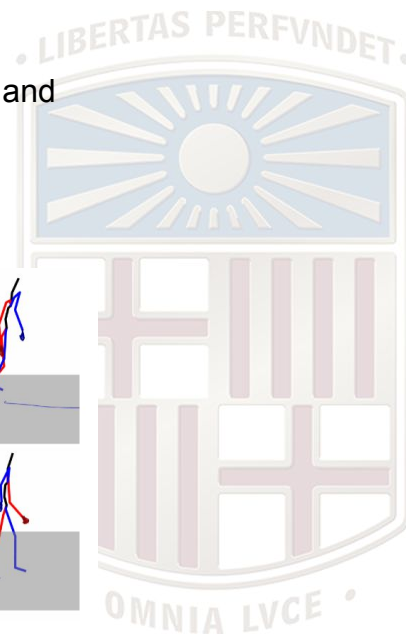
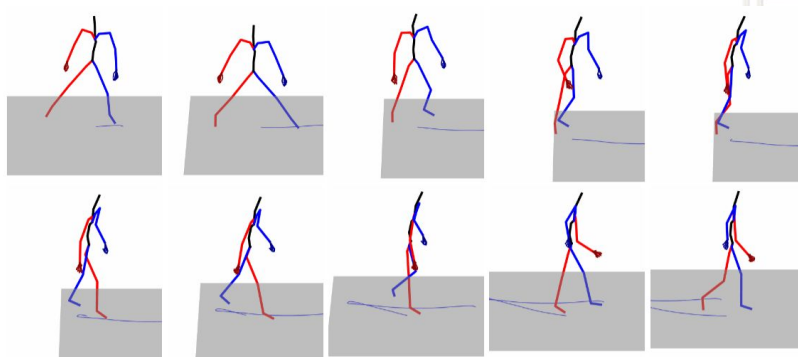


Motion Binary Latent Diffusion

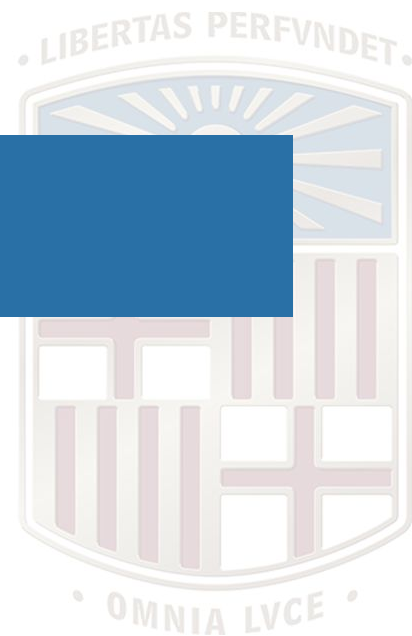
Àlex Pujol

Supervisors: Sergio Escalera, Germán Barquero and
Cristina Palmero

25 January 2024



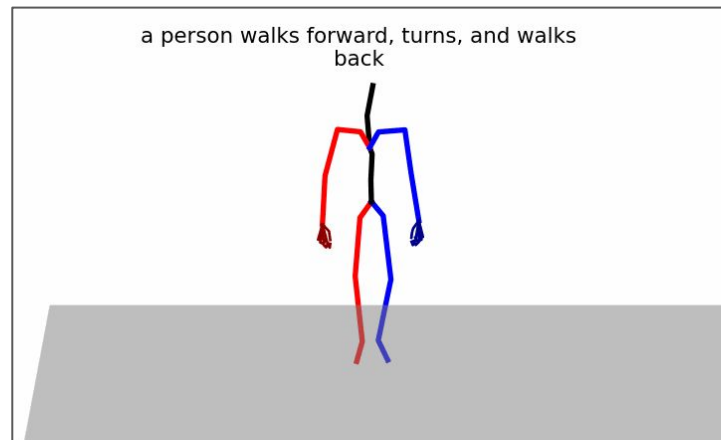
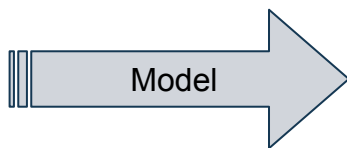
Introduction



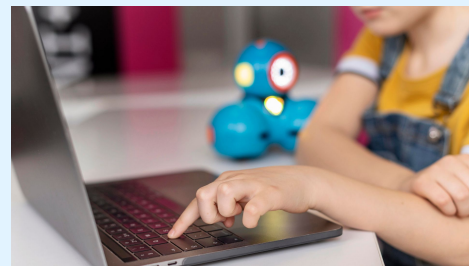
Problem definition

Text-to-motion: *“Design and train a model to generate plausible, natural and diverse range of human motion sequences, based on an input prompt text.”*

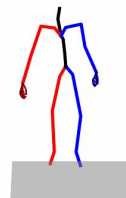
“a person walks
forward, turns,
and walks back”



Motivation



a person places their hands on their hips and stretches.



Challenges

Data Scarcity:

Expensive, requires expertise,
resource intensive, low accuracy ...

Motion Complexity:

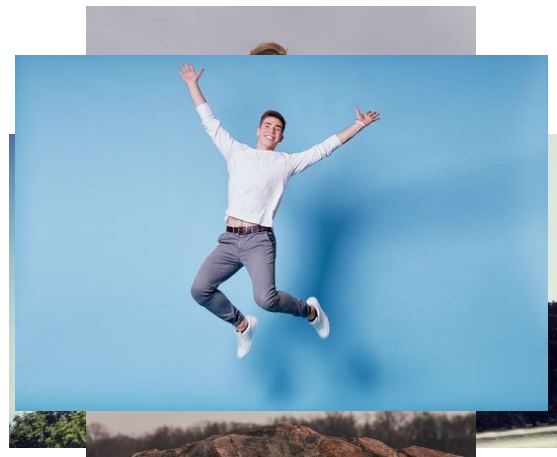
Physic laws, anatomy, intention,
personality, ...

NLP vs Motion:

Map between language and motion,
semantic and meaning

Many-to-many:

many motions $\longleftrightarrow$ many descriptions



"a person is jumping"

"guy jumping"

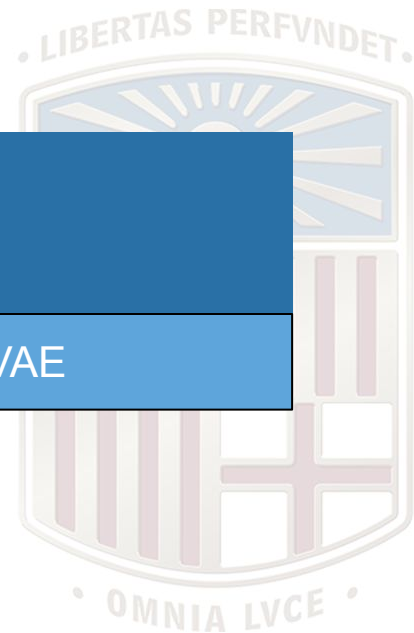
"doing a backflip"

"laying upside-down"

Background

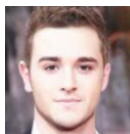
Diffusion Models

VQ-VAE



Diffusion Models

VQ-VAE

 $\mathbf{x}_0$

Gaussian

Forward trajectory:

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \mathcal{N}(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}\sqrt{1-\beta_t}, \mathbf{I}\beta_t)$$

Reverse trajectory:

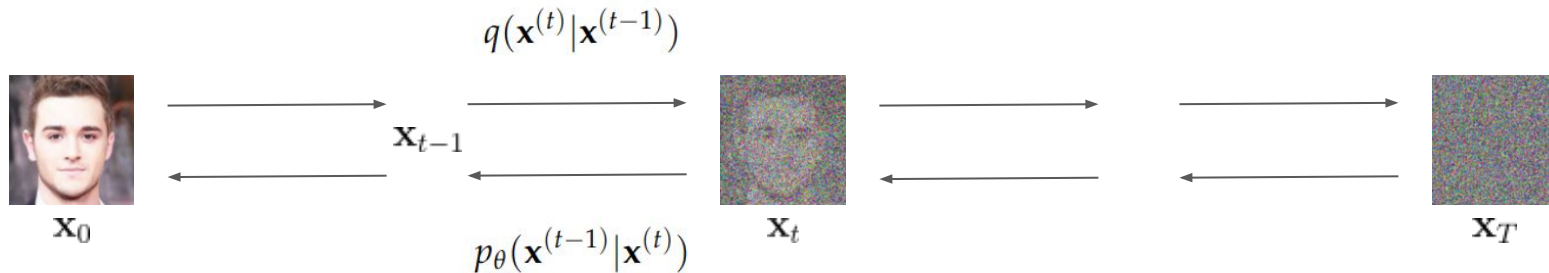
$$p_{\theta}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t-1)}; \mu_{\theta}(\mathbf{x}^{(t)}, t), \Sigma_{\theta}(\mathbf{x}^{(t)}, t))$$

Binomial

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \mathcal{B}(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}(1-\beta_t) + 0.5\beta_t)$$

$$p_{\theta}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \mathcal{B}(\mathbf{x}^{(t-1)}; \mathbf{b}_{\theta}(\mathbf{x}^{(t)}, t))$$

Diffusion Models



Bayes' Rule: $q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)})$

Gaussian

Binomial

Learning Objective:
Forward trajectory

$$D_{\text{KL}}(q(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}, \mathbf{x}^{(0)}) || p_\theta(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}))$$

Reverse trajectory:

$$p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t-1)}; \mu_\theta(\mathbf{x}^{(t)}, t), \Sigma_\theta(\mathbf{x}^{(t)}, t))$$

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \mathcal{B}(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}(1 - \beta_t) + 0.5\beta_t)$$

$$p_\theta(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \mathcal{B}(\mathbf{x}^{(t-1)}; \mathbf{b}_\theta(\mathbf{x}^{(t)}, t))$$

Diffusion Models

VQ-VAE

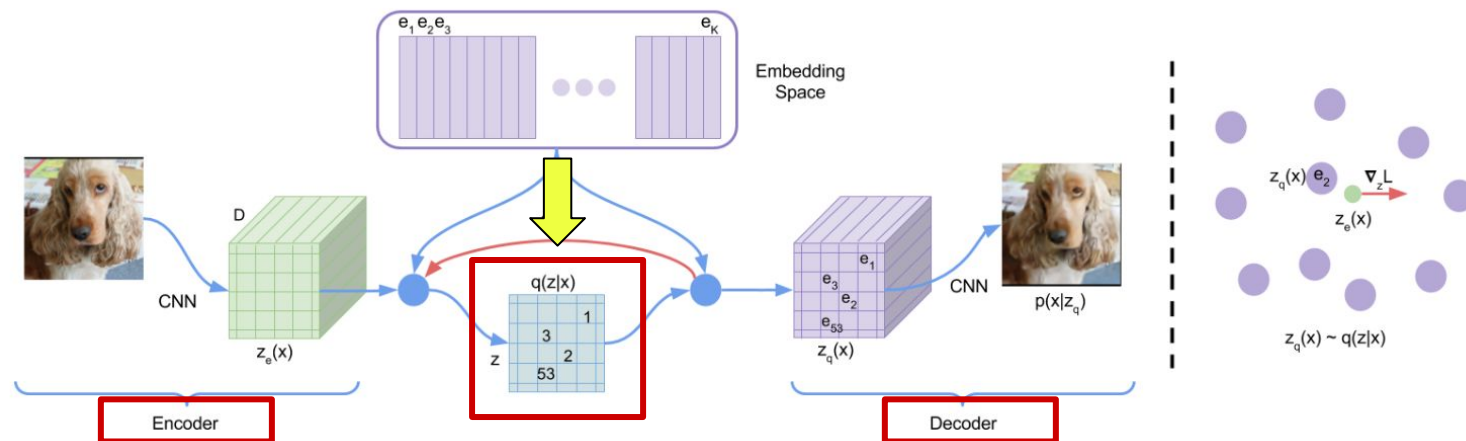


figure from [1]

Binary Latent Diffusion

figure from [1]

Binary Quantization

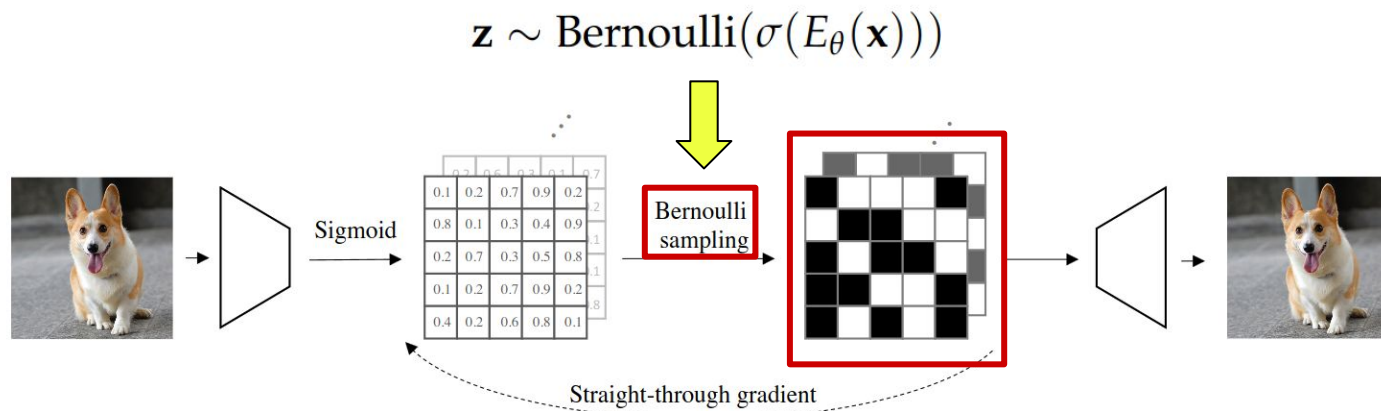
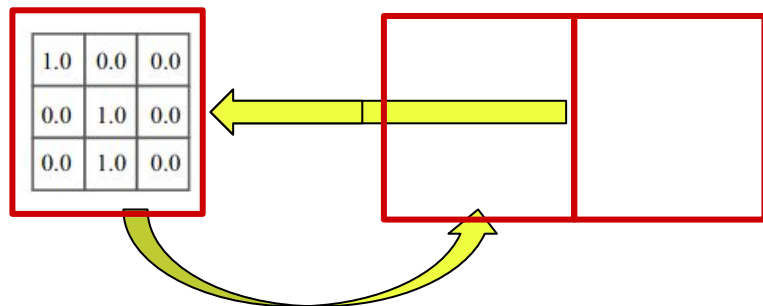


figure from [2]

Bernoulli Diffusion



Forward trajectory:

$$q(\mathbf{z}^{(t)}|\mathbf{z}^{(t-1)}) = \mathcal{B}(\mathbf{z}^{(t)}; \mathbf{z}^{(t-1)}(1 - \beta_t) + 0.5\beta_t)$$

Reverse trajectory:

$$p_{\theta}(\mathbf{z}^{(t-1)}|\mathbf{z}^{(t)}) = \mathcal{B}(\mathbf{z}^{(t-1)}; f_{\theta}(\mathbf{z}^{(t)}, t))$$

Target:

$$\text{Original: } \hat{\mathbf{z}}^{(0)} = f_{\theta}(\mathbf{z}^{(t)}, t)$$

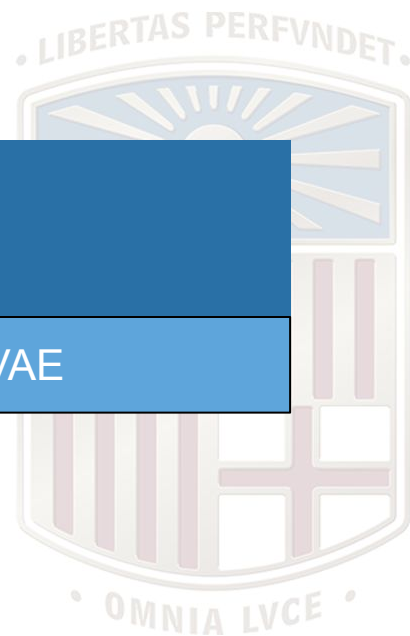
$$\text{Pflip: } \mathbf{z}^{(0)} \oplus \mathbf{z}^{(t)} = f_{\theta}(\mathbf{z}^{(t)}, t)$$

figure from [2]

State-of-the-art

Diffusion Models

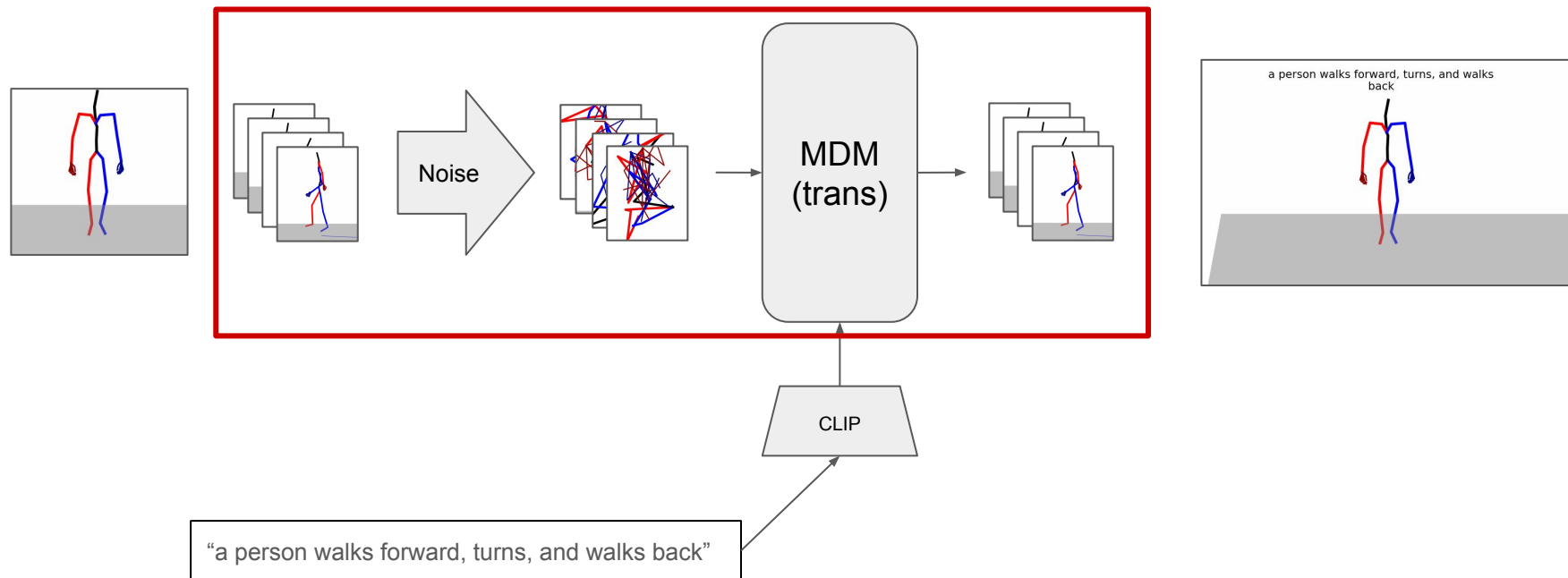
VQ-VAE



Diffusion Models

VQ-VAE

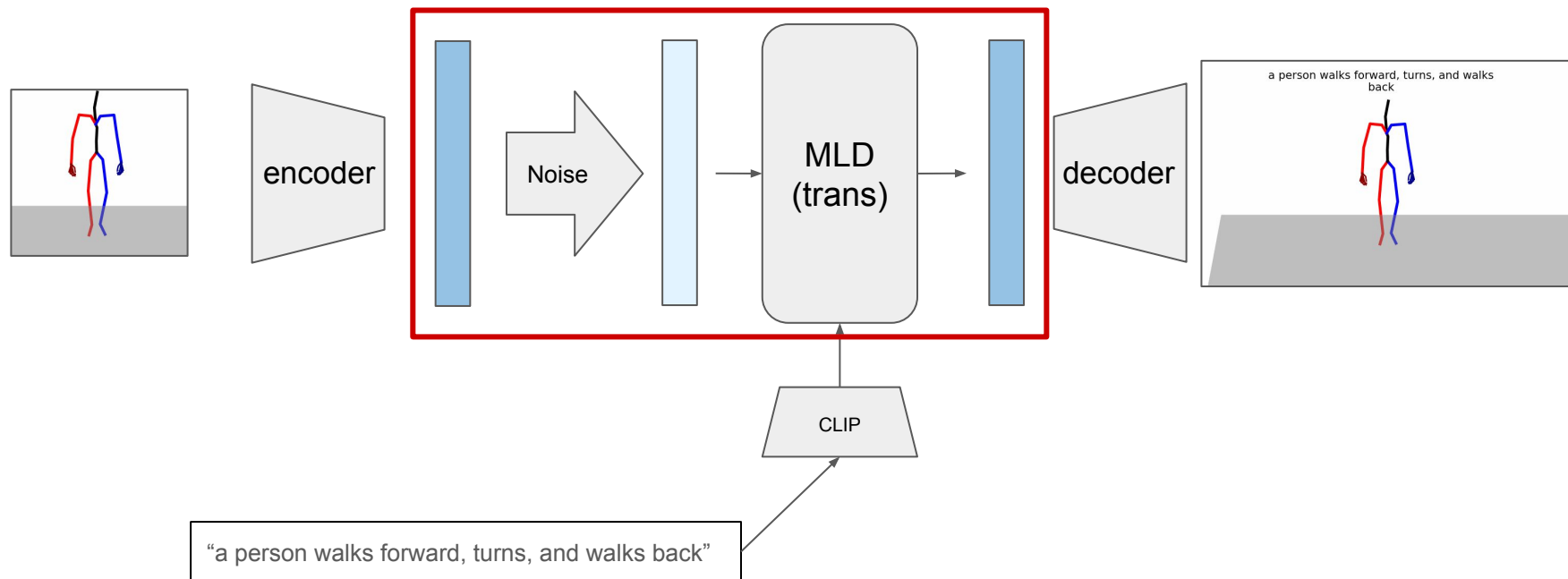
MDM:



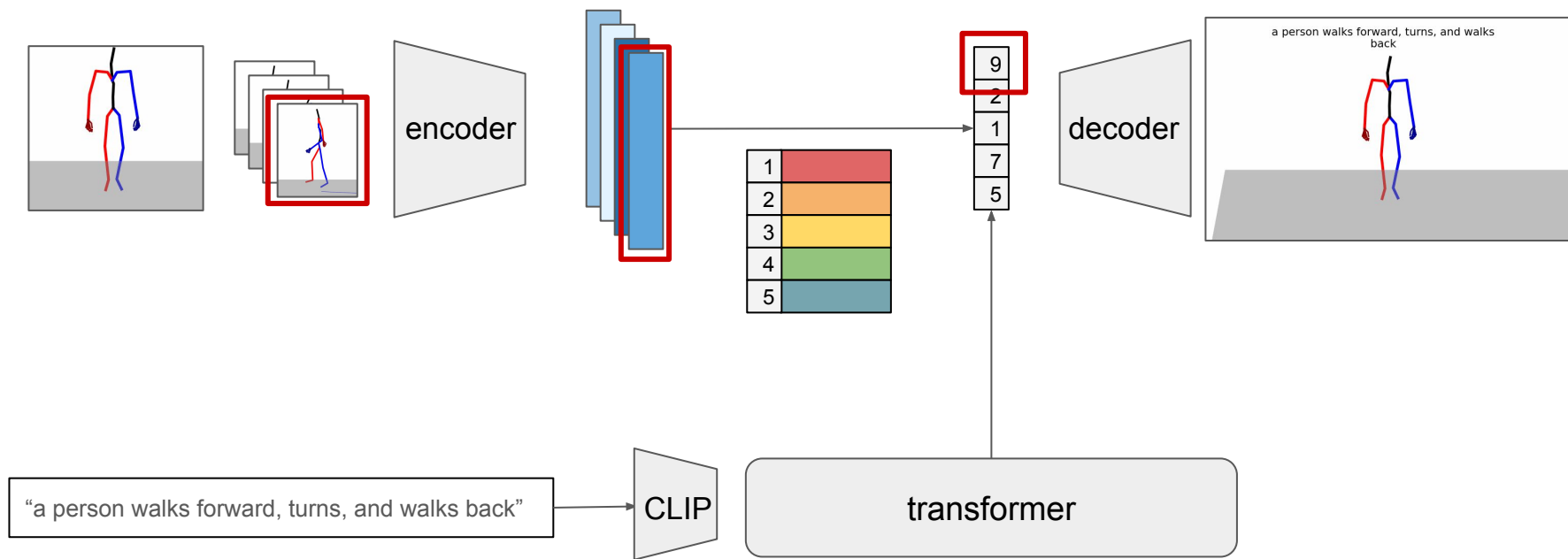
Diffusion Models

VQ-VAE

MLD:



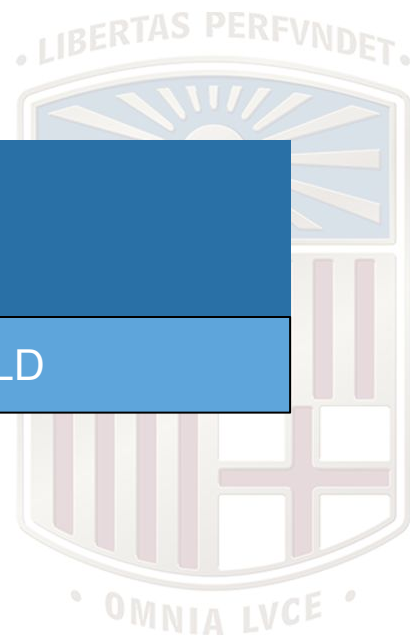
PoseGPT, T2M-GPT and MotionGPT:



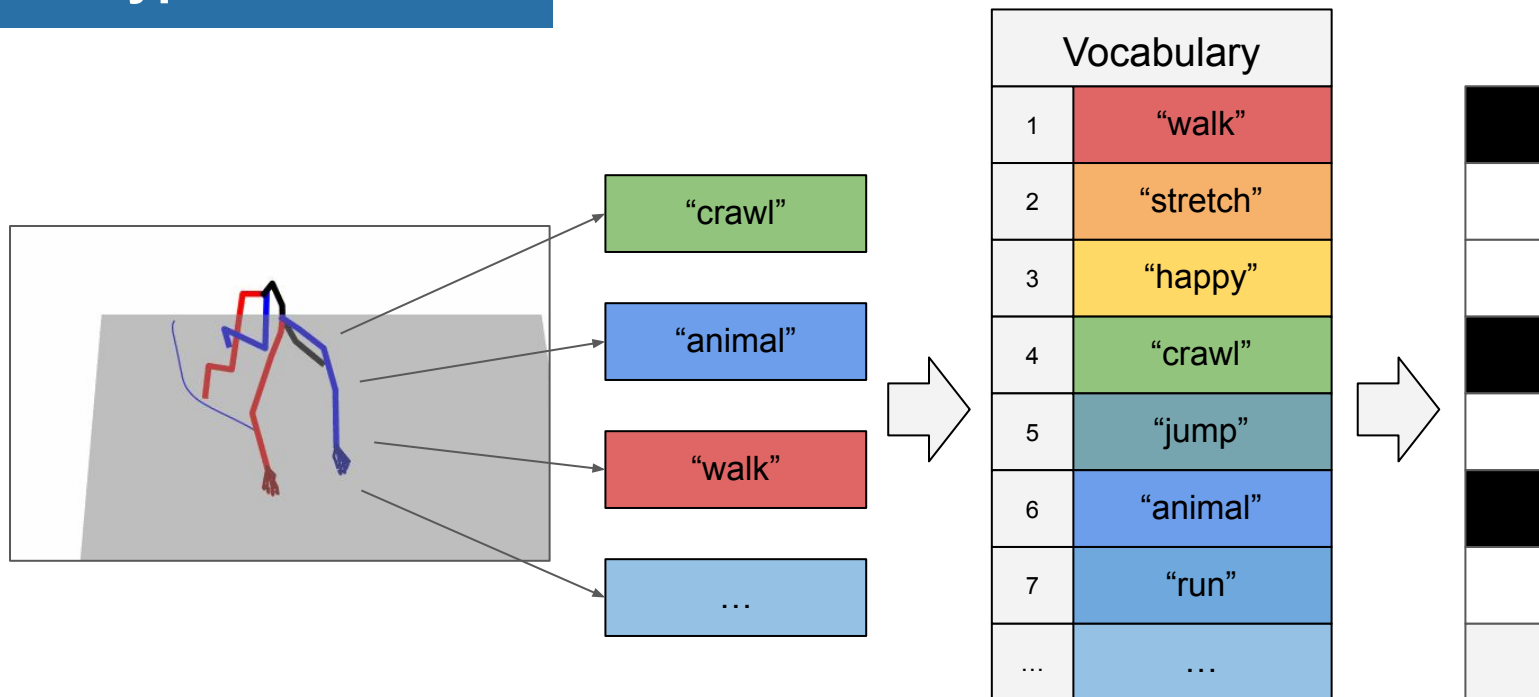
Methodology

MBVAE

MBLD

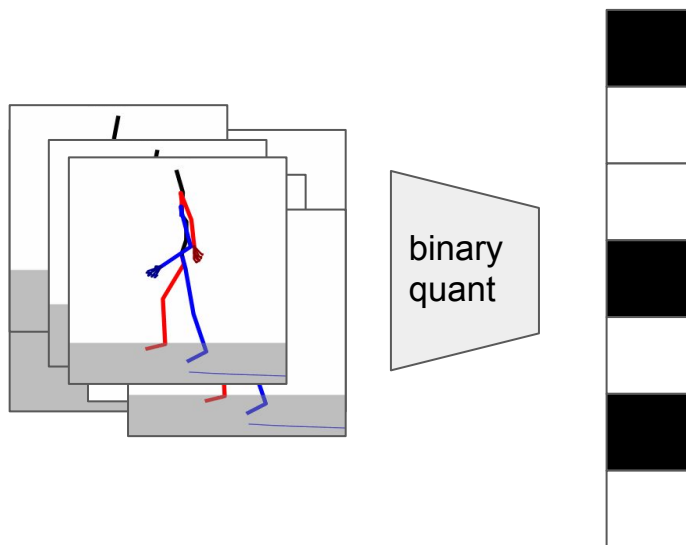


Hypothesis



MBVAE

MBLD



Frame-wise: per frame encoding.

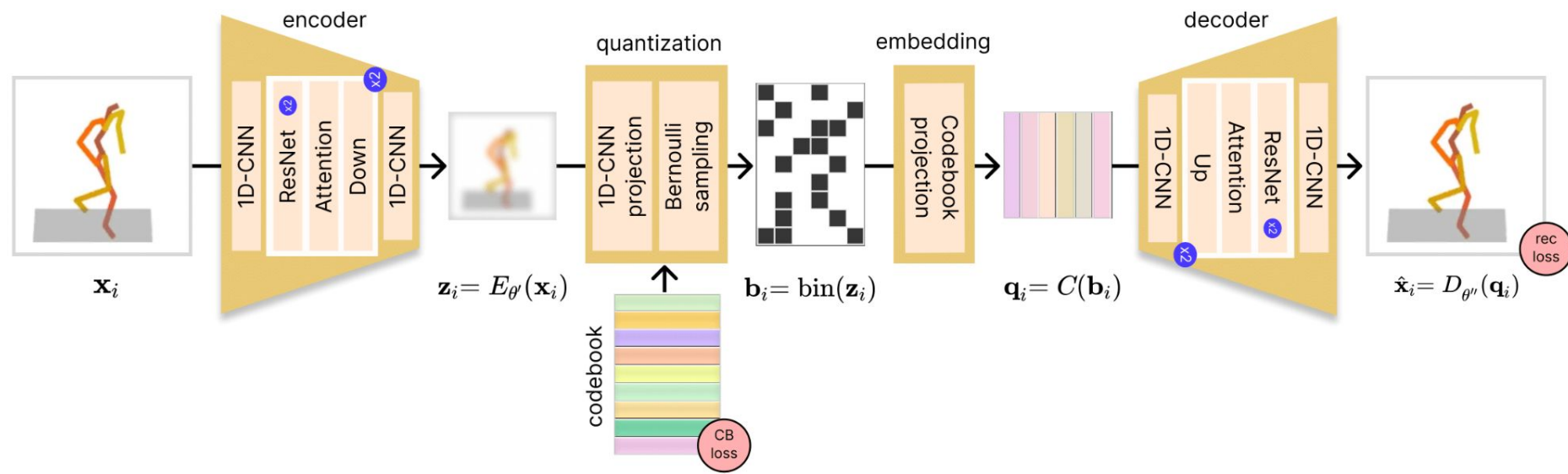
Some-frames: chunks of frames encoding.

Full-sequence: entire motion encoding.

MBVAE

MBLD

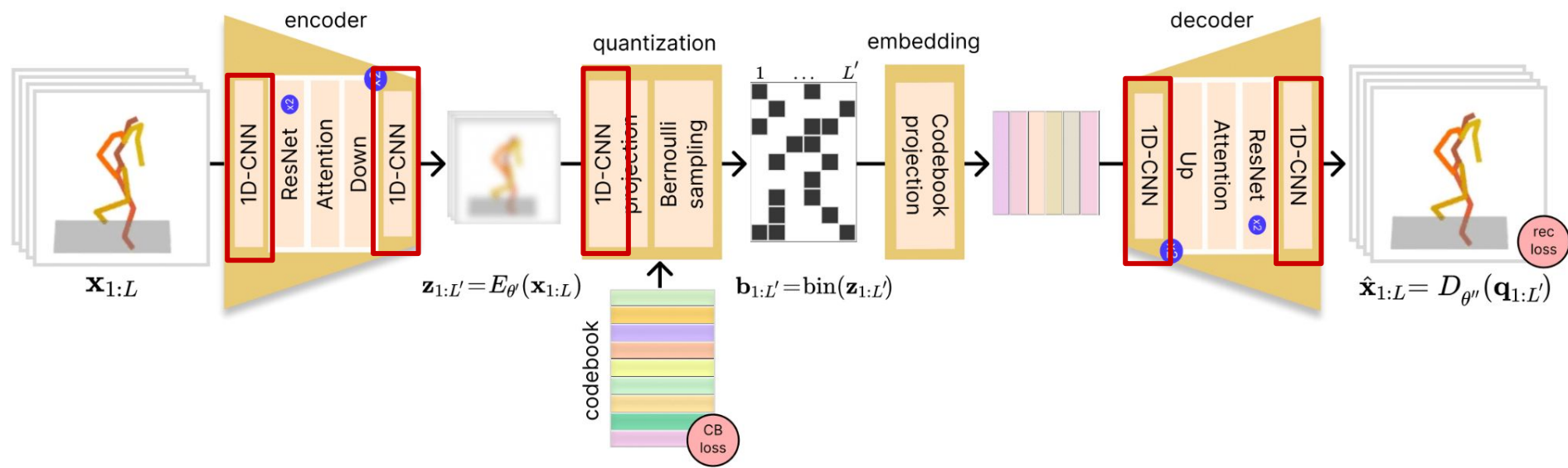
Frame-wise

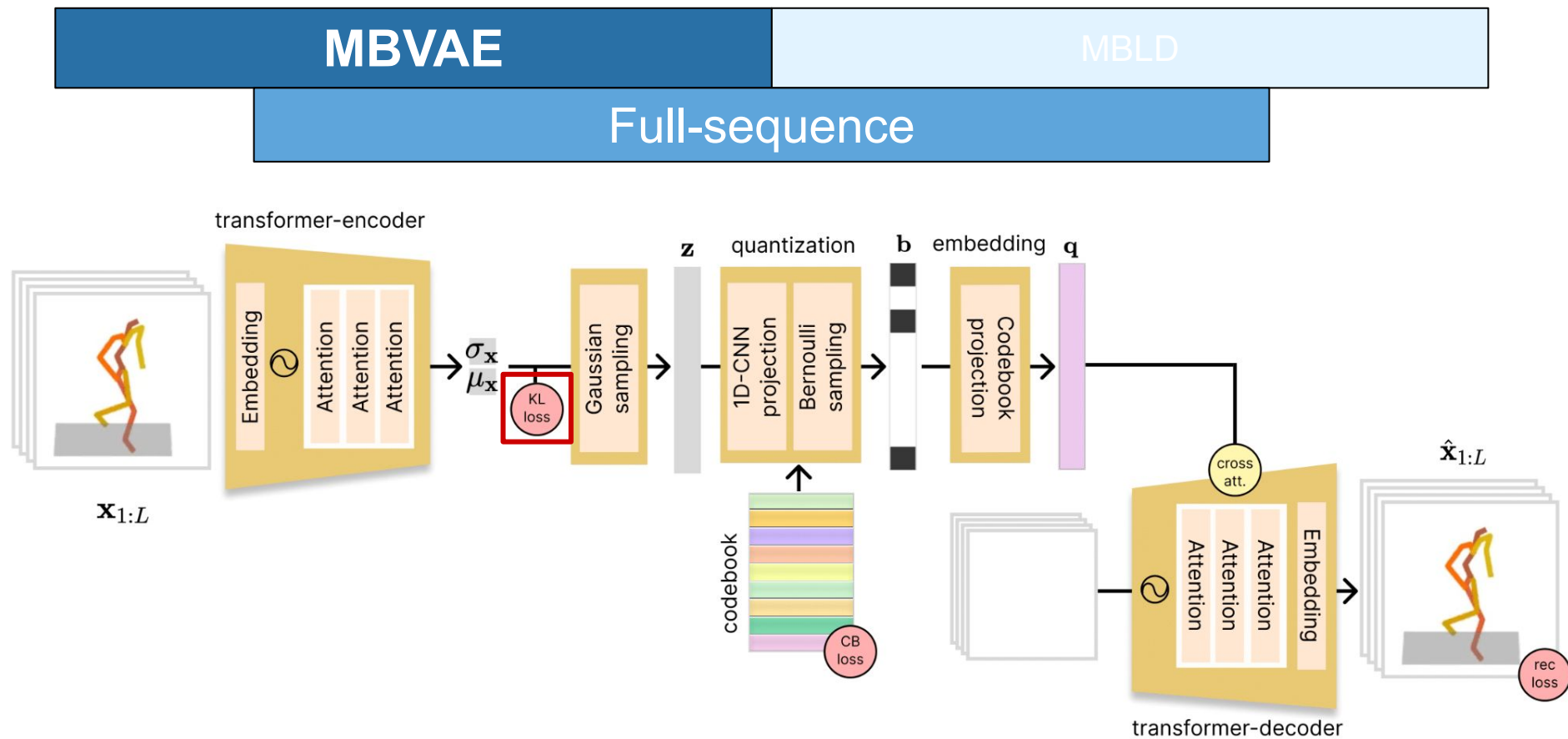


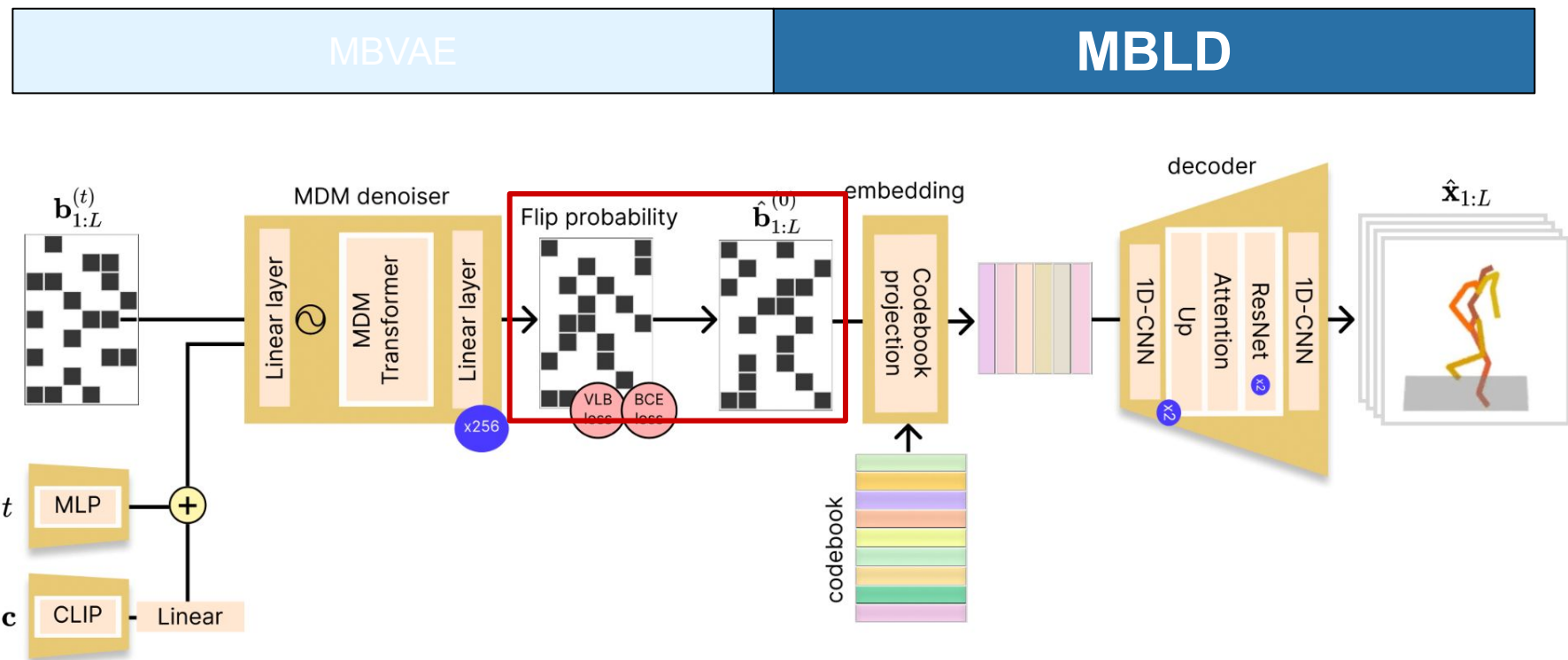
MBVAE

MBLD

Some-frames



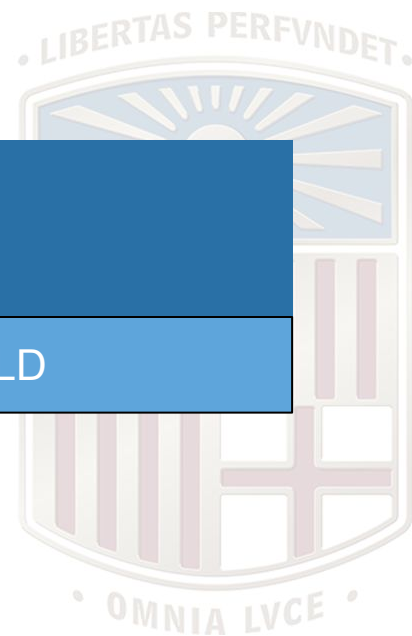




Experiments

MBVAE

MBLD

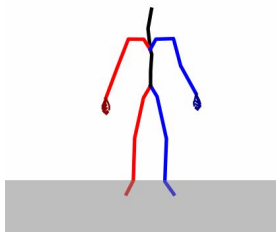


Dataset

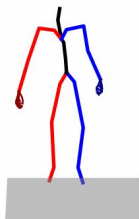
Subset of HumanML3D:

- Number of samples: 1,528 motion clips
- Text prompts: 4 text description per clip
- Length range: [150, 196] frames per clip
- Pose representation: 263-dimensional vector of position and rotation

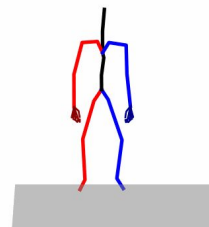
a figure paces confidently back and forth



a person places their hands on their hips and stretches.



a person balances along a beam



Evaluation

Fidelity:

MSE: *average of the squared differences between the prediction and the real motion.*

FID: *estimates the distance between the distribution of a feature space of the generated motion and the ground truth.*

Diversity:

DIV: *global diversity of the model.*

MM: *measures the diversity but conditioning to a set C of different text prompts.*

Condition

Consistency:

R-precision: *measures the accuracy of the model to generate motions that satisfy a given condition.*

Top-1

Top-2

Top-3

MBVAE

MBLD

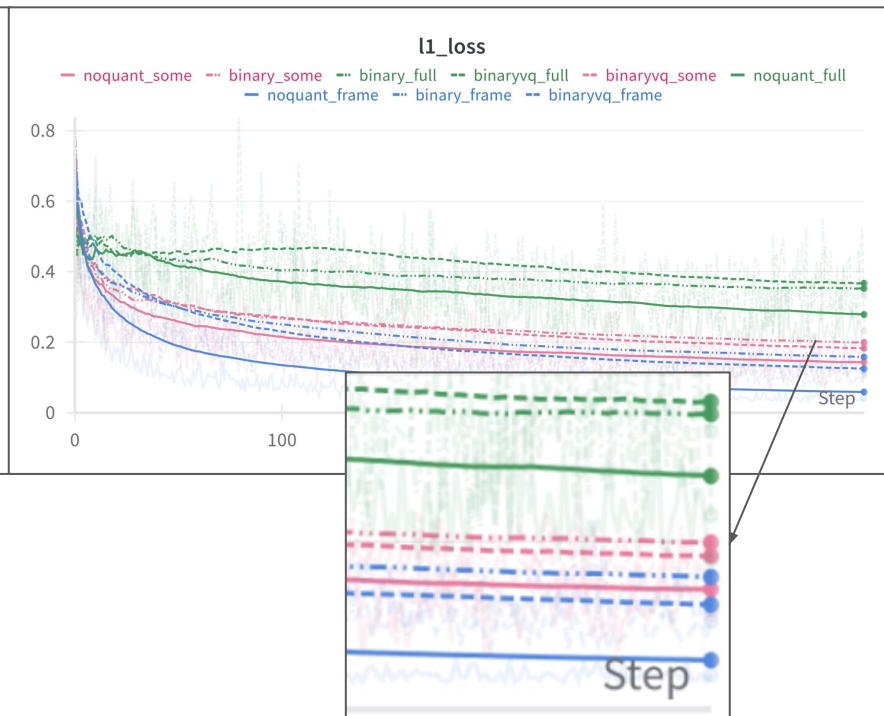
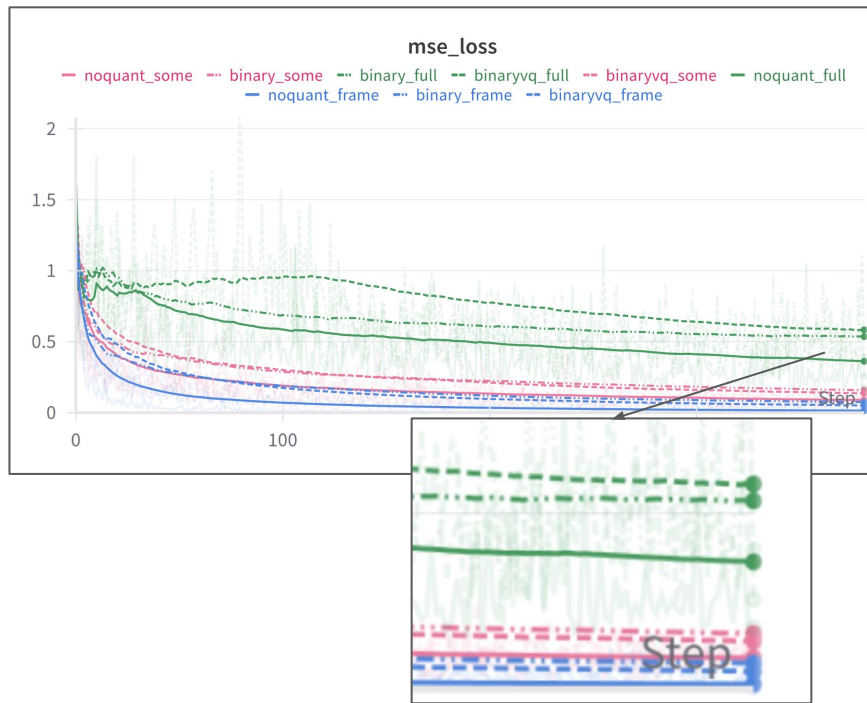
1. Is the binary latent space able to encode motion data?
2. Which is the model that better performs?
3. Is it memory efficient?

Model	Binary	Binary-VQ	No-Quant
Frame-wise	50,176bits $\simeq$ 6.3kB	100,352bits $\simeq$ 12.6kB	1,605,632bits $\simeq$ 200.7kB
Some-frames	784bits $\simeq$ 0.1kB	1,568bits $\simeq$ 0.2kB	25,088bits $\simeq$ 3.1kB
Full-sequence	256bits $\simeq$ 0.03kB	256bits $\simeq$ 0.03kB	8,192bits $\simeq$ 1.02kB

Memory footprint

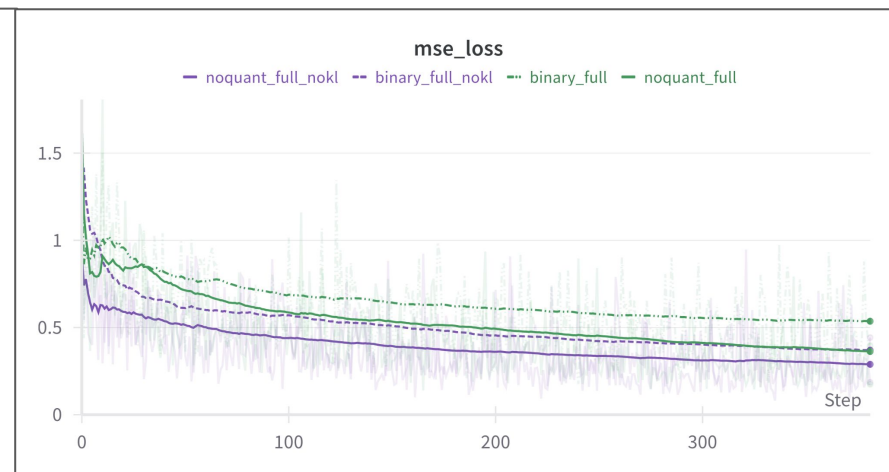
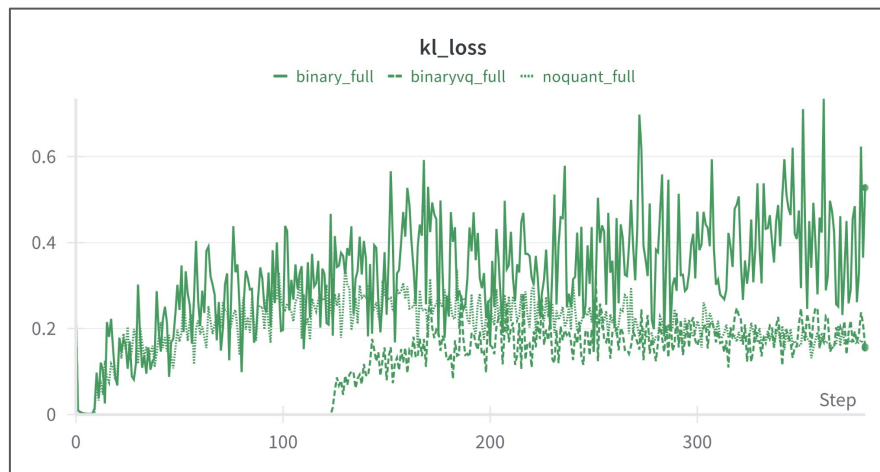
MBVAE

MBLD



MBVAE

MBLD



MBVAE

MBLD

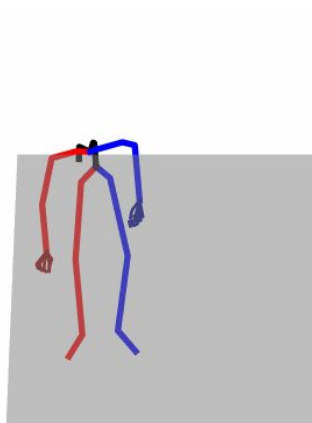


Around 21% of
bits are 1 for
every model!

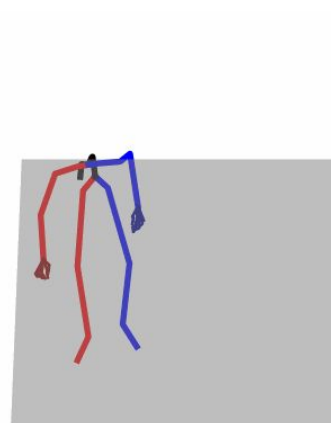
MBVAE

MBLD

Input



Output



MBVAE

MBLD

Exp. 1

1. Is the Bernoulli diffusion able to generate plausible motion with **Frame-wise**?
2. Which is the best target for the denoising function?

Target: **Original** **P-flip**

Exp. 2

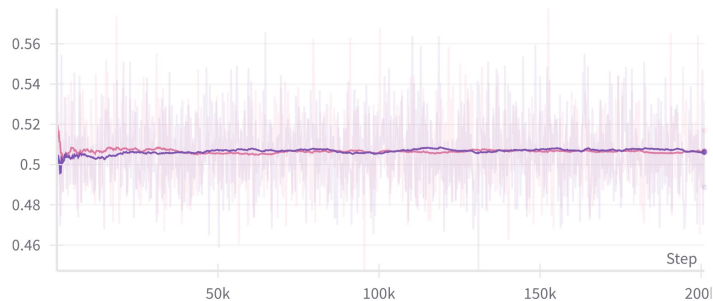
1. What about using **Full-sequence**?
2. Does the model properly fit the evaluation metrics?

Exp. 1

MBLD

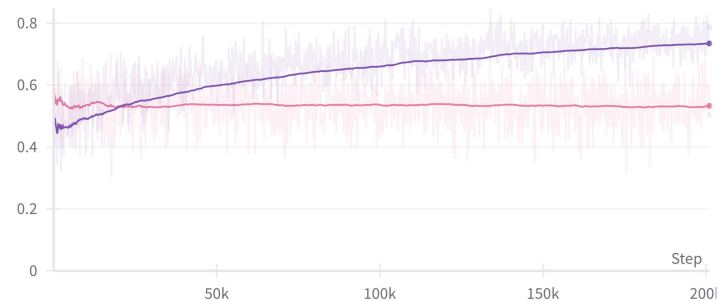
vib_loss

— p-flip-bdiff — origin-bdiff



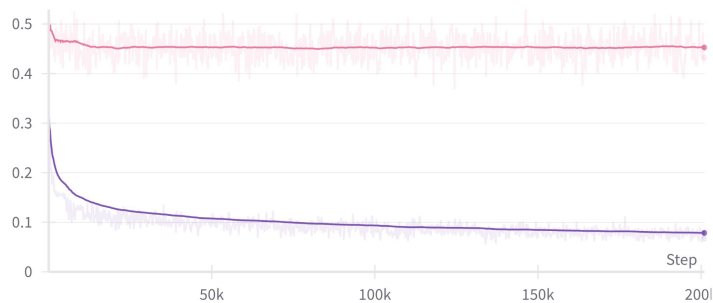
acc

— p-flip-bdiff — origin-bdiff

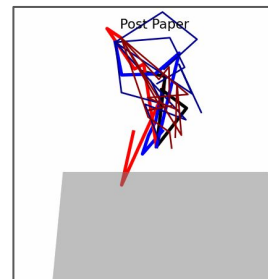


bce_loss

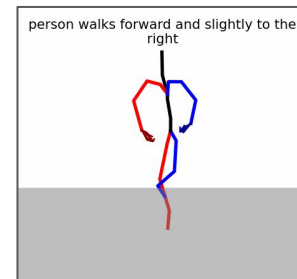
— p-flip-bdiff — origin-bdiff



P-flip:



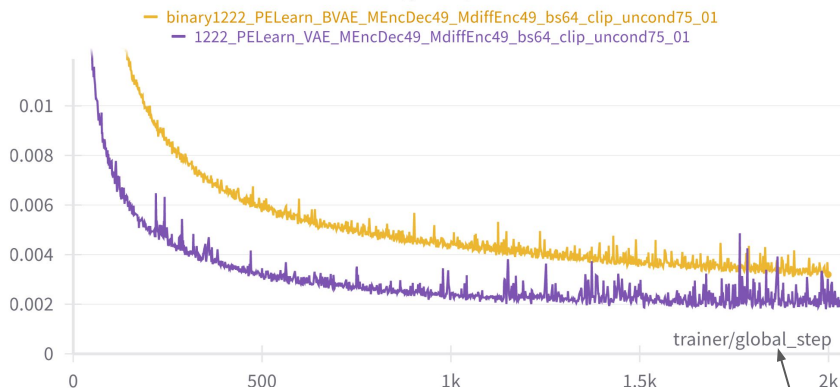
Original:



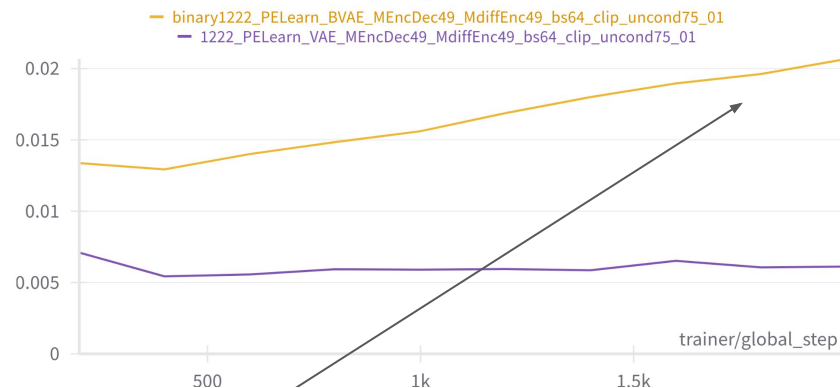
Exp. 2

MBLD

recons/joints/train



recons/joints/val



Model	FID	MM	DIV	Top-1	Top-2	Top-3
GT	-	-	9.5	0.514	0.705	0.797
MLD	17.504	4.744	5.284	0.03	0.062	0.093
MBLD	31.49	4.041	4.292	0.034	0.067	0.1

Conclusions and Future work



Conclusions

- ✗ Requires further regularization → to avoid shaky movement
- ✗ Requires more data → overfits rapidly
- ✓ Binary latent space successfully **encode pose data**
- ✓ Can be **memory efficient**
- ✓ Setting Original as target better than P-flip, as in MDM!
- ✓ Comparable to MLD in terms of Diversity and Condition Consistency

Future work

- Add extra regularization:
 - Adversarial loss
 - Perceptual losses (use joints, rotations, velocities)
- Train on larger dataset:
 - Bigger subset of HumanML3D
 - MotionX

References

- [1] Oord, Aaron van den, Oriol Vinyals, and Koray Kavukcuoglu (2018). Neural Discrete Representation Learning.
- [2] Wang, Ze et al. (2023). Binary Latent Diffusion.

Thank you for your attention

person sits down then puts left leg over right knee

