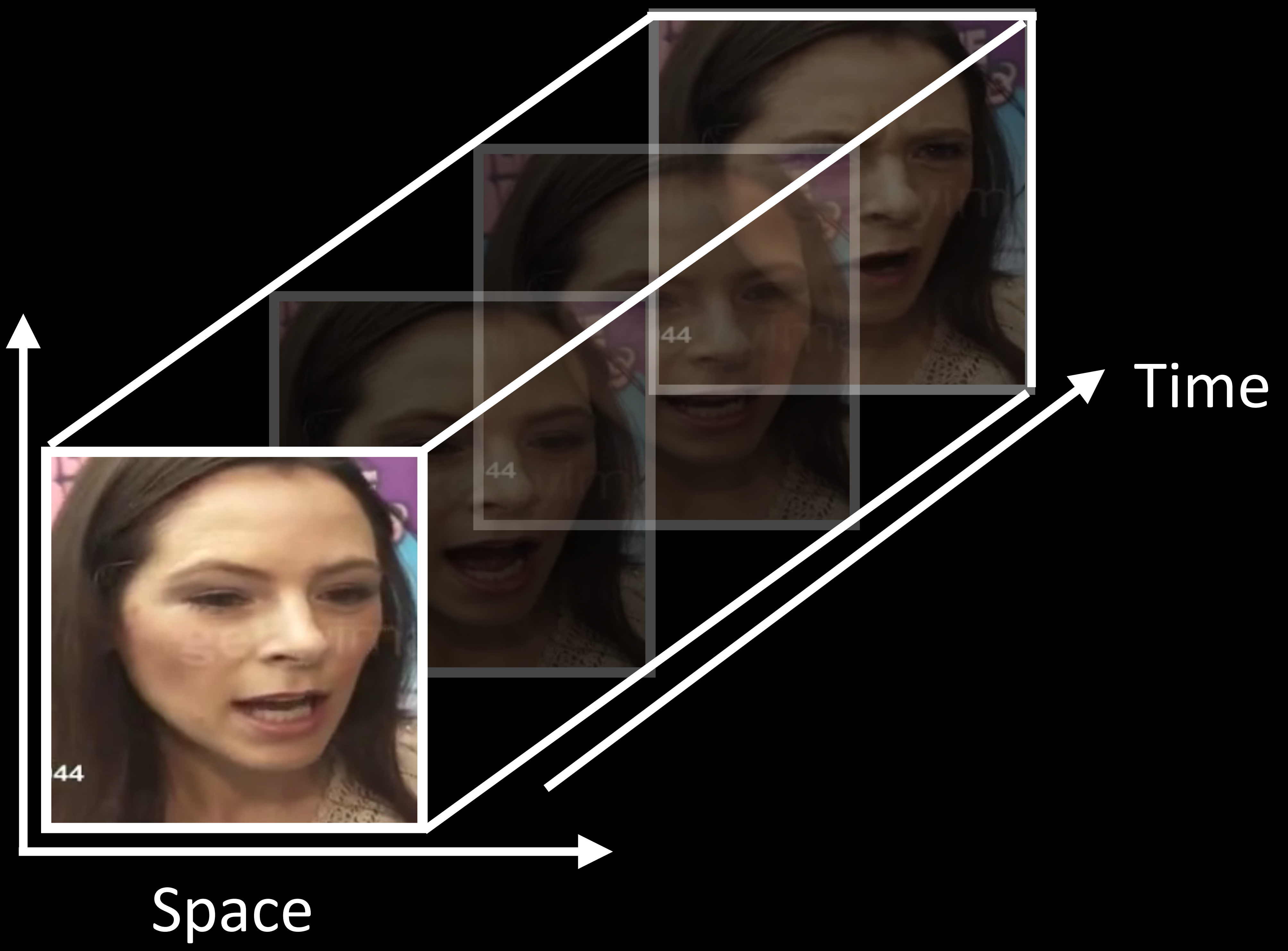
? → ? →

Latent Space

- Interpretable
- Controllable

Architecture of generator

- Sharp images
- Spatio-temporal consistency



Outline

1. Noise-to-video generation

- G^3AN [Wang et al., CVPR'20]
- InMoDeGAN [Wang et al., arXiv'21]

2. Image-to-video generation (Image Animation)

- ImaGINator [Wang et al., WACV'20]
- LIA [Wang et al., ICLR'22]

Outline

1. Noise-to-video generation

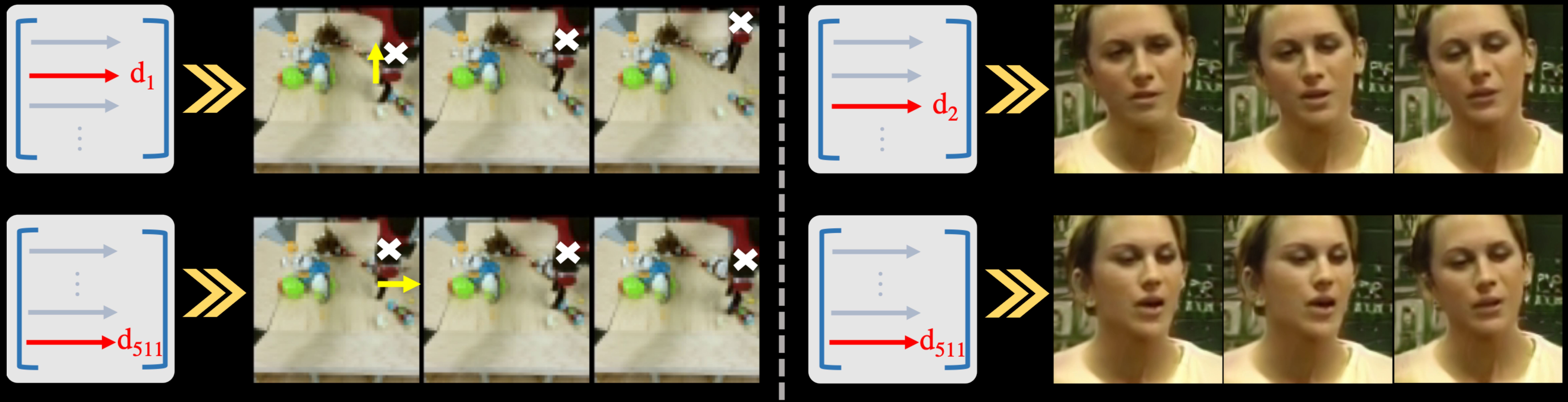
- G^{3}AN [Wang et al., CVPR'20]
- InMoDeGAN [Wang et al., arXiv'21]

2. Image-to-video generation (Image Animation)

- ImaGINator [Wang et al., WACV'20]
- LIA [Wang et al., ICLR'22]

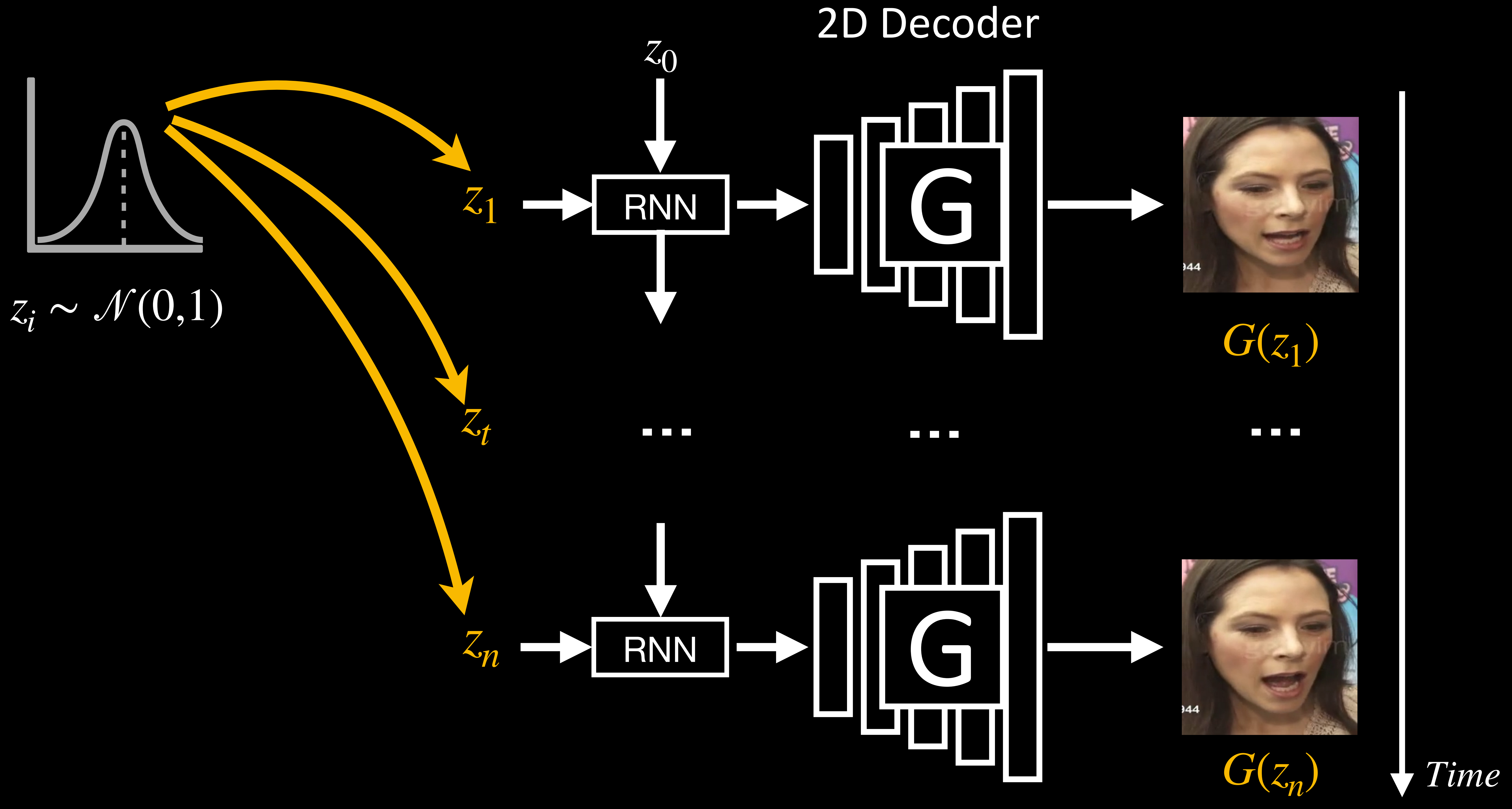
InMoDeGAN: Interpretable Motion Decomposition Generative Adversarial Network for Video Generation

[Wang et al., arXiv’21]



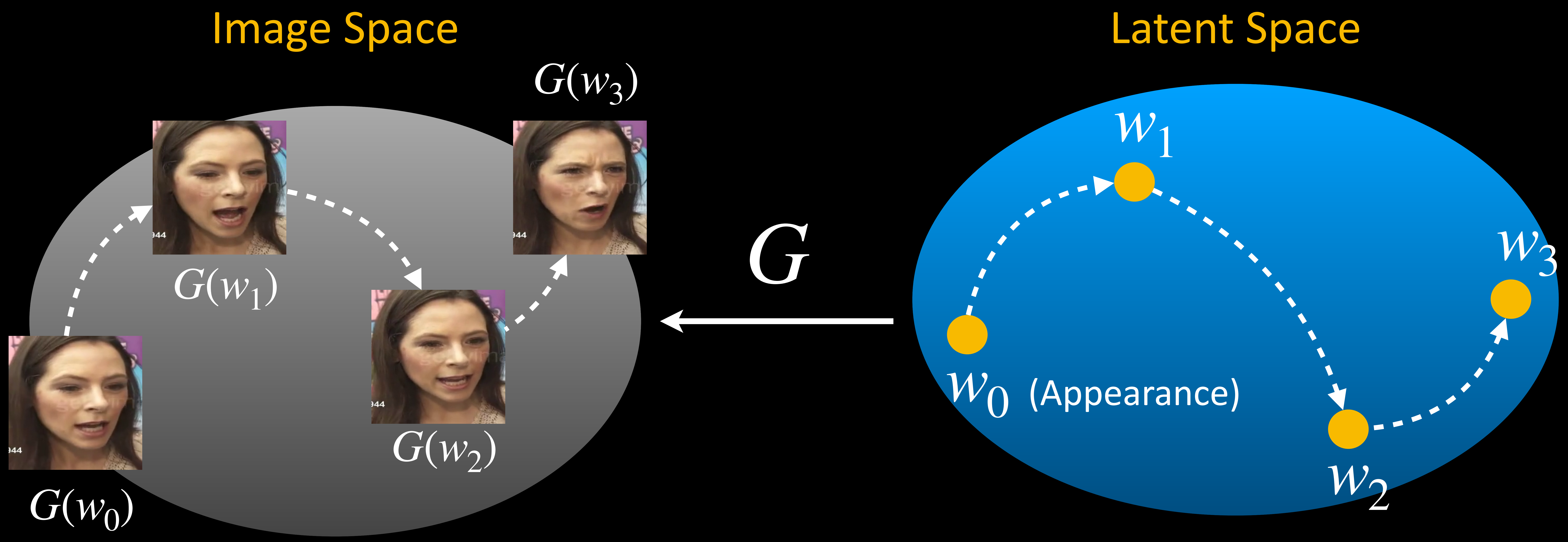
Goal:
1. High resolution generation
2. Interpretable motion space

InMoDeGAN: General model architecture



Sequence-to-sequence model

InMoDeGAN: From image space to latent space

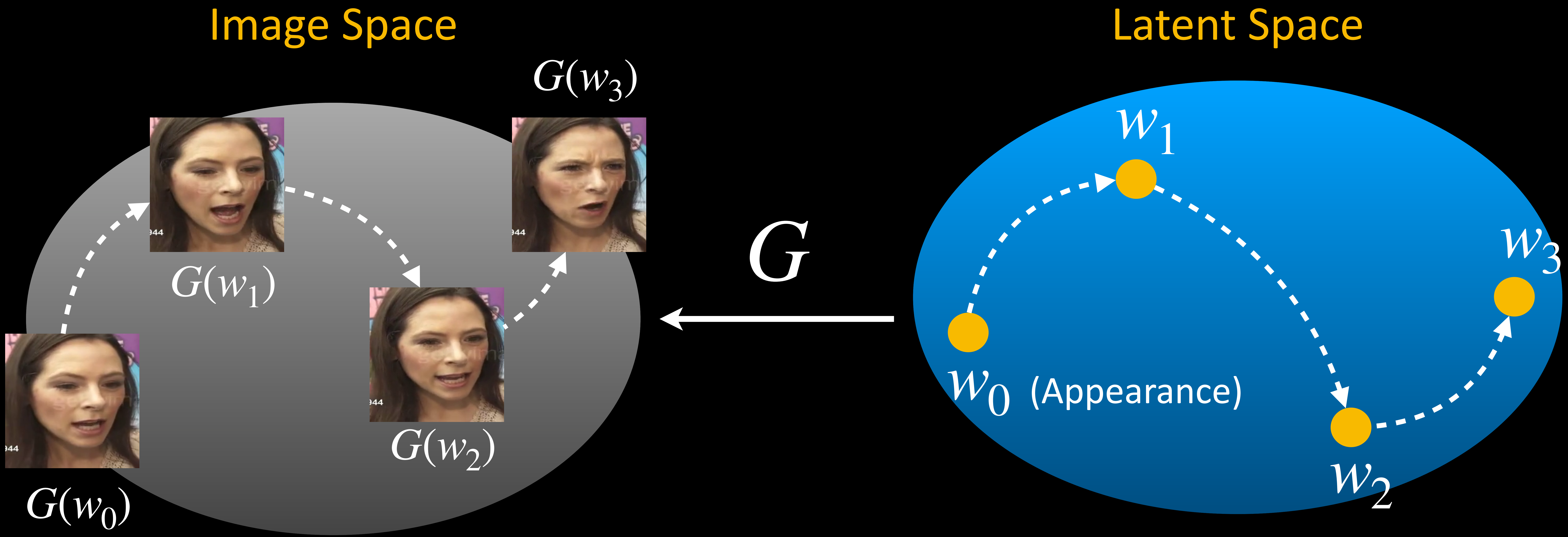


Latent transformation

$G(w_{t+1}) = \mathscr{T}_{t \to t+1}(G(w_t))$ $\xrightarrow{\text{Idea in equivariance}}$ $w_{t+1} = \boxed{\tau_{t \to t+1}}(w_t)$

Transformations in the <u>latent space</u> result in equivalent transformations in the <u>image space</u>

InMoDeGAN: From image space to latent space

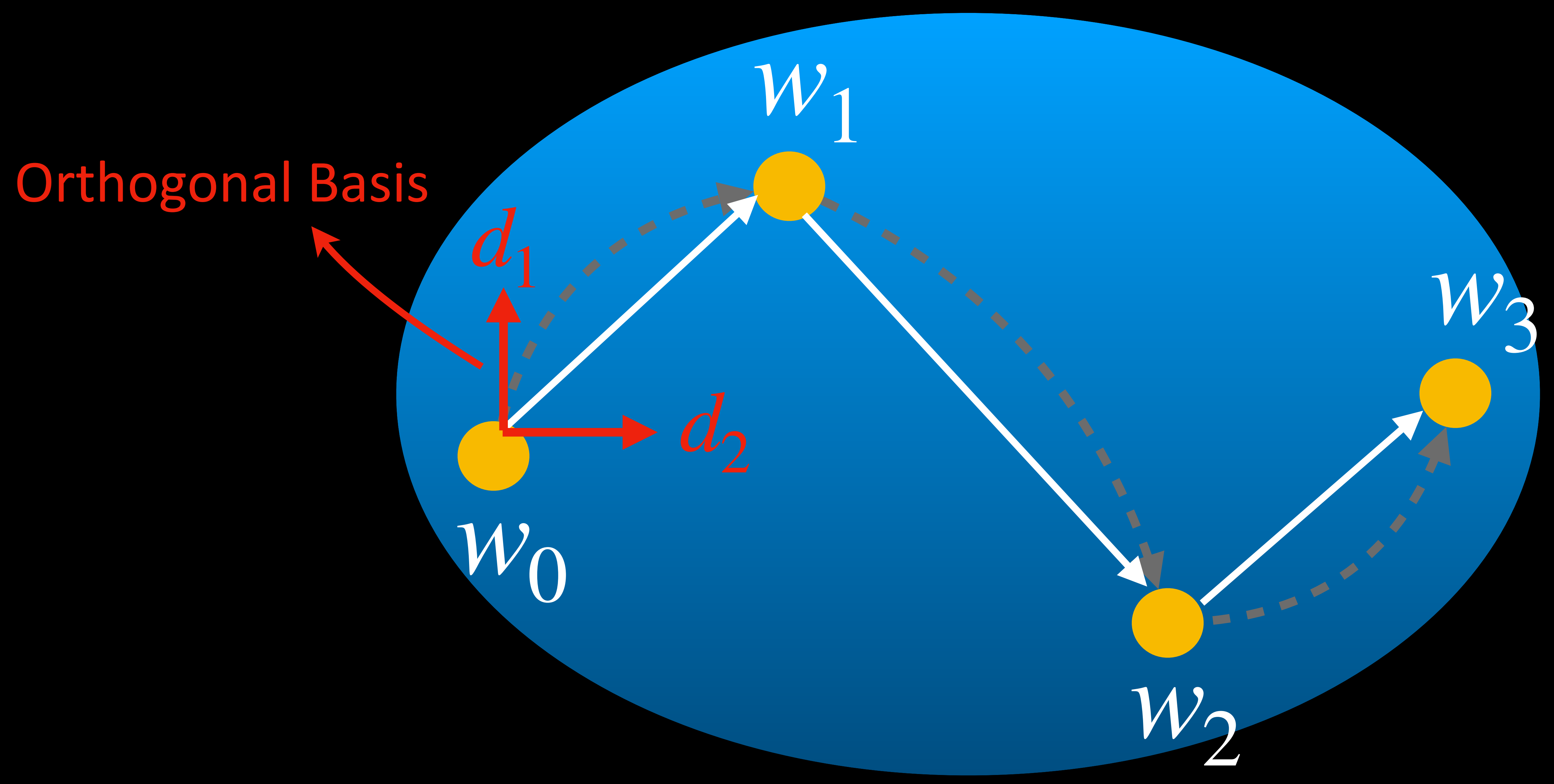


$$G(w_{t+1}) = \mathcal{T}_{t\to t+1}(G(w_t)) = G(\tau_{t\to t+1}(w_t))$$

$$w_{t+1} = \overset{?}{\boxed{\tau_{t\to t+1}}}(w_t)$$

InMoDeGAN: Linear Motion Decomposition (LMD)

Latent Space



$$w_{t+1} = \tau_{t \rightarrow t+1}(w_t)$$

$$\Downarrow$$

$$w_{t+1} = w_t + p_{t \rightarrow t+1}$$

Recurrence relation

$$\left.\begin{aligned} w_1 &= w_0 + \sum_{i=0}^{N-1} \alpha_{1,i}\, d_i \\ &\vdots \\ w_t &= w_{t-1} + \sum_{i=0}^{N-1} \alpha_{t,i}\, d_i \end{aligned}\right\} \sum$$

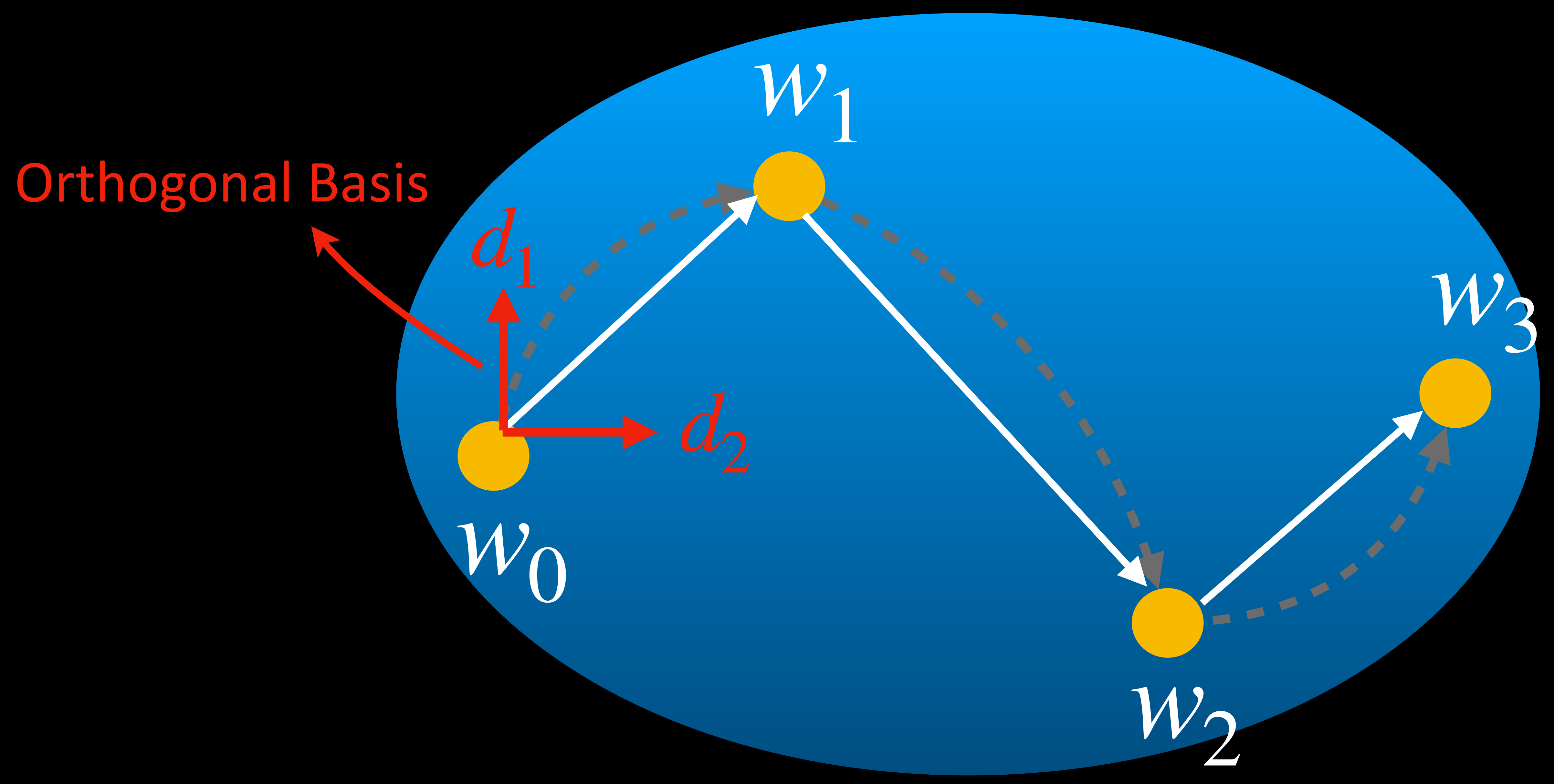
$$\Downarrow$$

$$w_t = w_0 + \sum_{t=1}^{t} \sum_{i=0}^{N-1} \alpha_{t,i}\, d_i$$

General formula of w_t

InMoDeGAN: Linear Motion Decomposition (LMD)

Latent Space



$$w_{t+1} = \tau_{t \to t+1}(w_t)$$

$$\Downarrow$$

$$w_{t+1} = w_t + p_{t \to t+1}$$

Recurrence relation

$$\left.\begin{aligned} w_1 &= w_0 + \sum_{i=0}^{N-1} \alpha_{1,i}\, d_i \\ &\vdots \\ w_t &= w_{t-1} + \sum_{i=0}^{N-1} \alpha_{t,i}\, d_i \end{aligned}\right\} \sum$$

$$\Downarrow$$

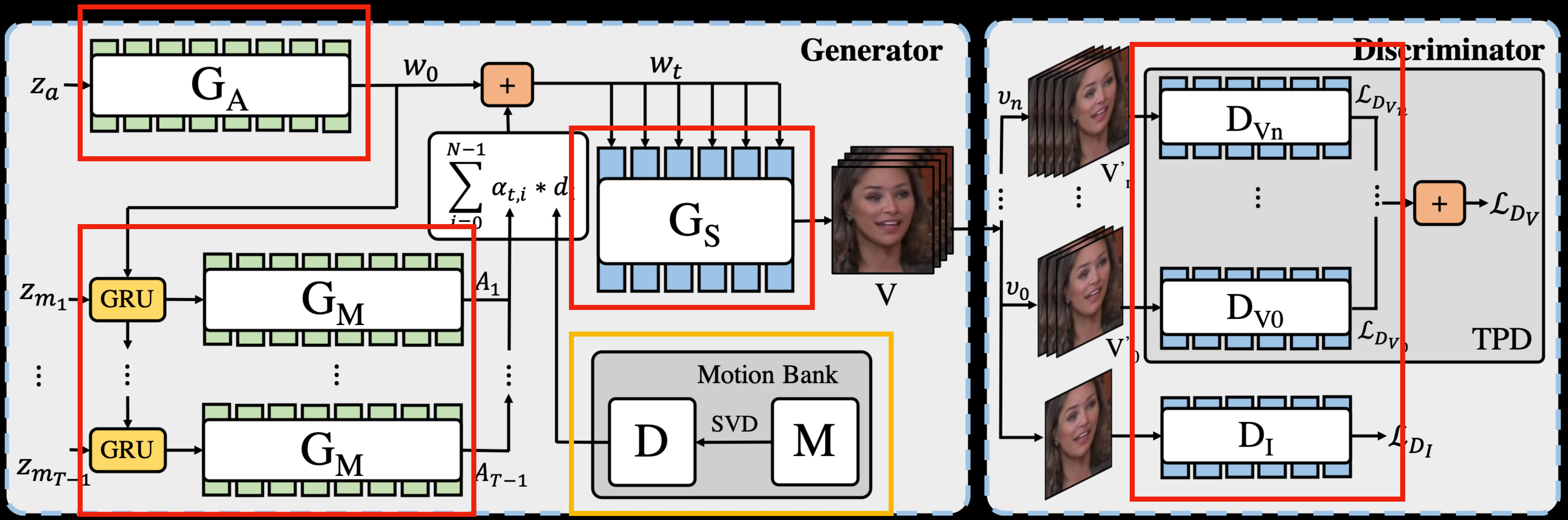
$$w_t = w_0 + \sum_{t=1}^{t} \sum_{i=0}^{N-1} \alpha_{t,i}\, d_i$$

Appearance

Motion magnitude

Motion direction

InMoDeGAN: Architecture

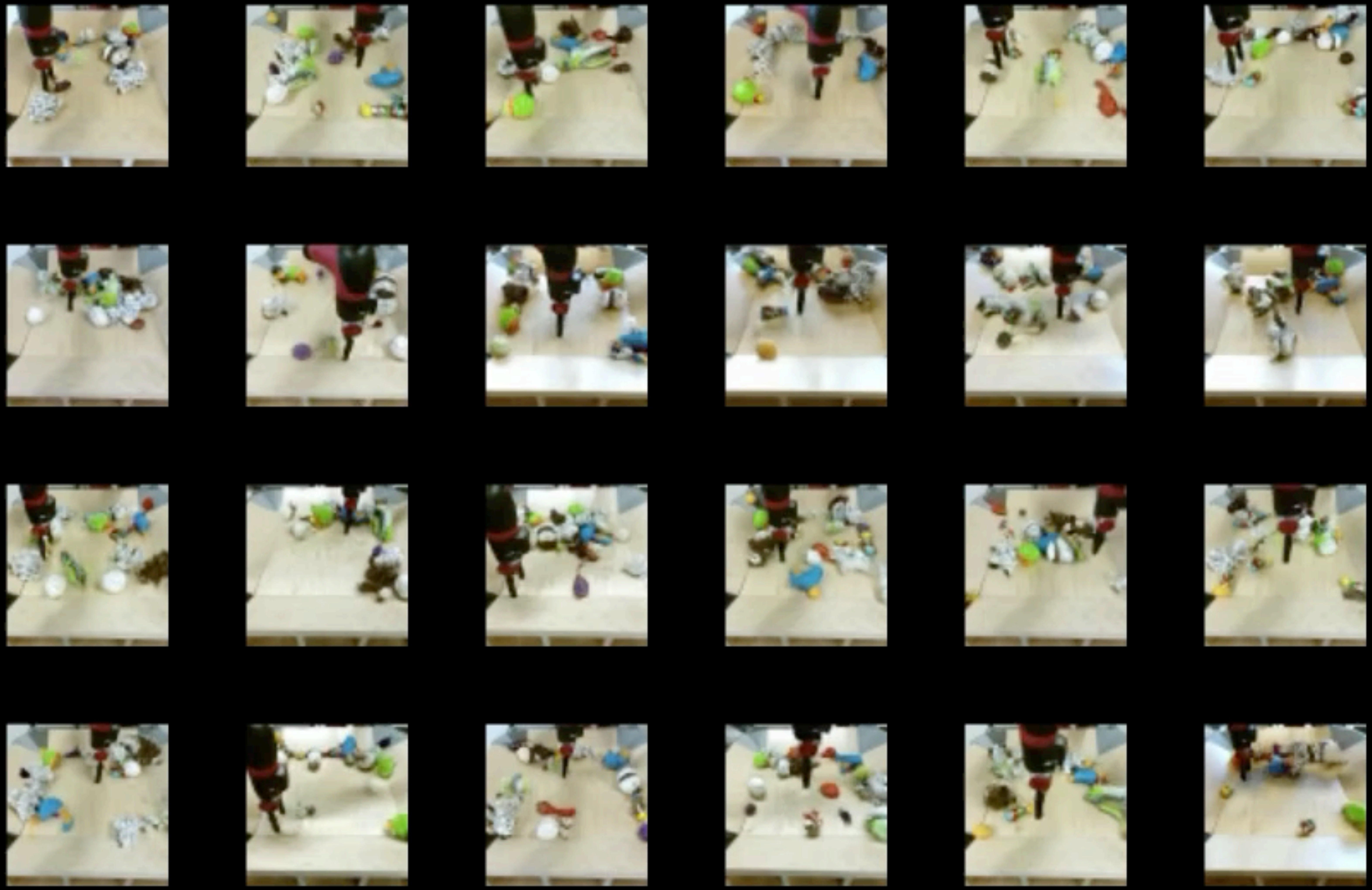


Orthogonal Basis (Motion Bank)

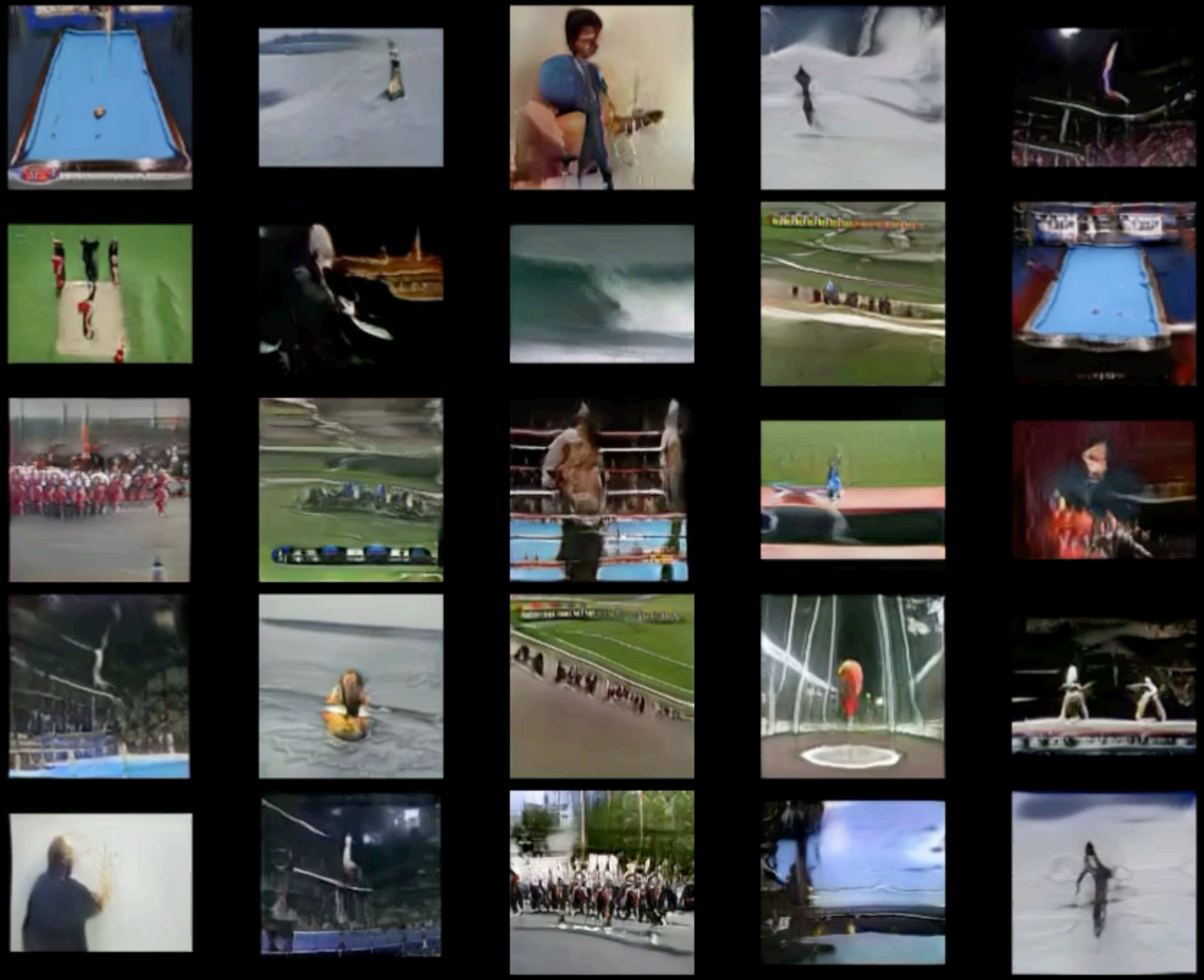
Learning model parameters and motion directions simultaneously

InMoDeGAN: Results (VoxCeleb)

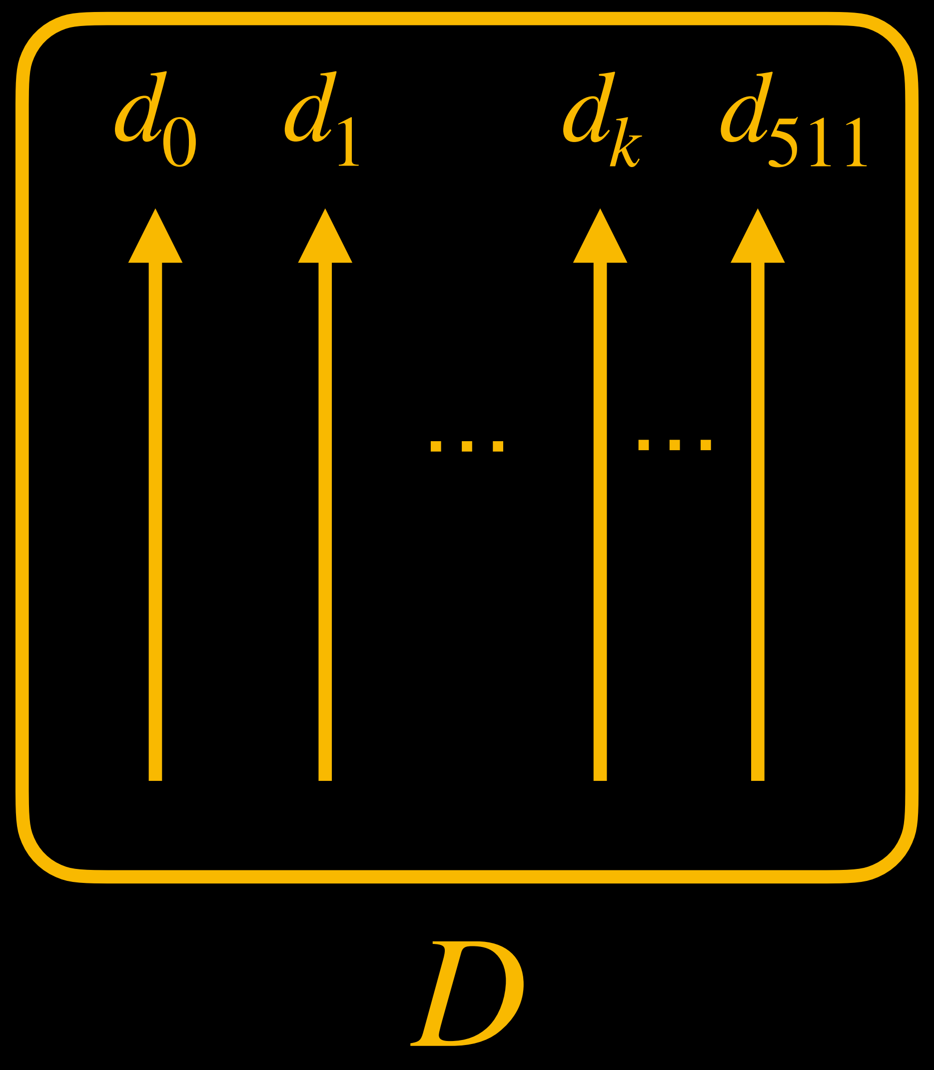
InMoDeGAN: Results (BAIR)



InMoDeGAN: Results (UCF101)

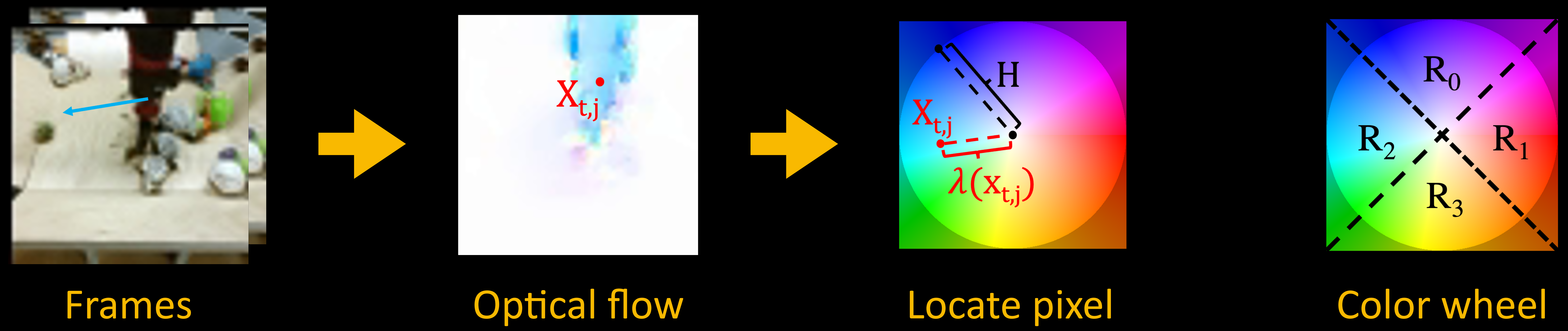


InMoDeGAN: Motion interpretation



What does d_i represent?

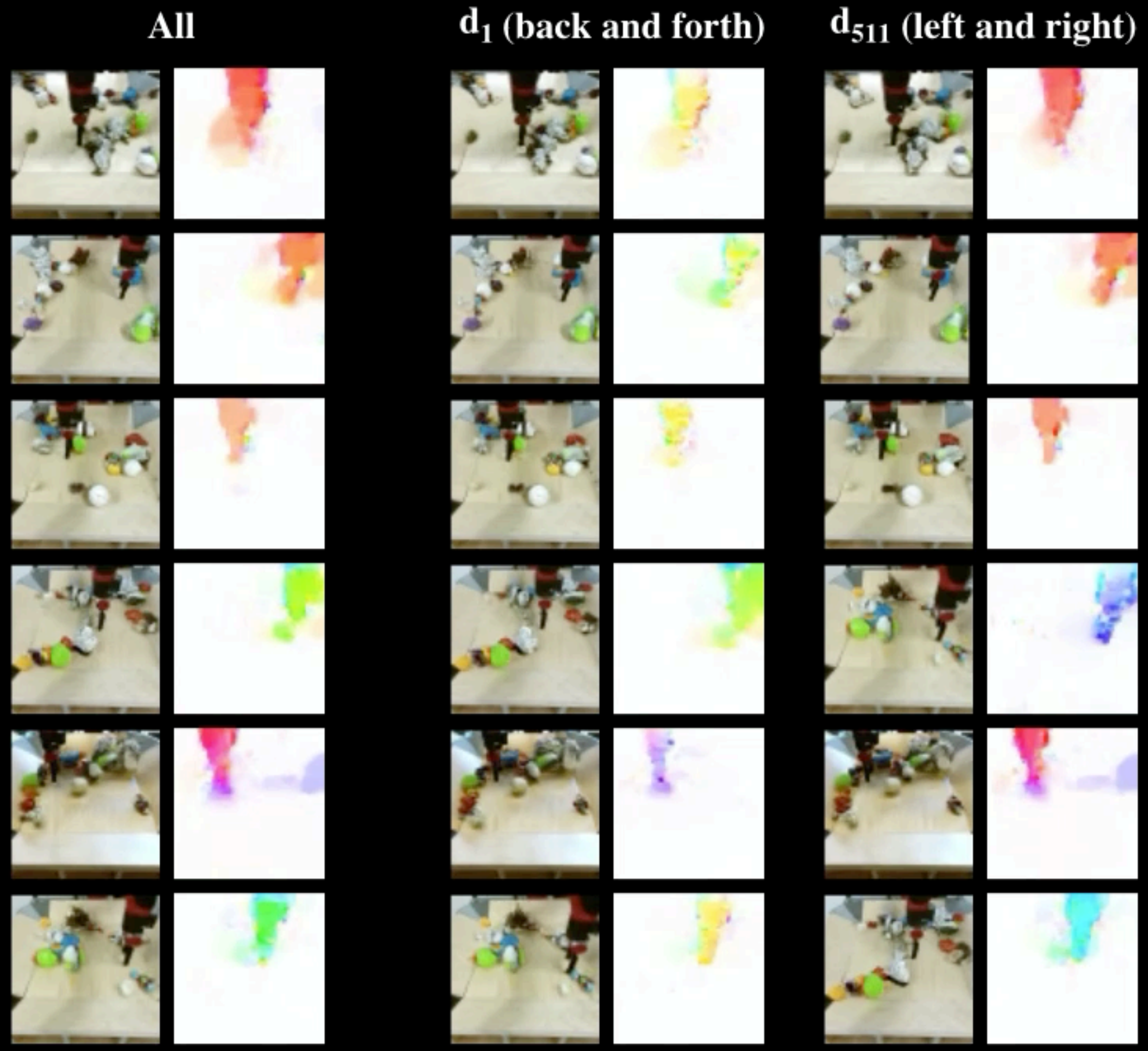
InMoDeGAN: Motion interpretation (BAIR) — leveraging optical flow



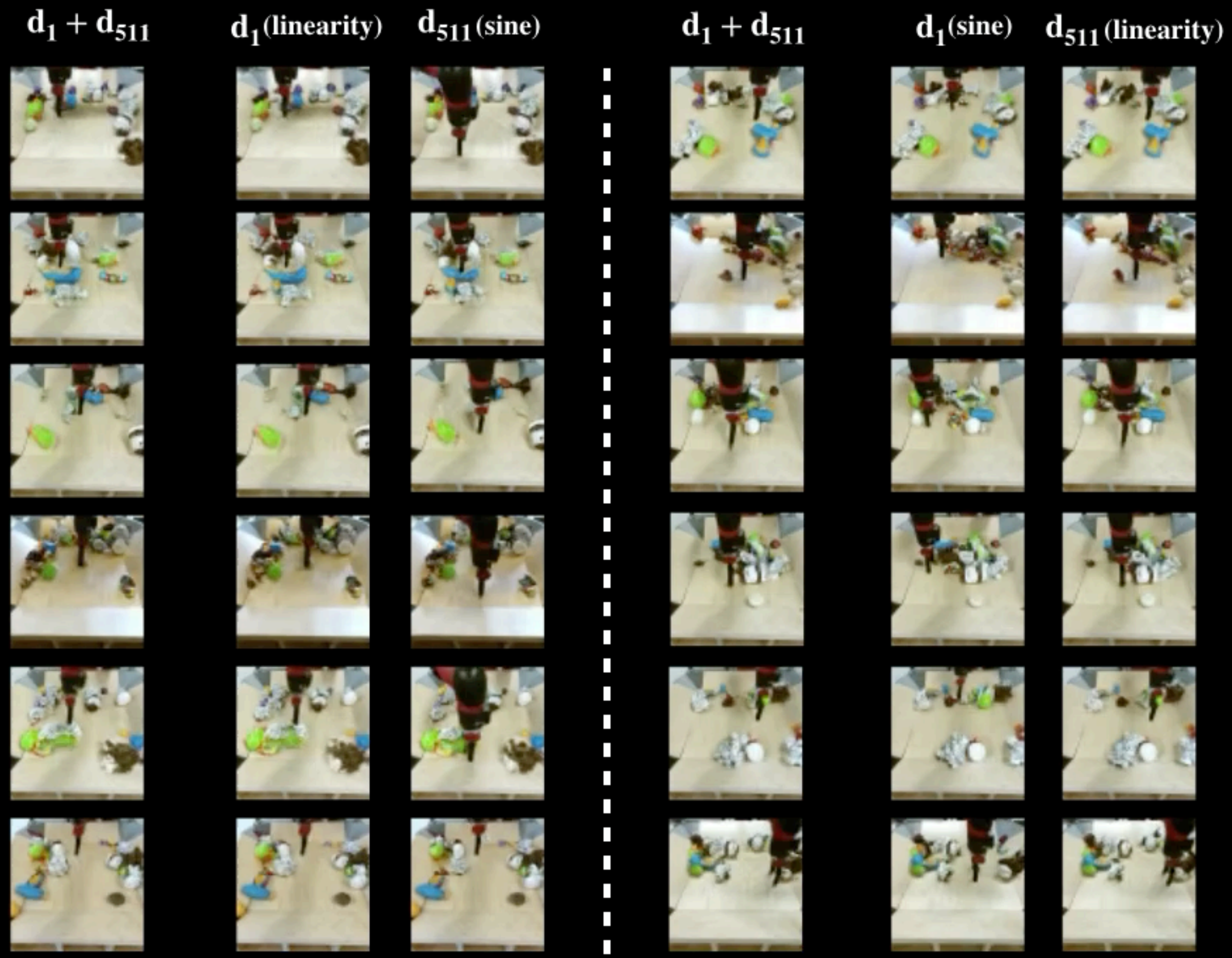
$$\phi_i = \frac{1}{N_i}\sum_{t=0}^{T-1}\sum_{j=0}^{N-1}\frac{\lambda(x_{t,j})}{H}1_{R_i}(x_{t,j}),\ i \in \{0,1,2,3\}$$

Quantify motion in R_0, R_1, R_2, R_3

InMoDeGAN: Results (BAIR) — Direction analysis



InMoDeGAN: Results (BAIR) — Controllable generation



Outline

1.Noise-to-video generation

- G3AN [Wang et al., CVPR'20]
- InMoDeGAN [Wang et al., arXiv'21]

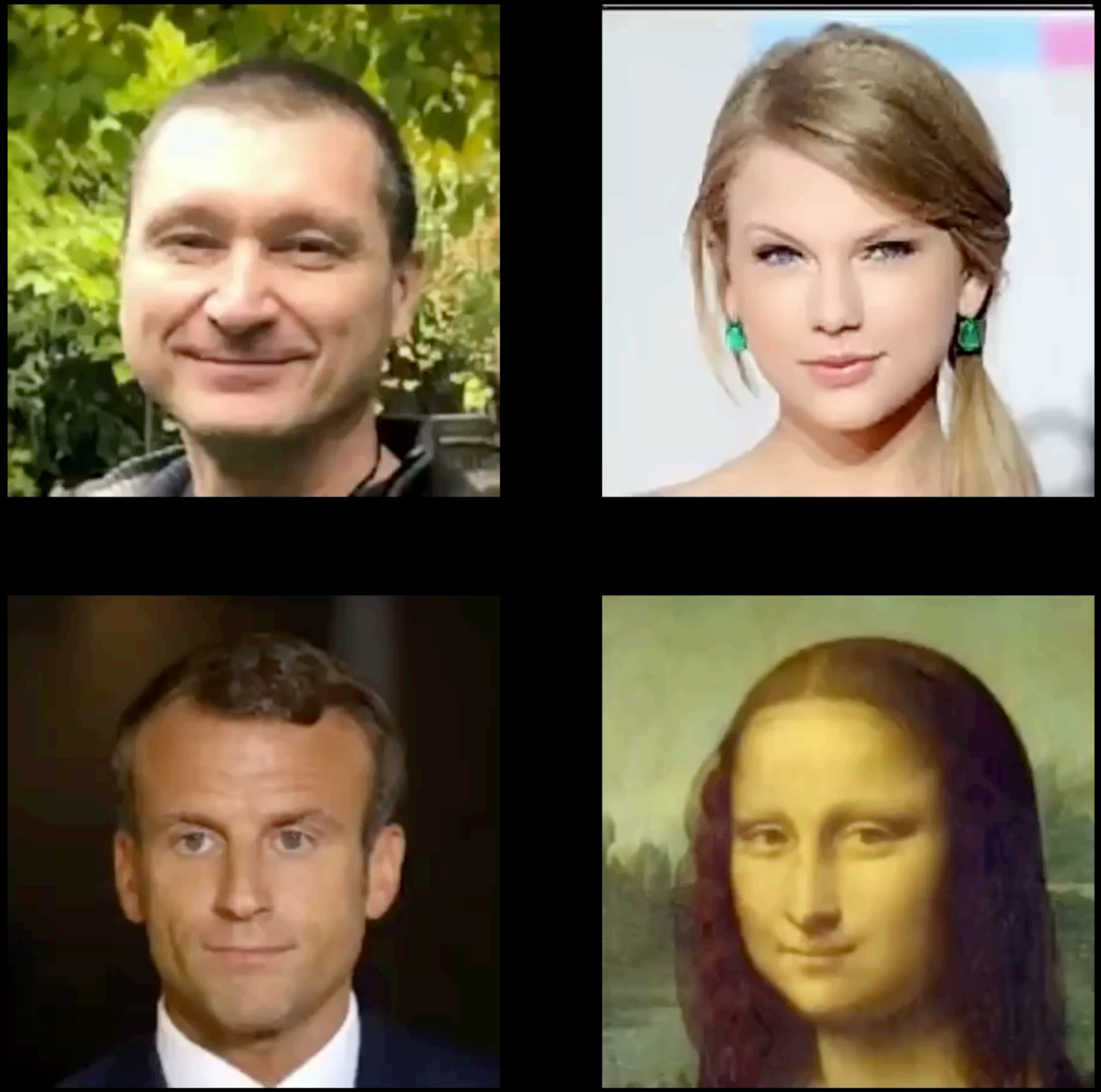
2.Image-to-video generation (Image Animation)

- ImaGINator [Wang et al., WACV'20]
- LIA [Wang et al., ICLR'22]

Latent Image Animator (LIA):Learning to Animate Images via Latent Space Navigation

[Wang et al., ICLR’22]

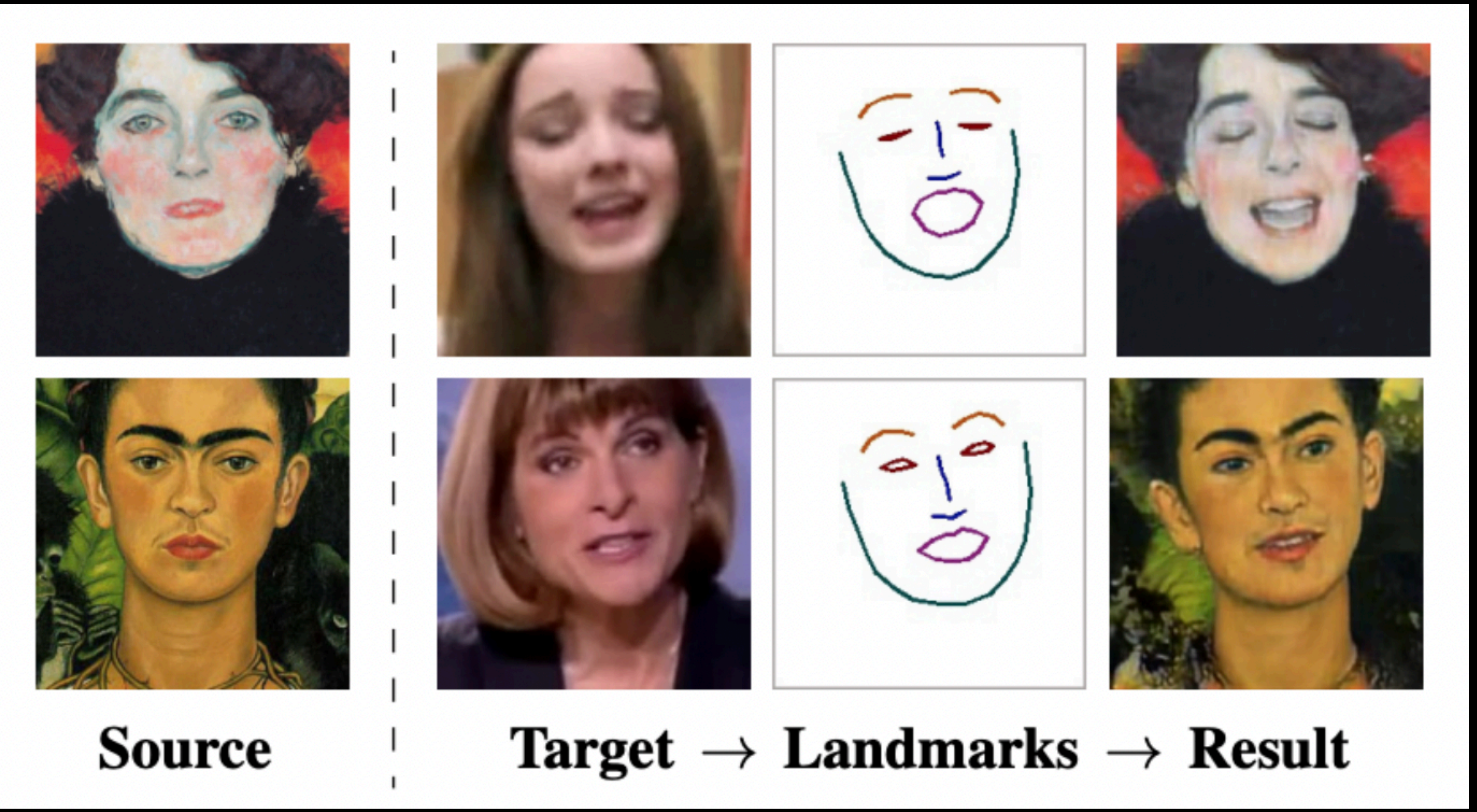
Driving video



Generated videos

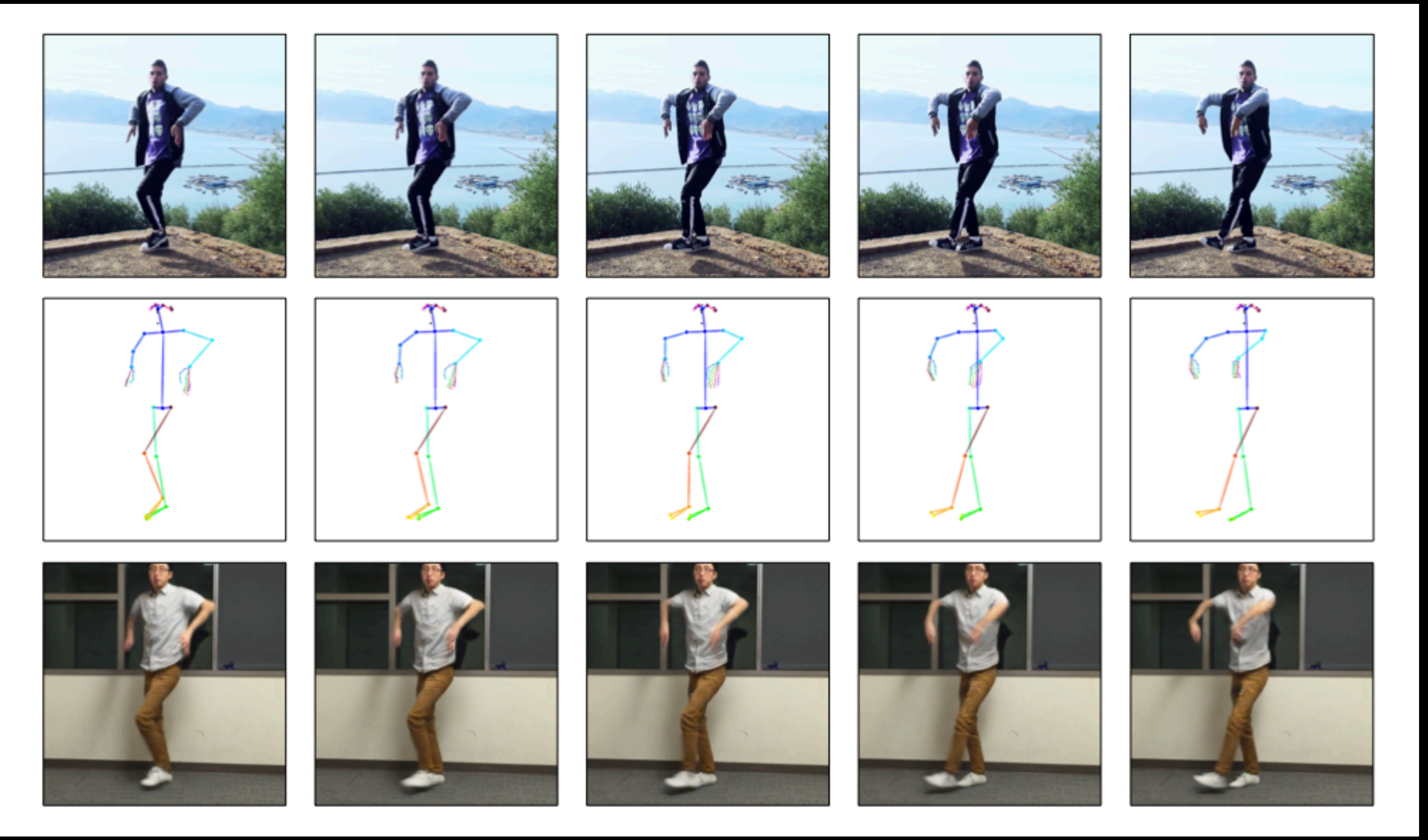
LIA: Related work

2D landmarks



[Thies et al., ICCV'19]

2D Human poses

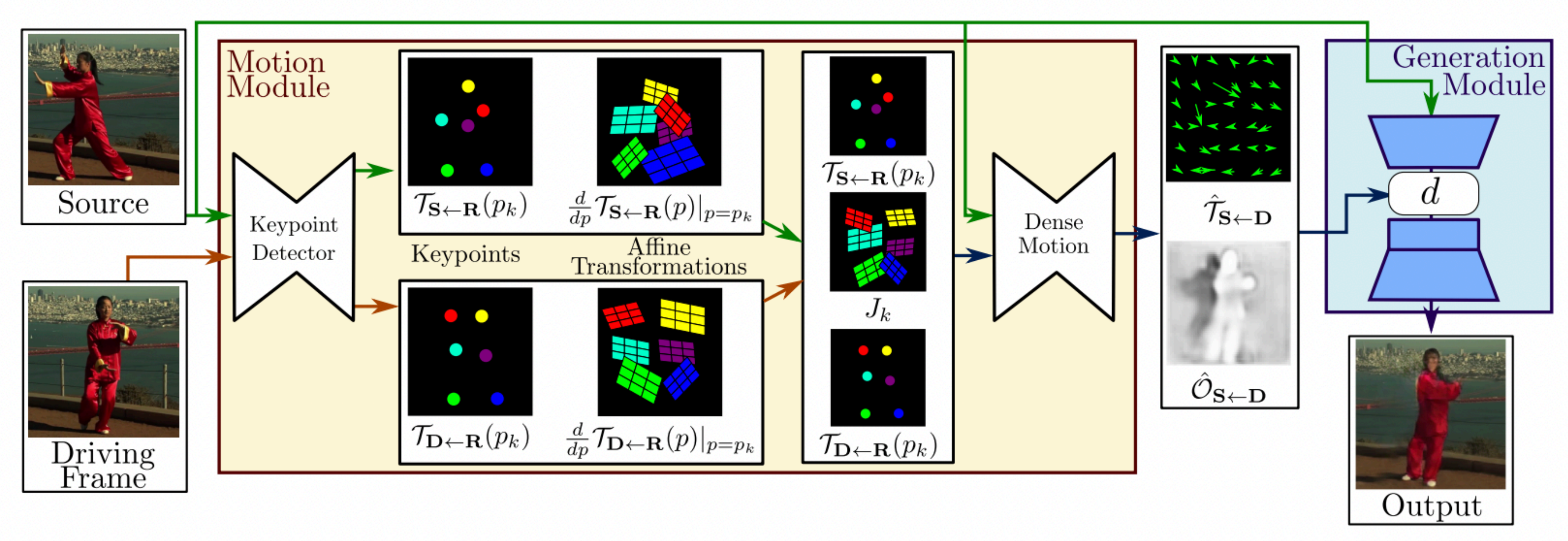


[CHAN et al., ICCV'19]

Offline extracting explicit structure representations, e.g., landmarks and poses

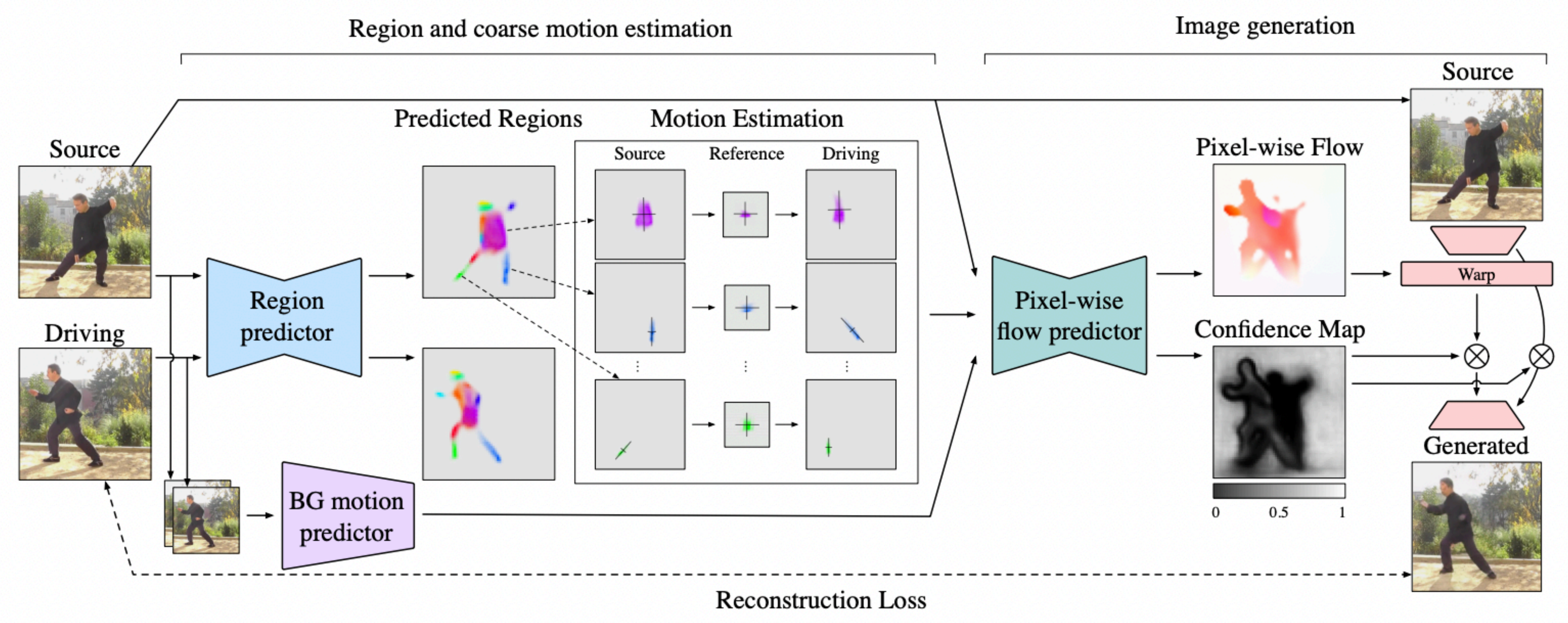
LIA: Related work

2D keypoints



FOMM [Siarohin et al., NeurIPS'19]

2D regions



MRAA [Siarohin et al., CVPR'21]

Online predicting explicit structure representations, e.g., landmarks and regions

Our goal:

Image animation without explicit structure representations

LIA: Transformation in image space?

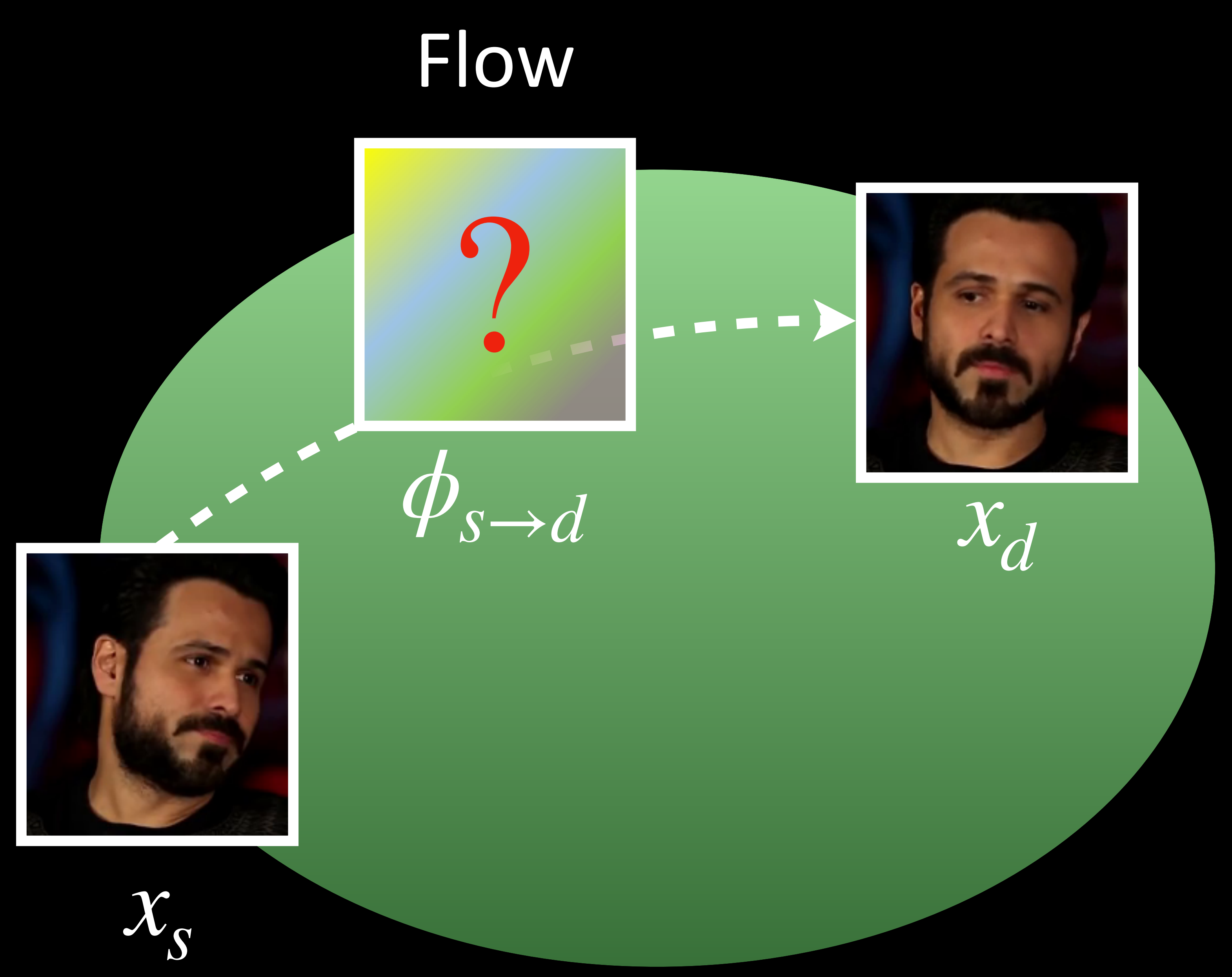


Image Space

LIA: From latent space to flow field space

LIA: Linear navigation

$$z_{s\to d} = z_{s\to r} + w_{r\to d}$$

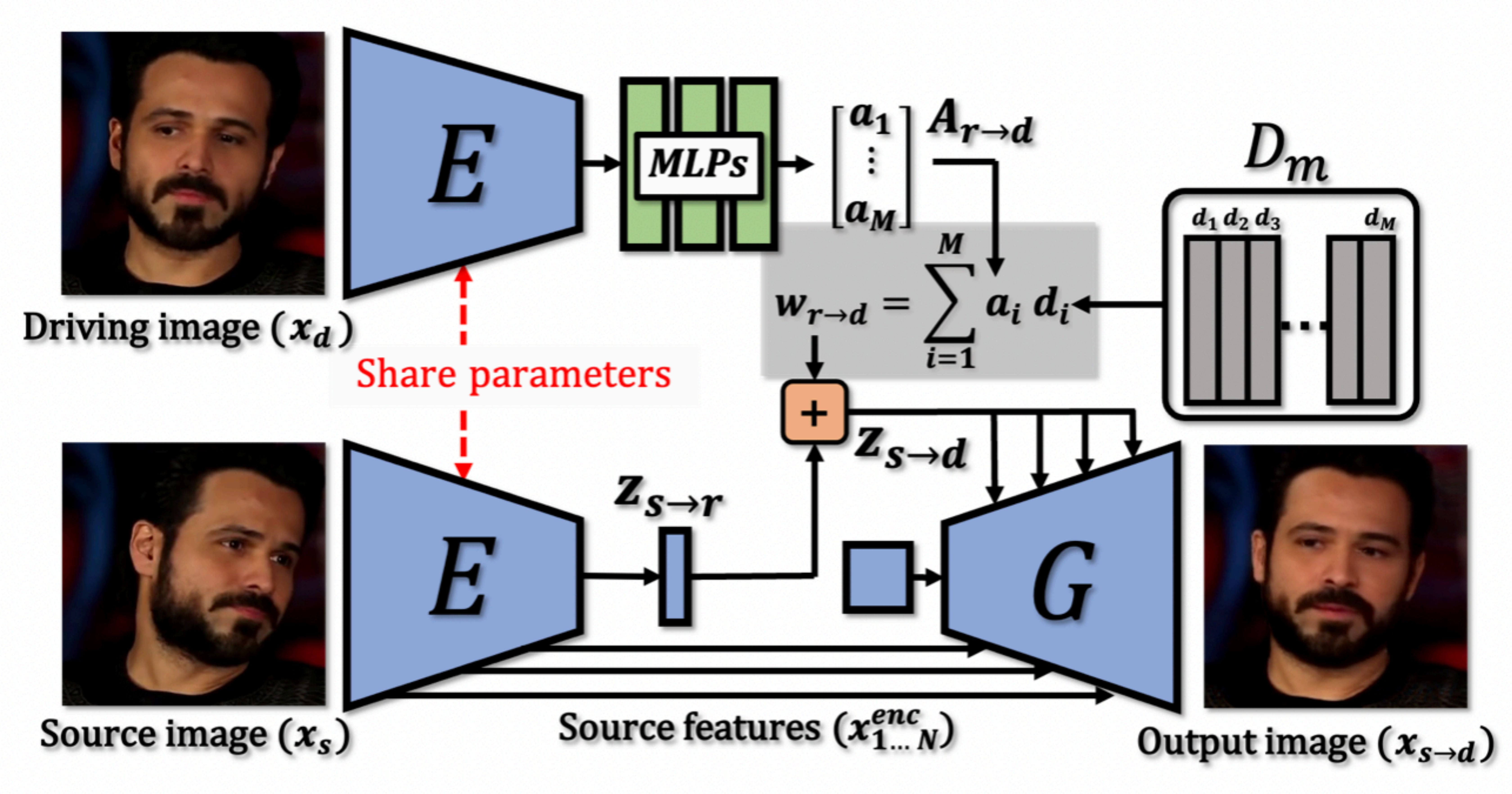
LIA: Linear Motion Decomposition (LMD) — InMoDeGAN

$$z_{s\to d} = z_{s\to r} + w_{r\to d} \qquad w_{r\to d} = \sum_{i=1}^{N} a_i d_i$$

LIA: Linear Motion Decomposition (LMD) — InMoDeGAN

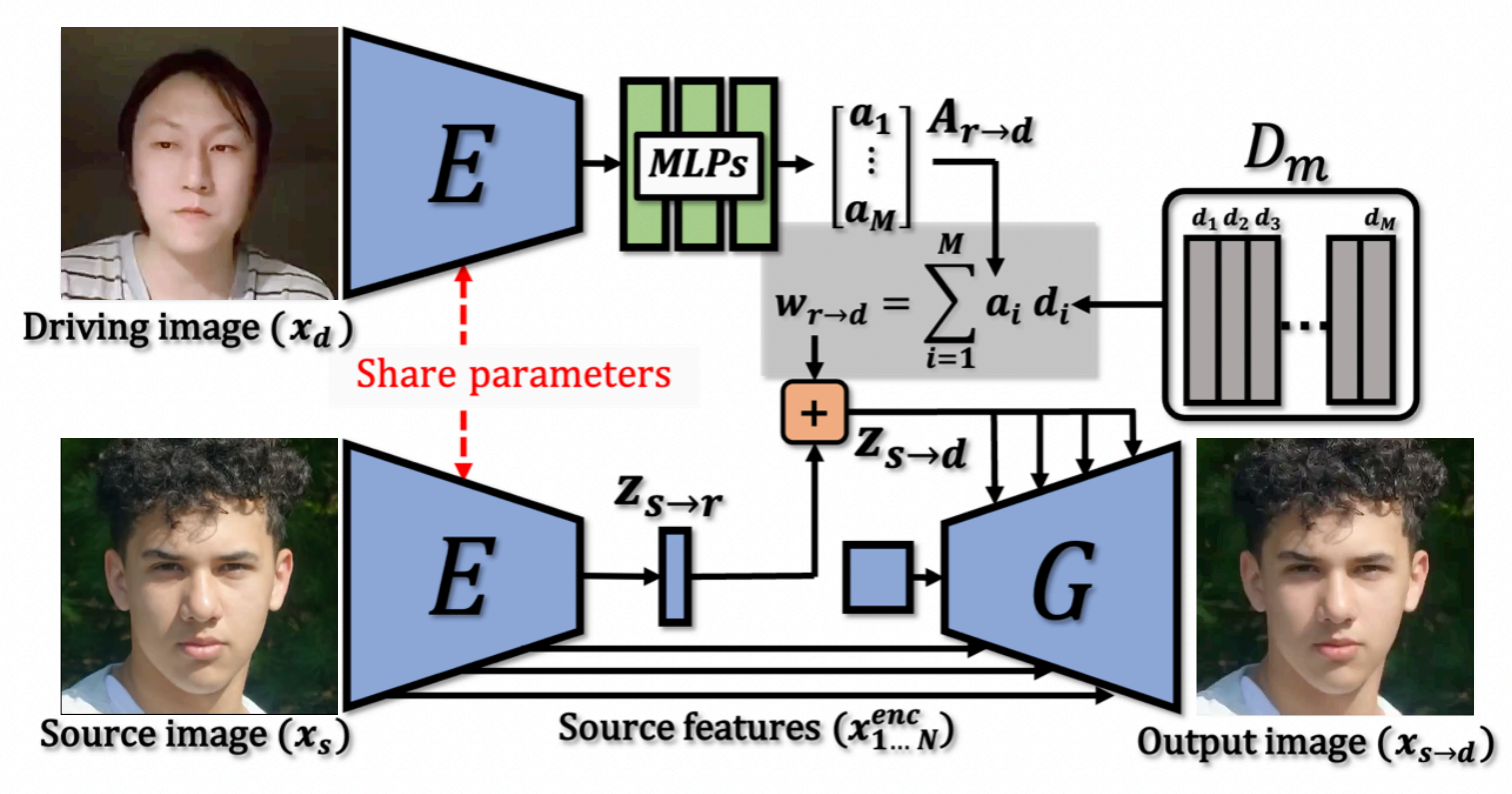
$$z_{s\to d} = z_{s\to r} + \sum_{i=1}^{N} a_i d_i$$

LIA: Overview — Training



Self-supervised learning

LIA: Overview — Inference



x_d and x_s can be different identities during inference

Comparison with SOTA

Ours MRAA FOMM Ours MRAA FOMM

LIA: Results (Taichi)

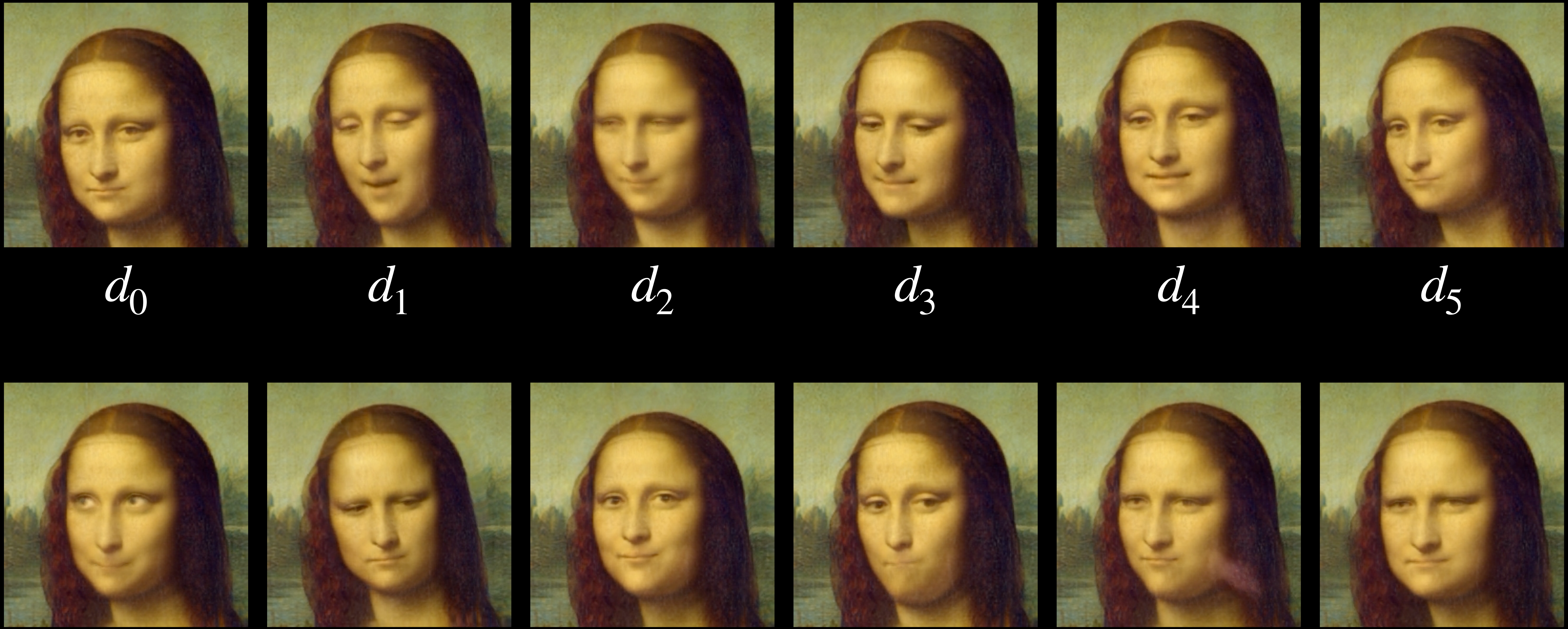
Driving video

Generated videos

Subject1 Subject2 Driving

LIA: Latent Space Interpretability

Manipulation of motion directions



d_0 d_1 d_2 d_3 d_4 d_5

d_6 d_7 d_8 d_9 d_{10} d_{11}

LIA: Latent Space Interpretability

Manipulation of motion directions

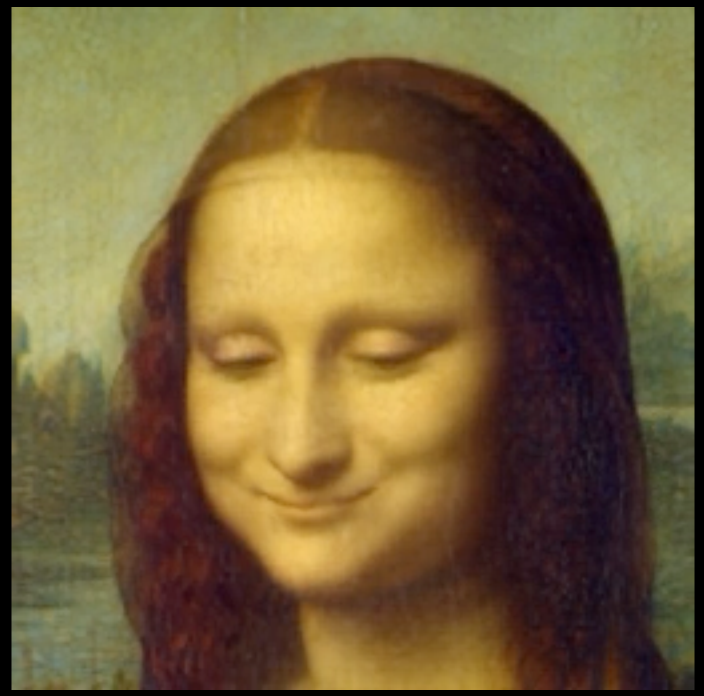

d_{12}


d_{13}


d_{14}


d_{15}


d_{16}


d_{17}


d_{18}


d_{19}

Conclusions

1. Noise-to-video generation

- InMoDeGAN [Wang et al., arXiv'21]
  - Sequence-to-sequence model (high-resolution)
  - The first method to interpret motion space

2. Image-to-video generation

- LIA [Wang et al., work in progress]
  - Sequence-to-sequence model
  - Image animation without relying on explicit structure representations

Linear Motion Decomposition (LMD)

Future directions

1. Controllability. 3D-aware, illumination, …
   - GIRAFFE (CVPR'21), CAMPARI(3DV'21), EG3D (3D-aware StyleGAN), …
2. Generalizability. Multiple scenes & objects, One-shot face & body reenactment
   - PixelNeRF (CVPR'21), LIA (ICLR'22), MRAA(CVPR'21), FOMM (NeurIPS'19), …
3. Scalability. City- or global-scale scenes. (e.g., Block-NeRF, City-NeRF)
   - Block-NeRF (CVPR'22), City-NeRF, …
4. Interpretability. Latent space & network
   - InterFaceGAN (CVPR'20), InMoDeGAN, …
5. Machine learning. GANs, VAEs, Diffusion Model (DDPM), Flow, …
6. Learning from synthetic data. Video understanding, robot learning, …
   - Varol et al. (IJCV'20), AVID (RSS'20), GCL (CVPR'21), …

Thank you !

We are hiring Interns/Engineers/Researchers at Shanghai AI Lab on deep generative models (GANs, Diffusion Models, …) for image/video generation, animation etc..
If you are interested, please contact

wangyaohui@pjlab.org.cn