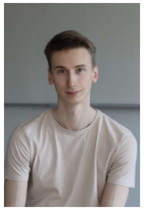
UNIVERSITÉ DE GENÈVE

Security of foundation models: implications for downstream tasks, content protection and tracking

Slava Voloshynovskiy

in collaboration with **Brian Pulfer**, **Yury Belousov** and **Vitaliy Kinakh**



stochastic information processing

2024, Rennes

Agenda

- Problem formulation
  - A need for content protection and tracking
  - Advancement of Foundation Models (FM)
  - Advancement of Digital Watermarking Systems
  - Advancement of Content Tracking Systems
- Security of Foundation Models
- Security of Digital Watermarking
- Variability of Security of Various Foundation Models
- Conclusions

Problem formulation: The Importance of Data Origin and Protection

Data Origin: the source of data

- **Physical Observations**: Data collected from real-world phenomena, including environmental readings, health statistics, and economic indicators.

Multimedia devices

Medical imaging (X-ray, MRI, CT, US)

Remote sensing (satellites, drones)

Data from sensors (weather, traffic, wearables)

Lab instruments (microscopes)

- **Generative AI/ML**: Data produced by artificial intelligence (AI) or machine learning (ML) models, such as synthetic images, text, or sounds, simulating real-world data.

OpenAI midjourney Meta MISTRAL AI_ ANTHROP\C Google AI

Problem formulation: The Importance of Data Origin and Protection

Data Use: applications and implications

- **By Humans**: utilized in various sectors including educational content, news dissemination, entertainment, research, and decision-making.
- **By Machines**: employed for training and enhancing new AI/ML models.

Problem formulation: The Importance of Data Origin and Protection

Data Use: applications and implications

- **By Humans**: utilized in various sectors including news dissemination, educational content, entertainment, research, and decision making.

  Associated Problems:

  - Disinformation: the spread of false information under the guise of being legitimate.
  - Deep Fakes: highly realistic and convincing digital manipulations of audio or video content, often used maliciously.

- **By Machines**: employed for training and enhancing new AI/ML models.

Problem formulation: The Importance of Data Origin and Protection

Data Use: applications and implications

- **By Humans**: utilized in various sectors including news dissemination, educational content, entertainment, research, and decision making.
  Associated Problems:
  - Disinformation: the spread of false information under the guise of being legitimate.
  - Deep Fakes: highly realistic and convincing digital manipulations of audio or video content, often used maliciously.

- **By Machines**: employed for training and enhancing new AI/ML models.
  Associated Problems:
  - Copyright violation: unauthorized use of copyrighted data.
  - Bias: training data with inherent biases can result in biased models.
  - Adversarial attacks:
    - Poisoning: deliberately manipulating training data to compromise the model's integrity.
    - Adversarial examples: adversarial inputs designed to cause failure.

Problem formulation: The Importance of Data Origin and Protection

Challenges

- **Data Provenance:** Ensuring integrity, authenticity, and security.
- **Concerns**: Trust in information, misinformation prevention, adversarial attack protection, legal evidence integrity, ethical standards.

Regulatory Perspective

- **EU AI Act:** Acknowledges risks of modern ML models and generated content.

Technical Perspective

- **Necessity:** Robust methods for content protection and tracking.

Agenda

- Problem formulation
  - A need of content protection
  - **Advancement of Foundation Models (FM)**
  - Advancement of Digital Watermarking Systems
  - Advancement of Content Tracking Systems
- Security of Foundation models
- Security of Digital Watermarking and Active Image Indexing
- Variability of Security of Various Foundation Models
- Conclusions

Advancement of foundation models

Definition of Foundation Models

Foundation Models: Large-scale machine learning models trained on diverse and extensive datasets collected from the internet.

- **Data Modalities**: Incorporates multiple types of data including images, text, audio, and video.
- **Data Provenance**: The origins of the training data are not clearly known.
- **Applications:** Representation learning and generative models.

Development of Foundation Models

- **Developers**: Primarily developed by major technology companies with substantial computational resources: Meta (DINO, MAE, VICreg, I-JEPA, Llama), OpenAI (CLIP, ChatGPT, DALL-E, Sora).
- **Parameters**: These models contain millions to billions of parameters, requiring significant computational power.
- **Transparency Issues**: Not all companies disclose the specifics of the **training data** and processes.

Advancement of foundation models

Definition of Foundation Models

Architectural Diversity and Training Techniques

- **Model Architectures**: Varies widely, including different network structures (CNN, ViT, Mamba) and learning paradigms
- **Training Techniques**: Utilizes both contrastive and non-contrastive learning methods
- **Augmentations**: Various image manipulations for better generalization
- **Masked image modeling (MIM)**: to force models to learn powerful representations

Advancement of foundation models

Main concept of represenation learning

Foundation Model Training

- **Given:** large amount of training data (both public and proprietary, mainly without labels) and significant compute resources
- **Develop an encoder/embedder** that can project high-dimensional data into an informative low-dimensional space (**self-supervised learning (SSL)**)
- **Utility/versality for tasks should** ensure the resulting embeddings are effective and applicable to a variety of downstream tasks

Main concept of represenation learning

Foundation Model Training

- **Given:** large amount of training data (both public and proprietary, mainly without labels) and significant compute resources
- **Develop an encoder/embedder** that can project high-dimensional data into an informative low-dimensional space (**self-supervised learning (SSL)**)
- **Utility/versality for tasks should** ensure the resulting embeddings are effective and applicable to a variety of downstream tasks

Model Utilization

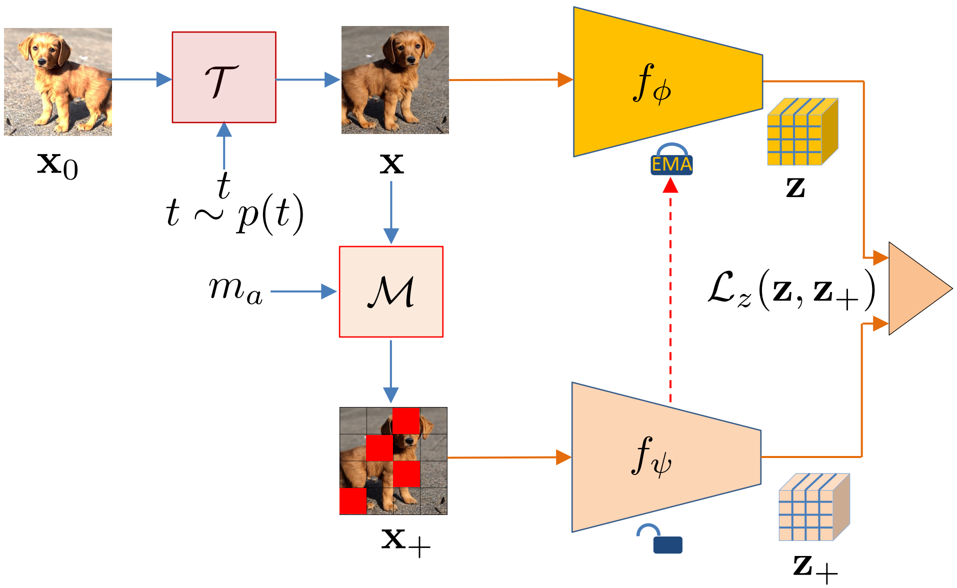
- **Base for enhancement**: Acts as a foundational platform for integrating specific neural network layers or projectors
- **Tailored fine-tuning**: Enables customization for particular applications using smaller, specialized datasets

Modes

- **Unimodal:** trained on one modality (ex: images)
- **Multimodal (CLIP)**: trained on several modalities (ex: images-text)

Main concept of represenation learning

1 Foundation Model Training

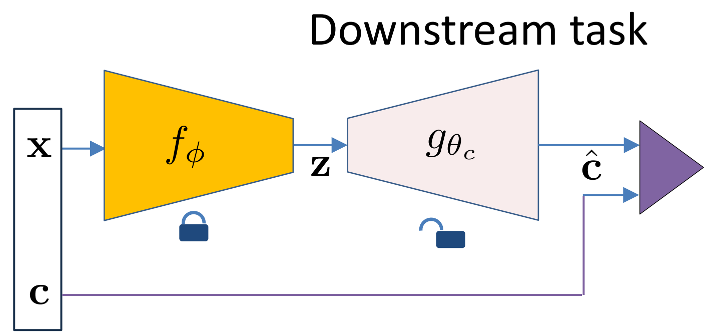


Example of joint embedding architecture

Training is done w/o labels

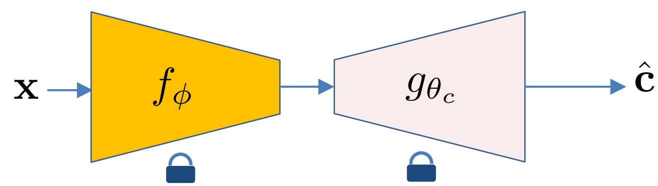
2 Model Utilization

Model fine-tuning



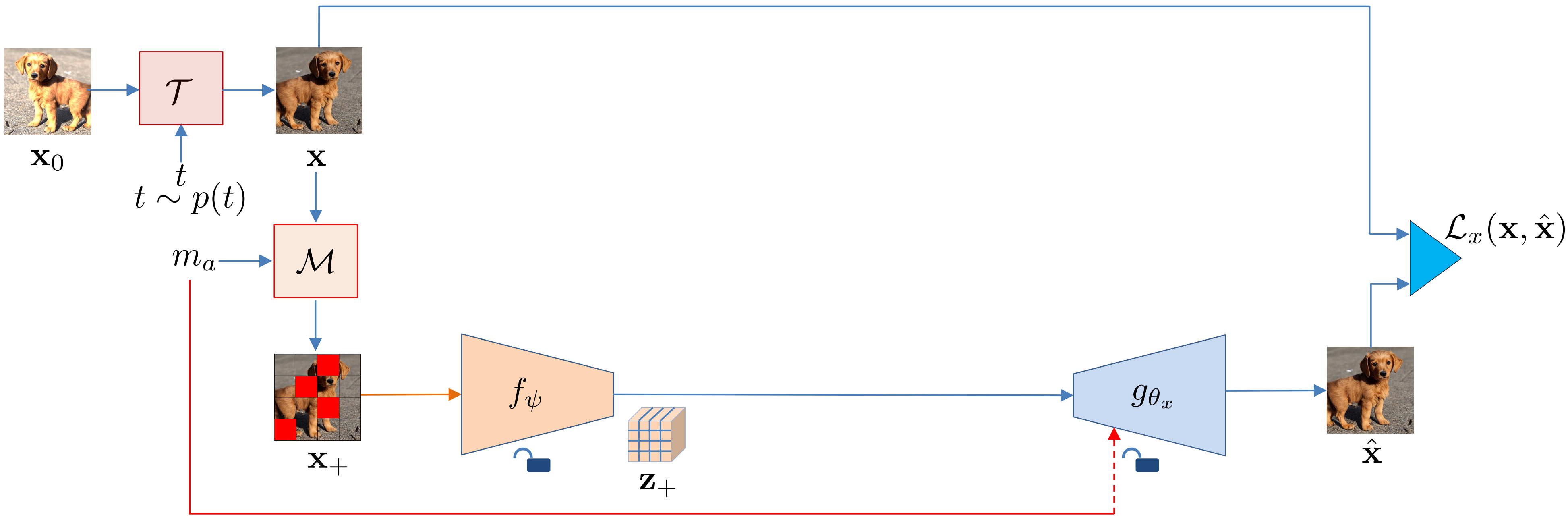
Fune-tuning is done with labeled data

3 Model deployment



Advancement of foundation models

Modern FM architectures



Embedding-reconstruction (AE):

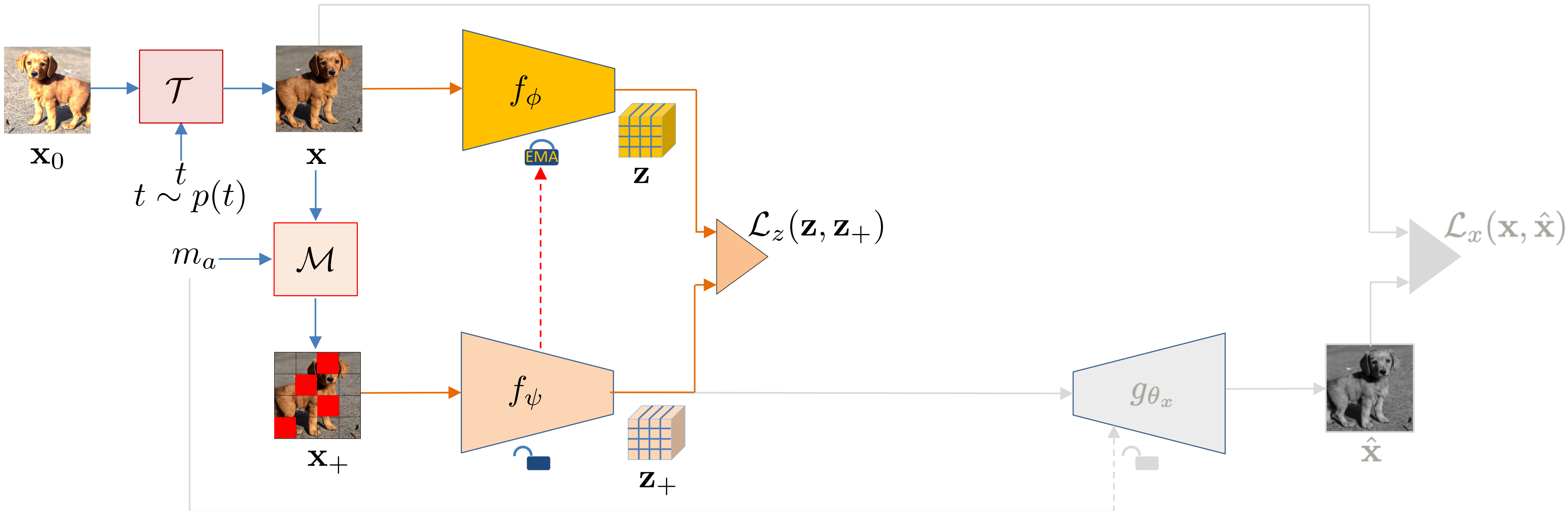
1. No mode collapse
2. High complexity
3. No good loss for pixel space

Denoising-AE, MAE

Training variations: scope (patches, aggregated patches, CLS); projectors; quantization

Advancement of foundation models

Modern FM architectures



Embedding-reconstruction (AE):

1. No mode collapse
2. High complexity
3. No good loss for pixel space

Denoising-AE, MAE

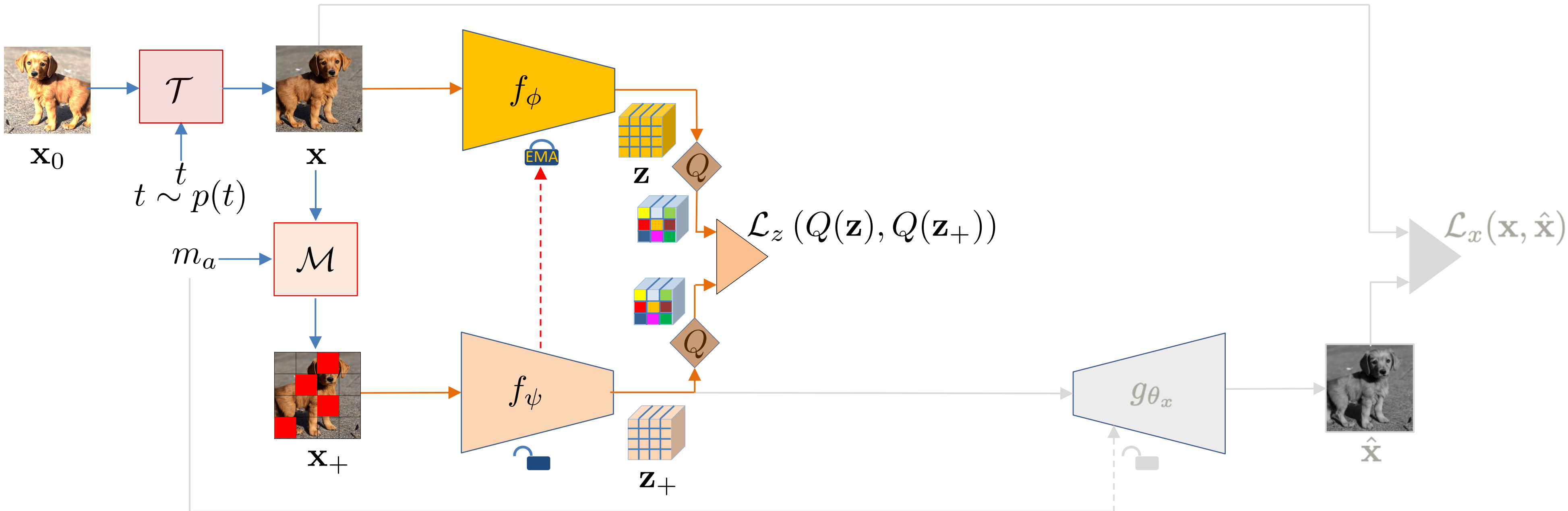
Joint embedding:

1. Possible mode collapse
2. Low complexity
3. Some potentially good losses for latent space

SimCLR, BYOL, Swav, MSN, DINO
VicReg, BarlowTwins

Training variations: scope (patches, aggregated patches, CLS); projectors; quantization

Modern FM architectures



Embedding-reconstruction (AE):
1. No mode collapse
2. High complexity
3. No good loss for pixel space

Denoising-AE, MAE

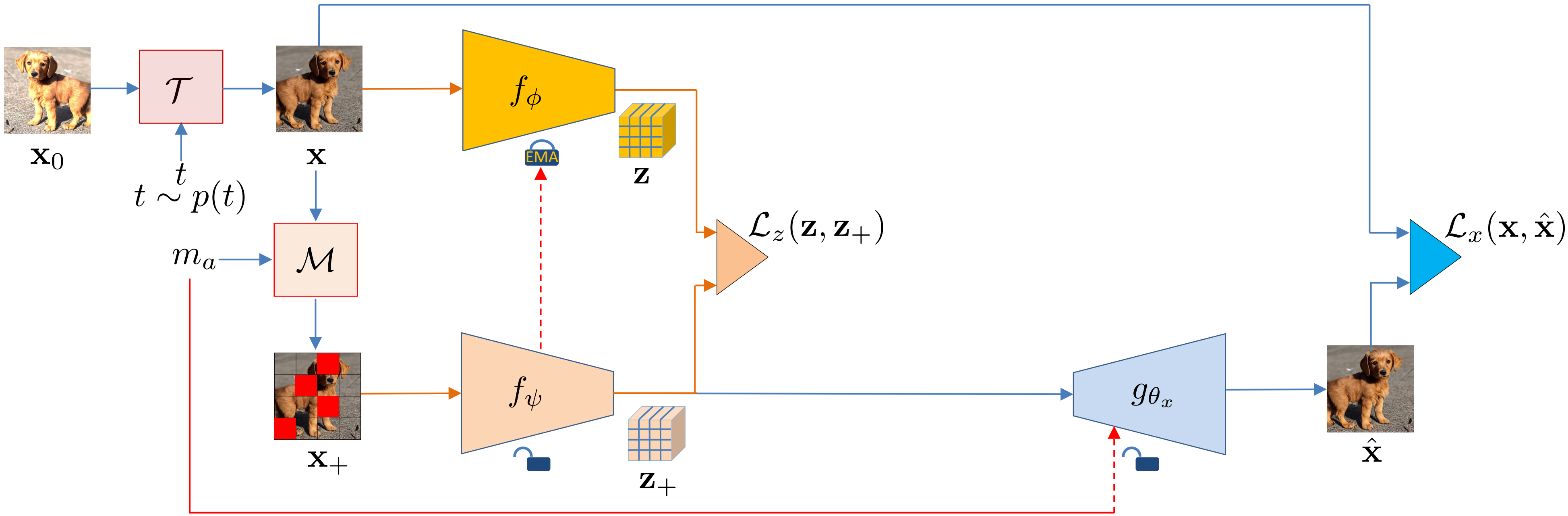
Joint embedding:
1. Possible mode collapse
2. Low complexity
3. Some potentially good losses for latent space

SimCLR, BYOL, Swav, MSN, DINO
VicReg, BarlowTwins

Training variations: scope (patches, aggregated patches, CLS); projectors; quantization

Advancement of foundation models

Modern FM architectures



Hybrid versions: CAN, CAE, CMAE, BeIT

Embedding-reconstruction (AE):

1. No mode collapse
2. High complexity
3. No good loss for pixel space

Denoising-AE, MAE

Joint embedding:

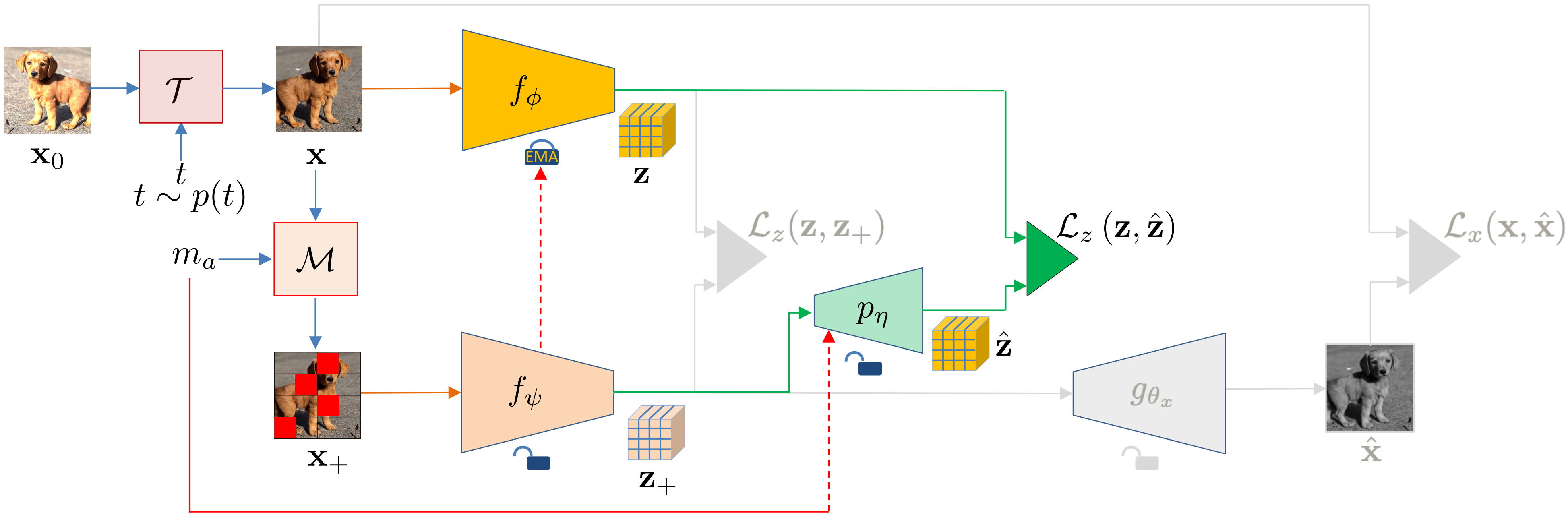
1. Possible mode collapse
2. Low complexity
3. Some potentially good losses for latent space

SimCLR, BYOL, Swav, MSN, DINO
VicReg, BarlowTwins

Training variations: scope (patches, aggregated patches, CLS); projectors; quantization

Advancement of foundation models

Modern FM architectures



Embedding-reconstruction (AE):
1. No mode collapse
2. High complexity
3. No good loss for pixel space

Denoising-AE, MAE

Joint embedding:
1. Possible mode collapse
2. Low complexity
3. Some potentially good losses for latent space

SimCLR, BYOL, Swav, MSN, DINO
VicReg, BarlowTwins

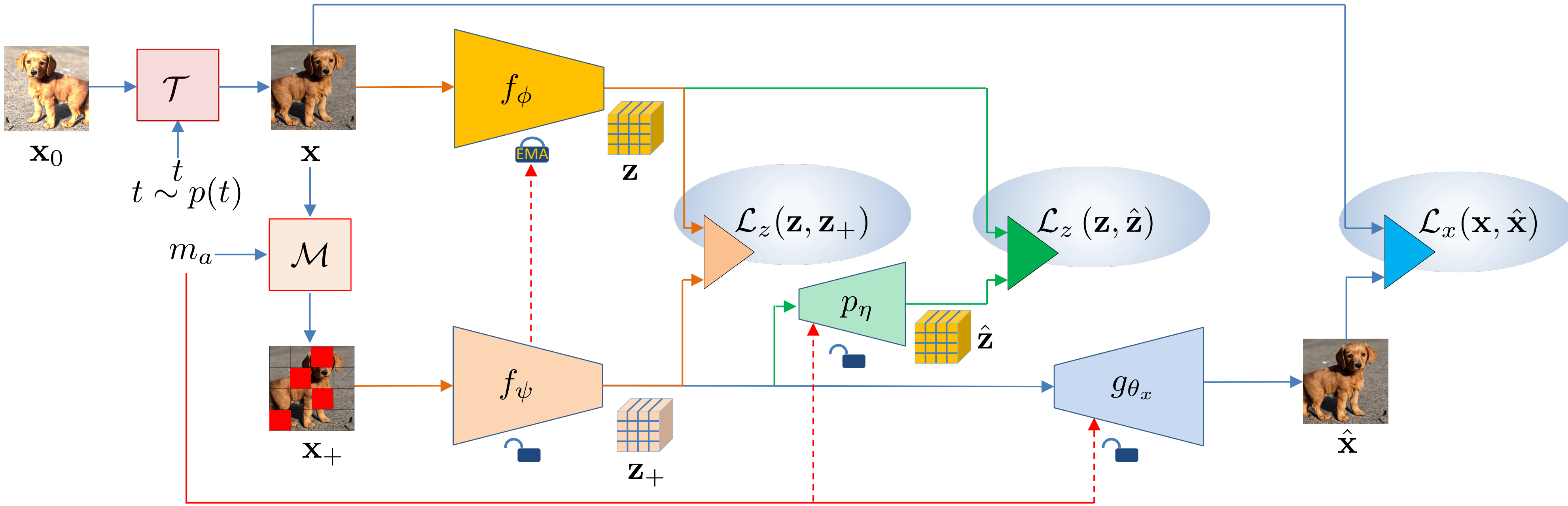
Joint embedding-prediction:
1. Possible mode collapse
2. Relatively low complexity
3. Some potentially good losses for latent space

I-JEPA, World Model

Training variations: scope (patches, aggregated patches, CLS); projectors; quantization

Advancement of foundation models

Modern FM architectures



Embedding-reconstruction (AE):

1. No mode collapse
2. High complexity
3. No good loss for pixel space

Denoising-AE, MAE

Joint embedding:

1. Possible mode collapse
2. Low complexity
3. Some potentially good losses for latent space

SimCLR, BYOL, Swav, MSN, DINO
VicReg, BarlowTwins

Joint embedding-prediction:

1. Possible mode collapse
2. Relatively low complexity
3. Some potentially good losses for latent space

I-JEPA, World Model

Training variations: scope (patches, aggregated patches, CLS); projectors; quantization

FM Losses

$\mathcal{L}_z(\mathbf{z}, \mathbf{z}_+)$

$\mathcal{L}_z(\mathbf{z}, \hat{\mathbf{z}})$

$\mathcal{L}_x(\mathbf{x}, \hat{\mathbf{x}})$

$$\max_{\psi} I_{\phi,\psi}(\mathbf{Z}; \mathbf{Z}_+) = H_\phi(\mathbf{Z}) - H_{\phi,\psi}(\mathbf{Z}|\mathbf{Z}_+)$$

FM Losses

$\mathcal{L}_z(\mathbf{z}, \mathbf{z}_+)$

$\mathcal{L}_z(\mathbf{z}, \hat{\mathbf{z}})$

$\mathcal{L}_x(\mathbf{x}, \hat{\mathbf{x}})$

$$\max_{\psi} I_{\phi,\psi}(\mathbf{Z}; \mathbf{Z}_+) = H_\phi(\mathbf{Z}) - H_{\phi,\psi}(\mathbf{Z}|\mathbf{Z}_+)$$

$$p_{\phi,\psi}(\mathbf{z}|\mathbf{z}_+) = \frac{1}{C} e^{\beta<\mathbf{z},\mathbf{z}_+>}$$

$$H_{\phi,\psi}(\mathbf{Z} \mid \mathbf{Z}_+) = -\mathbb{E}_{p_{\phi,\psi}(\mathbf{z},\mathbf{z}_+)}\left[\log p_{\phi,\psi}(\mathbf{z}|\mathbf{z}_+)\right]$$

$$= -\mathbb{E}_{p_{\phi,\psi}(\mathbf{z},\mathbf{z}_+)}\left[\log \frac{1}{C} e^{\beta<\mathbf{z},\mathbf{z}_+>\}}\right]$$

Advancement of foundation models

FM Losses

$\mathcal{L}_z(\mathbf{z}, \mathbf{z}_+)$

$\mathcal{L}_z(\mathbf{z}, \hat{\mathbf{z}})$

$\mathcal{L}_x(\mathbf{x}, \hat{\mathbf{x}})$

$$\max_{\psi} I_{\phi,\psi}(\mathbf{Z}; \mathbf{Z}_+) = \boxed{H_\phi(\mathbf{Z})} - H_{\phi,\psi}(\mathbf{Z}|\mathbf{Z}_+)$$

Contrastive

$$p_\phi(\mathbf{z}) = \mathbb{E}_{p_\psi(\mathbf{z}_+)}\left[p_{\psi,\phi}(\mathbf{z}|\mathbf{z}_+)\right]$$

$$H_\phi(\mathbf{Z}) = -\mathbb{E}_{p_\phi(\mathbf{z})}\left[\log p_\phi(\mathbf{z})\right]$$

$$= -\mathbb{E}_{p_{\phi,\psi}(\mathbf{z},\mathbf{z}_+)}\left[\log \mathbb{E}_{p_\psi(\mathbf{z}'_+)}\left[p_{\psi,\phi}(\mathbf{z}|\mathbf{z}'_+)\right]\right]$$

$$= -\mathbb{E}_{p_{\phi,\psi}(\mathbf{z},\mathbf{z}_+)}\left[\log \mathbb{E}_{p_\psi(\mathbf{z}'_+)}\left[\frac{1}{C} e^{\beta<\mathbf{z},\mathbf{z}'_+>\}}\right]\right]$$

Non-Contrastive Adversarial

$$H_\phi(\mathbf{Z}) = -\mathbb{E}_{p_\phi(\mathbf{z})}\left[\log p_\phi(\mathbf{z}) \frac{p_\psi(\mathbf{z}_+)}{p_\psi(\mathbf{z}_+)}\right]$$

$$= -\mathbb{D}_{KL}\left(p_\phi(\mathbf{z}) \| p_\psi(\mathbf{z}_+)\right) + H\left(p_\phi(\mathbf{z}); p_\psi(\mathbf{z}_+)\right)$$

$$p_{\phi,\psi}(\mathbf{z}|\mathbf{z}_+) = \frac{1}{C} e^{\beta<\mathbf{z},\mathbf{z}_+>}$$

$$H_{\phi,\psi}(\mathbf{Z} \mid \mathbf{Z}_+) = -\mathbb{E}_{p_{\phi,\psi}(\mathbf{z},\mathbf{z}_+)}\left[\log p_{\phi,\psi}(\mathbf{z}|\mathbf{z}_+)\right]$$

$$= -\mathbb{E}_{p_{\phi,\psi}(\mathbf{z},\mathbf{z}_+)}\left[\log \frac{1}{C} e^{\beta<\mathbf{z},\mathbf{z}_+>\}}\right]$$

InfoNCE

$$\mathbb{E}_{p_\psi(\mathbf{z}'_+)} \to \frac{1}{K}\sum_{k=1}^{K}$$

Regularization /VicREG/

$$\max_{\phi,\psi,} I_{\phi,\psi}(\mathbf{X}; \mathbf{Z}_+) + I_{\phi,\psi}(\mathbf{X}_+; \mathbf{Z})$$

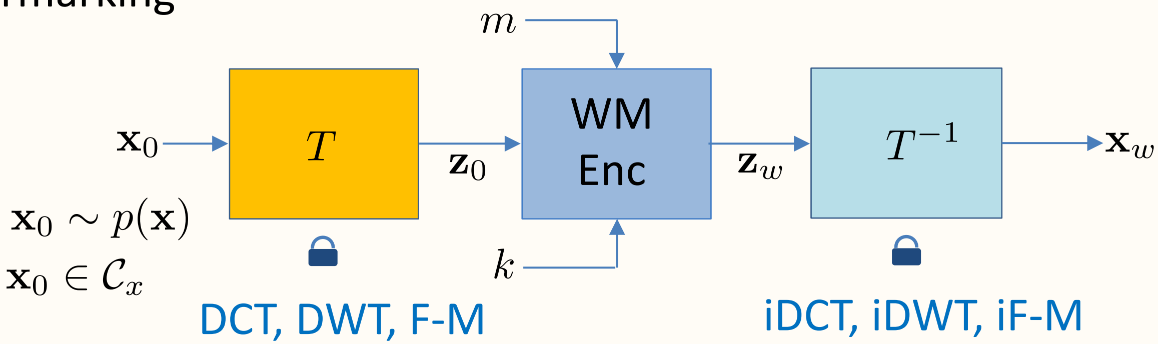
Agenda

- Problem formulation
  - A need of content protection
  - Advancement of Foundation Models (FM)
  - **Advancement of Digital Watermarking Systems**
  - Advancement of Content Tracking Systems
- Security of Foundation models
- Security of Digital Watermarking and Active Image Indexing
- Variability of Security of Various Foundation Models
- Conclusions

Advancement of digital watermarking systems

$\mathcal{DW}_1$ Hand-crafted architectures

Watermarking



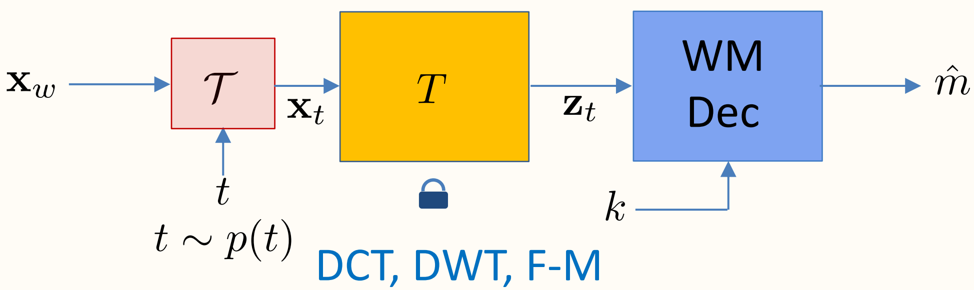
WM Enc:

- additive
- multiplicative
- quantization

Message m :

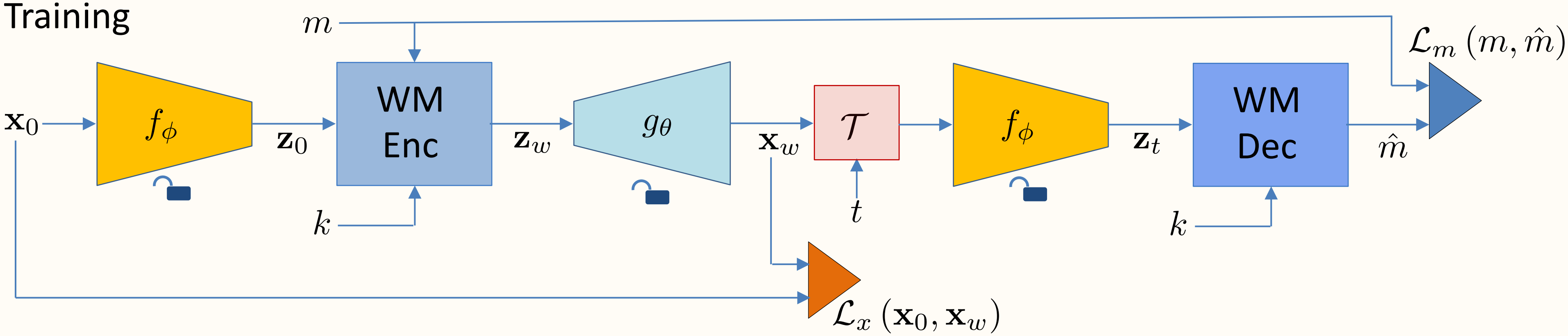
- zero-bit
- multi-bit

Testing/deployment

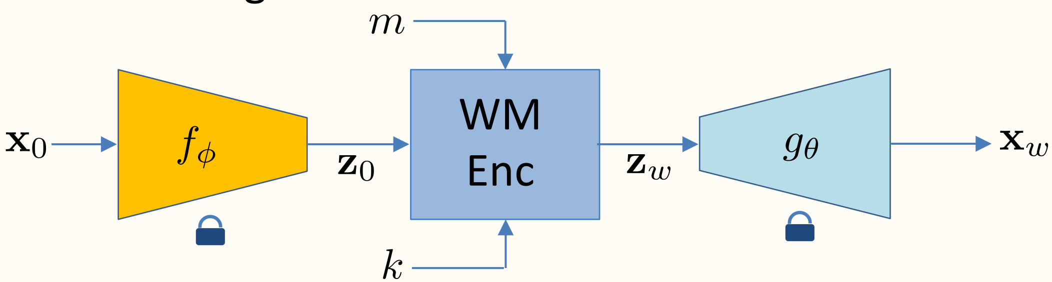


$\mathcal{DW}_2$ AE-based architectures

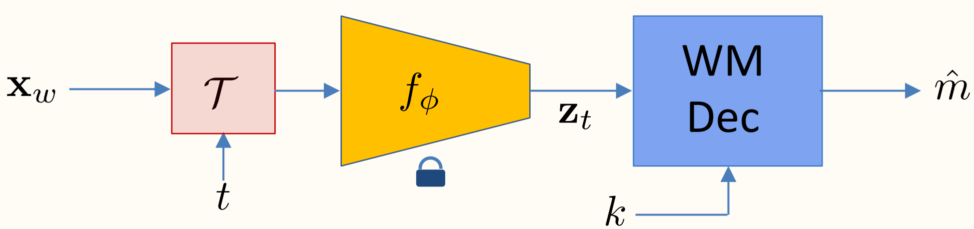
Training



Watermarking



Testing/deployment

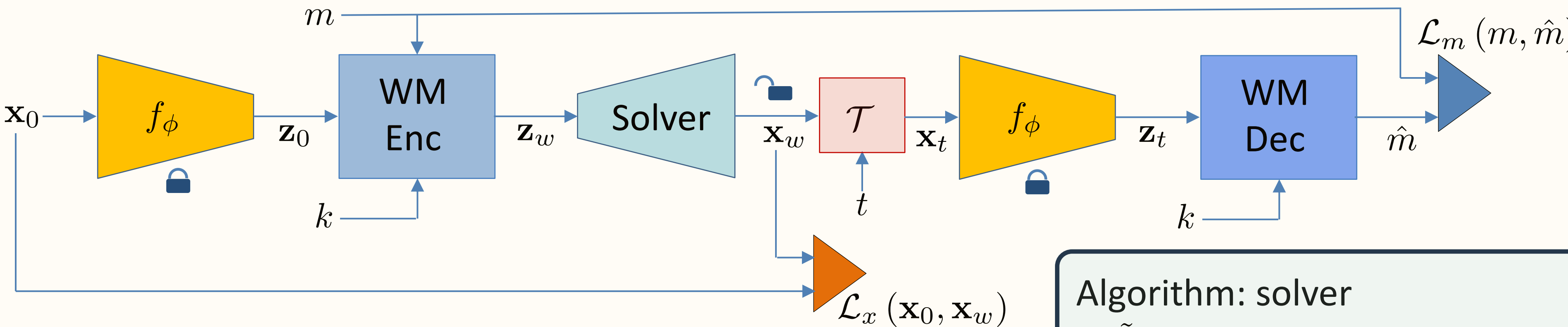


https://arxiv.org/pdf/1807.09937

Advancement of digital watermarking systems

$\mathcal{DW}_3$ Adversarial embedding architectures based on foundation models

Watermarking



$$\mathcal{L}_\mathcal{E}\left(\mathbf{x}_0, \mathbf{x}_a\right) = \mathcal{L}_x\left(\mathbf{x}_0, \mathbf{x}_w\right) + \lambda \mathcal{L}_m(m, \hat{m})$$

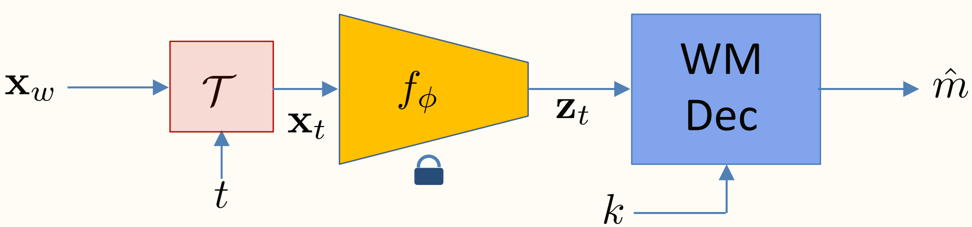
Algorithm: solver

$\tilde{\mathbf{x}} = \mathbf{x}_0$

$\tilde{\mathbf{x}} \leftarrow \tilde{\mathbf{x}} + \beta\, \mathrm{Optimizer}\left(\mathcal{L}_\mathcal{E}\right)$

$\mathbf{x}_w \leftarrow \ \mathrm{Constraints}\ \left(\mathbf{x}_0, \tilde{\mathbf{x}}\right)$

Testing/deployment



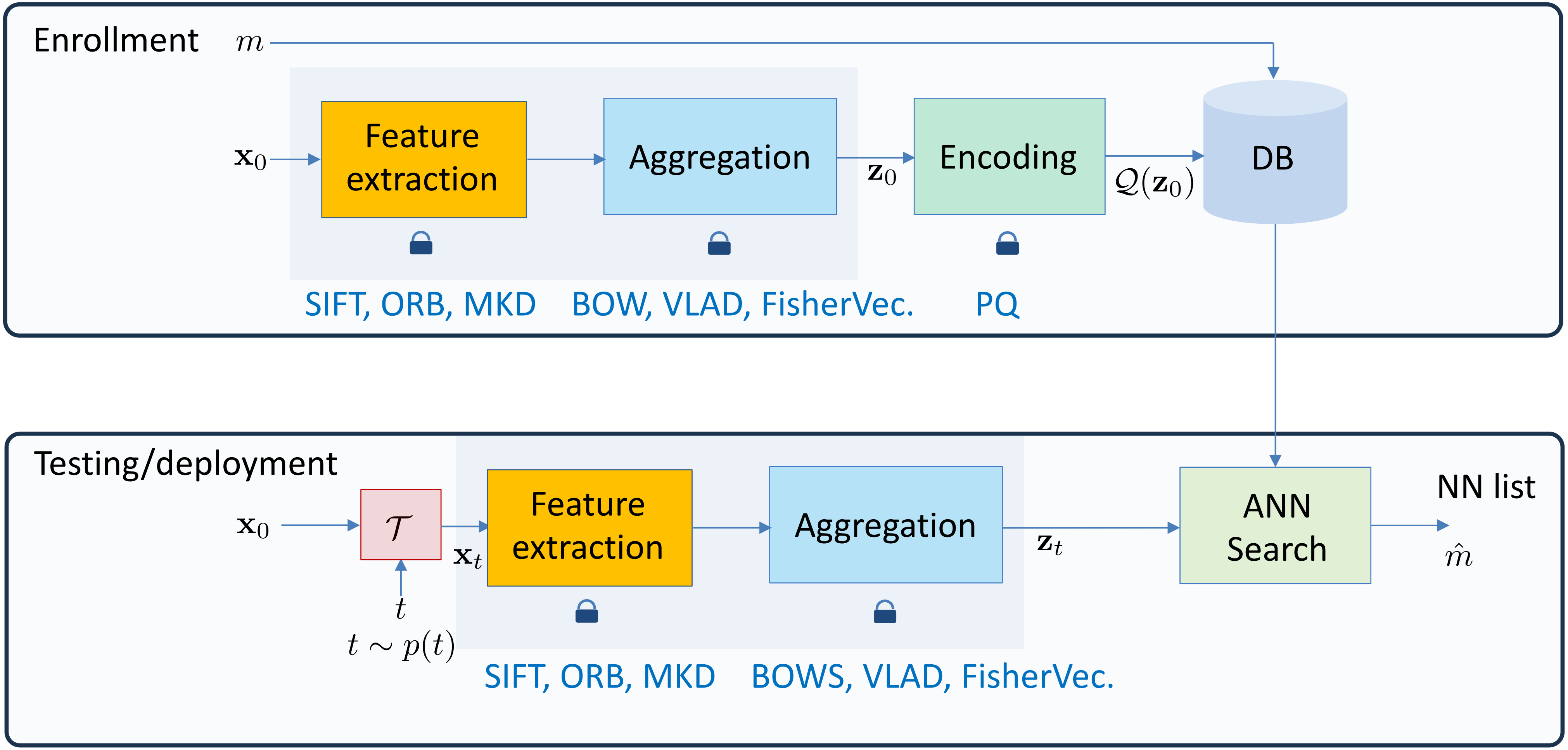
https://arxiv.org/pdf/2112.09581

Agenda

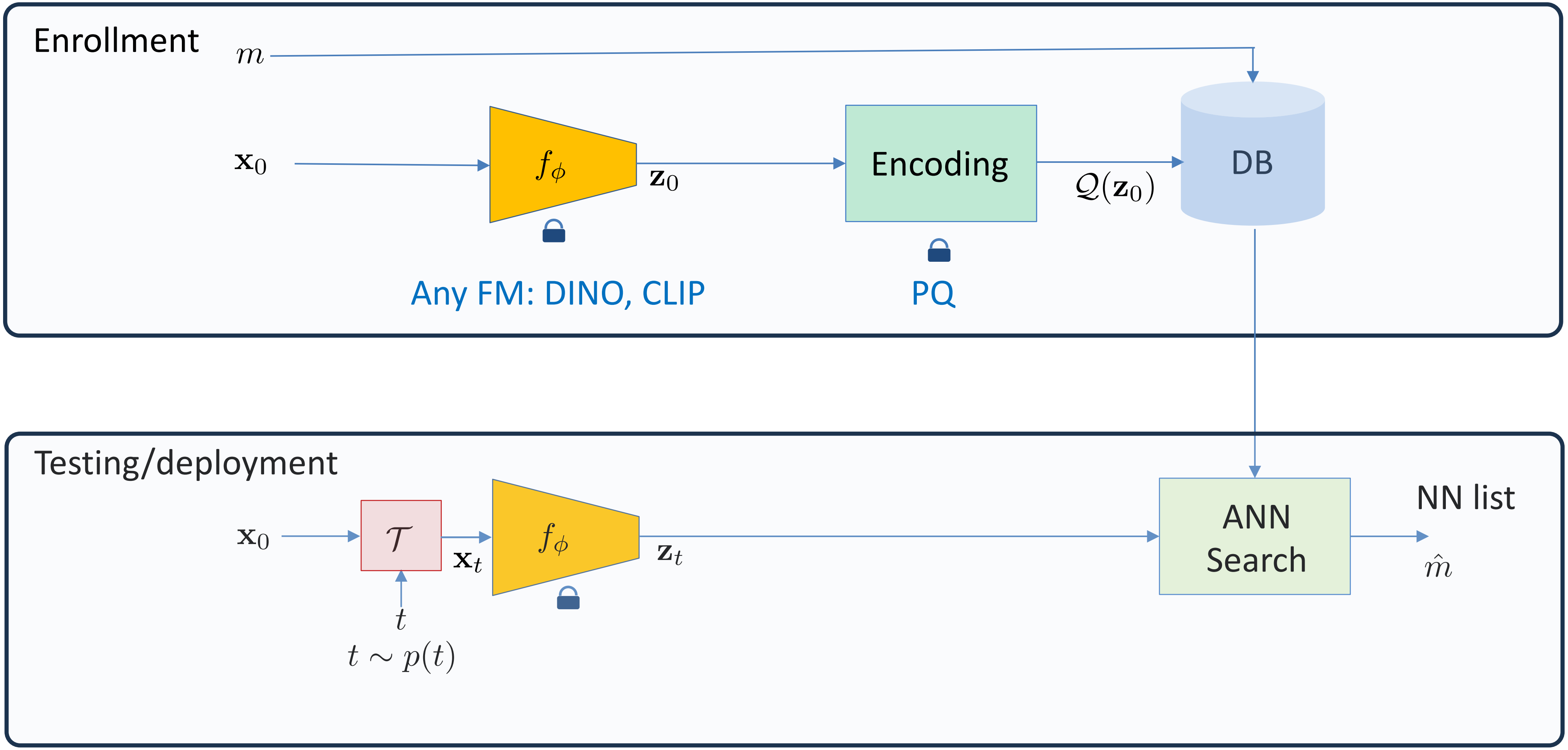
- Problem formulation
  - A need of content protection
  - Advancement of Foundation Models (FM)
  - Advancement of Digital Watermarking Systems
  - **Advancement of Content Tracking Systems**
- Security of Foundation models
- Security of Digital Watermarking and Active Image Indexing
- Variability of Security of Various Foundation Models
- Conclusions

Advancement of content tracking systems

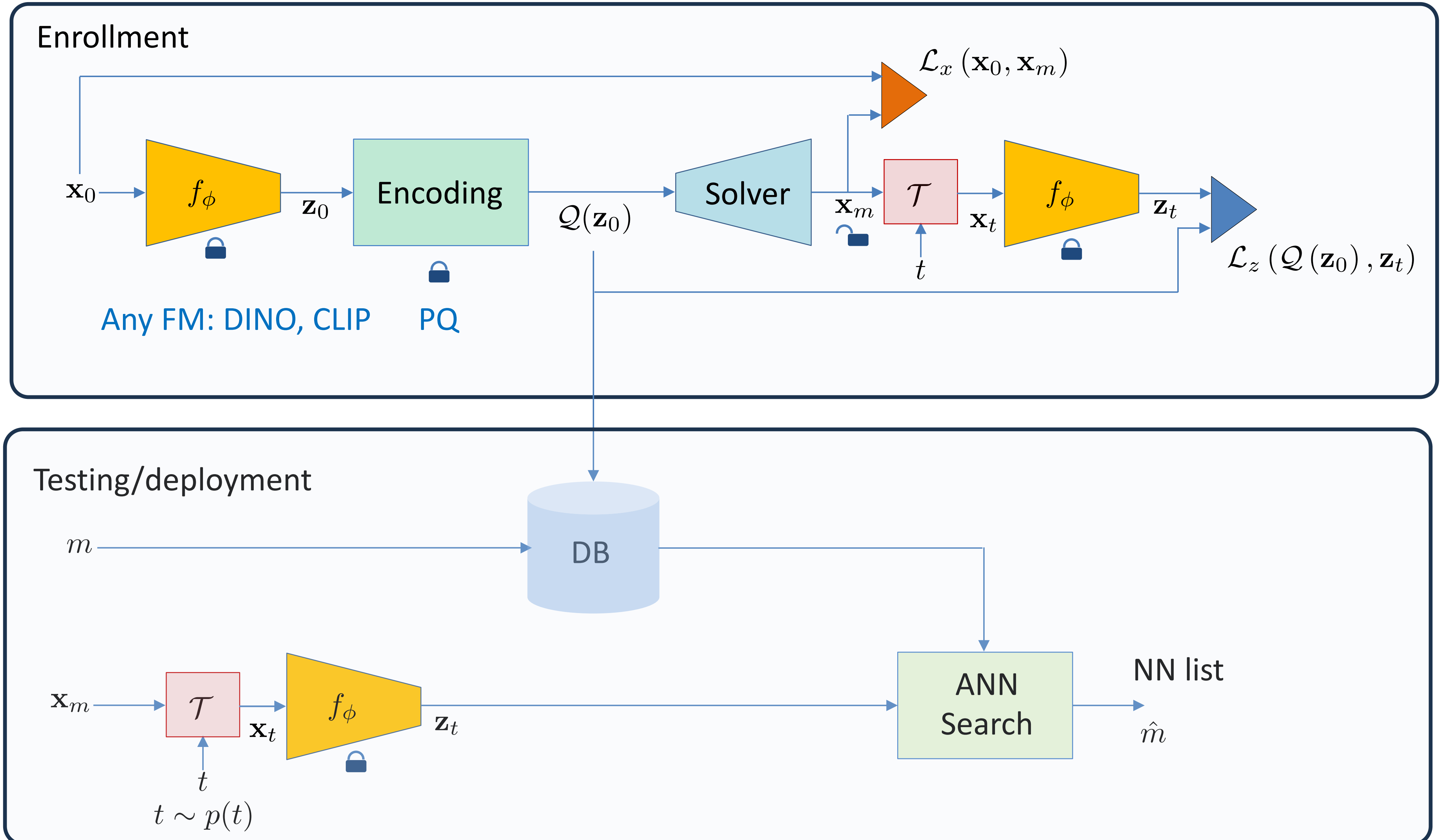
$\mathcal{CT}_1$ Hand-crafted architectures



$\mathcal{CT}_2$ Foundation model-based architectures



$\mathcal{CT}_3$ Active image indexing/active fingerprinting



https://arxiv.org/pdf/2210.10620

Concluding remark

Foundation models

- ML/AI Downstream Tasks
  - Classification
  - Segmentation
  - Retrieval
  - Object detection
  - Conditioning for GenAI
  - …
- Content protection
  - Backbone models
    - Embedder
- Content tracking
  - Backbone models
    - Feature extraction

What about the security of the foundation model?

Agenda

- Problem formulation
  - A need of content protection
  - Advancement of Foundation Models (FM)
  - Advancement of Digital Watermarking Systems
  - Advancement of Content Tracking Systems
- **Security of Foundation models**
- Security of Digital Watermarking Systems
- Variability of Security of Various Foundation Models
- Conclusions

Security of Foundation Models (task agnostic)

The rest of the slides will come soon...

Thank you!