The Role of Entropy and Reconstruction for Multi-View Self-Supervised Learning

Borja Rodríguez-Gálvez, Arno Blaas, Pau Rodríguez, Adam Goliński, Xavier Suau, Jason Ramapuram, Dan Busbridge, Luca Zappella

ICML 2023

Motivation: MVSSL progressing rapidly

… and many more

Motivation: But poorly understood theoretically

🔎 Lens of Information Theory

✅ Some contrastive MVSSL methods optimize InfoNCE, a lower bound on the Mutual Information (MI)

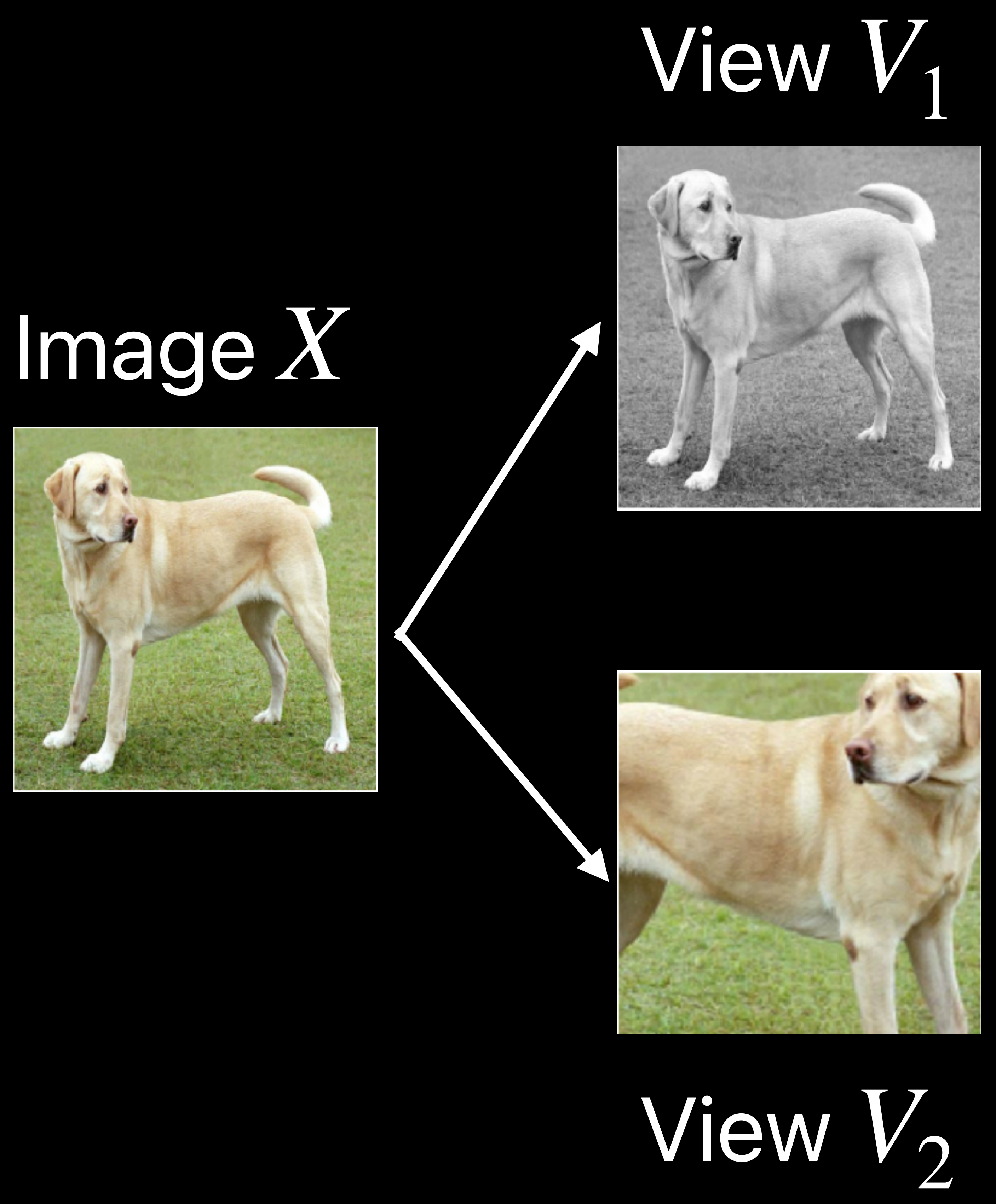
❓ What about the other MVSSL methods?

Background: Multi-view self-supervised learning (MVSSL)

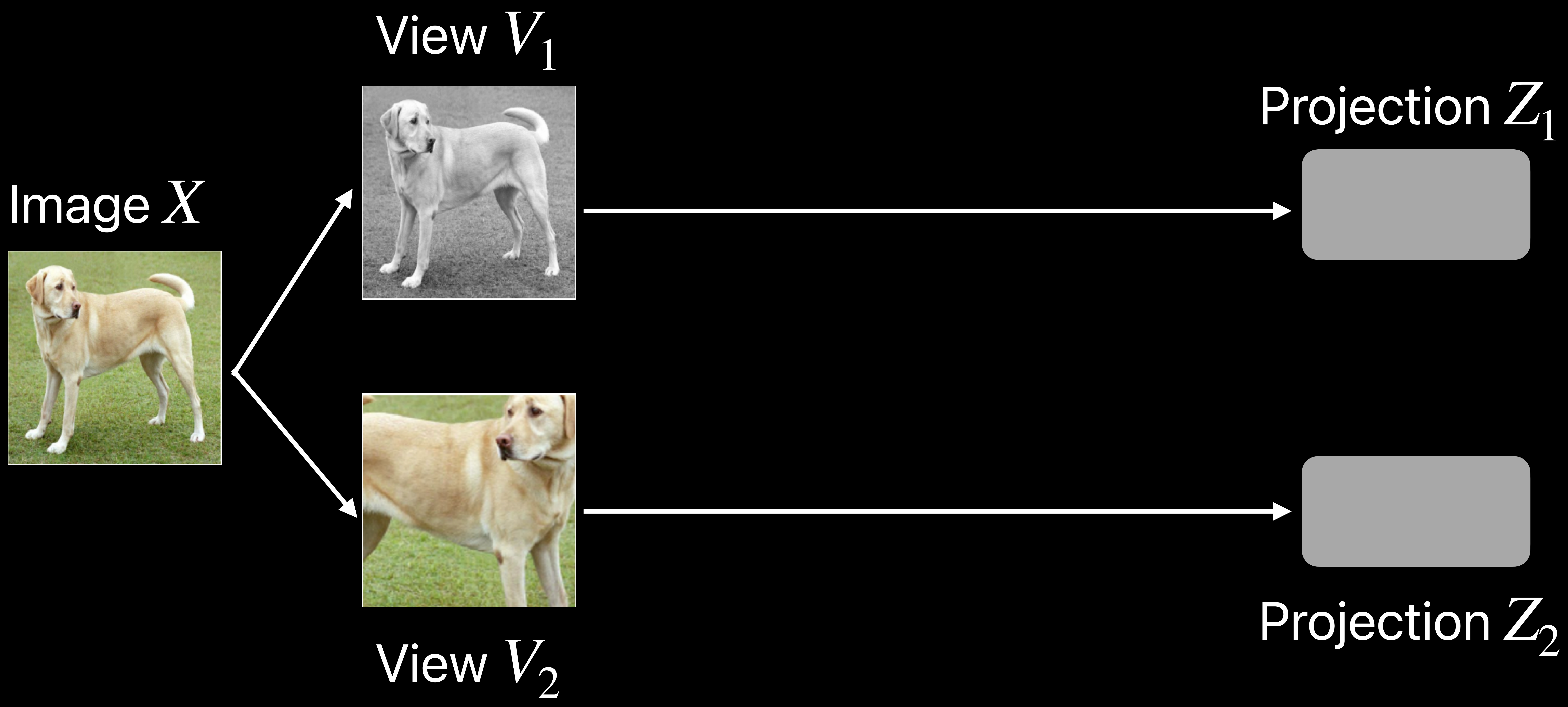
Background: Multi-view self-supervised learning (MVSSL)

Image X

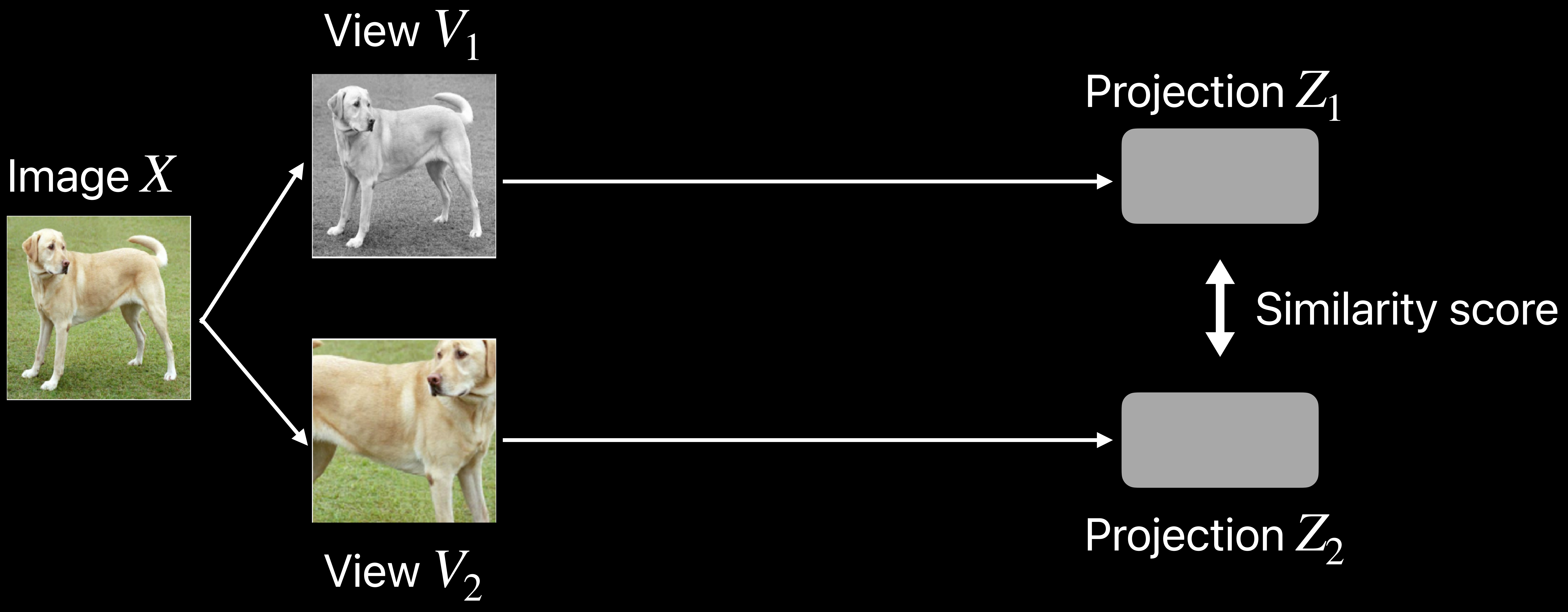
Background: Multi-view self-supervised learning (MVSSL)



Background: Multi-view self-supervised learning (MVSSL)

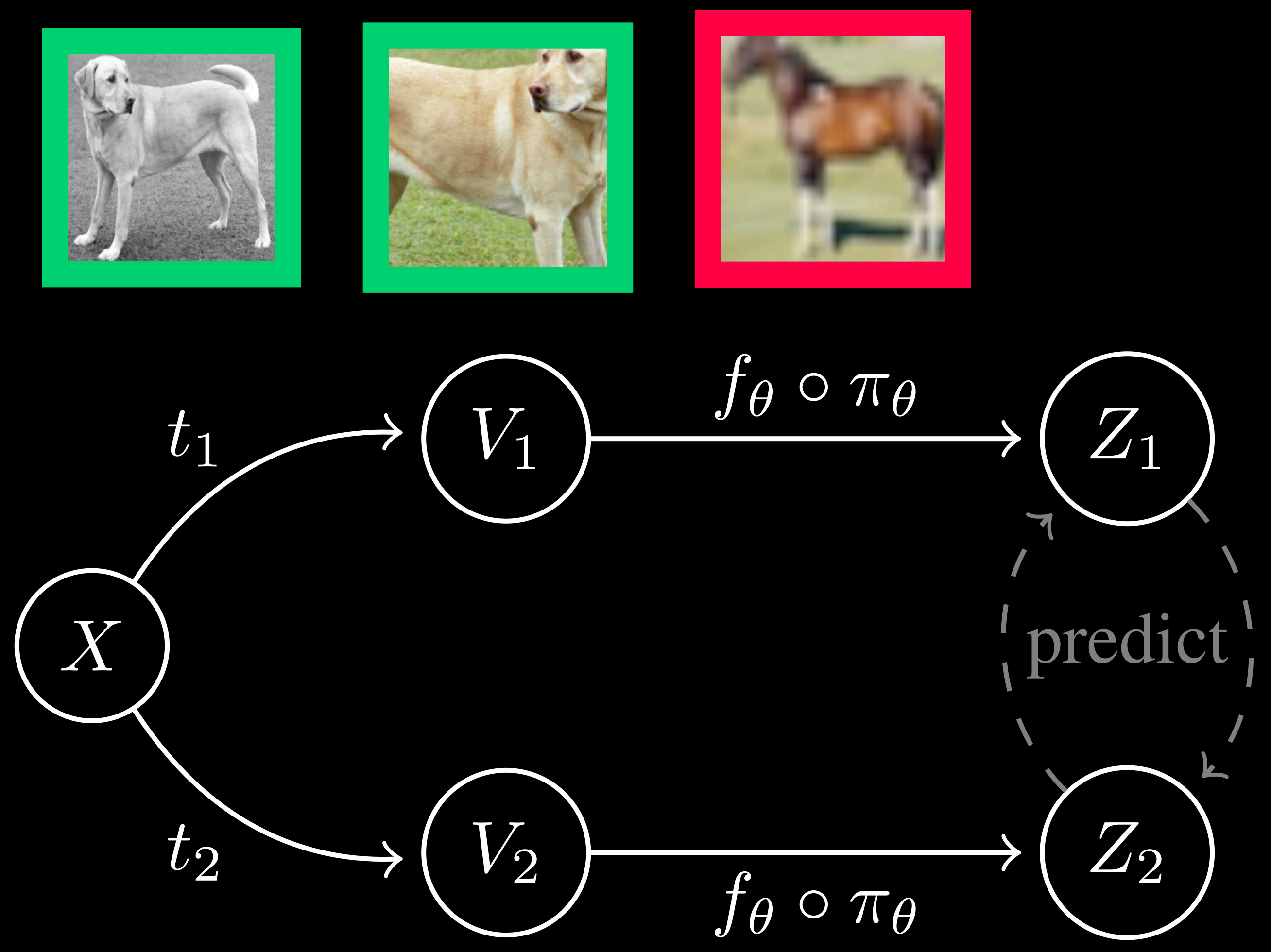


Background: Multi-view self-supervised learning (MVSSL)



Background: MVSSL families

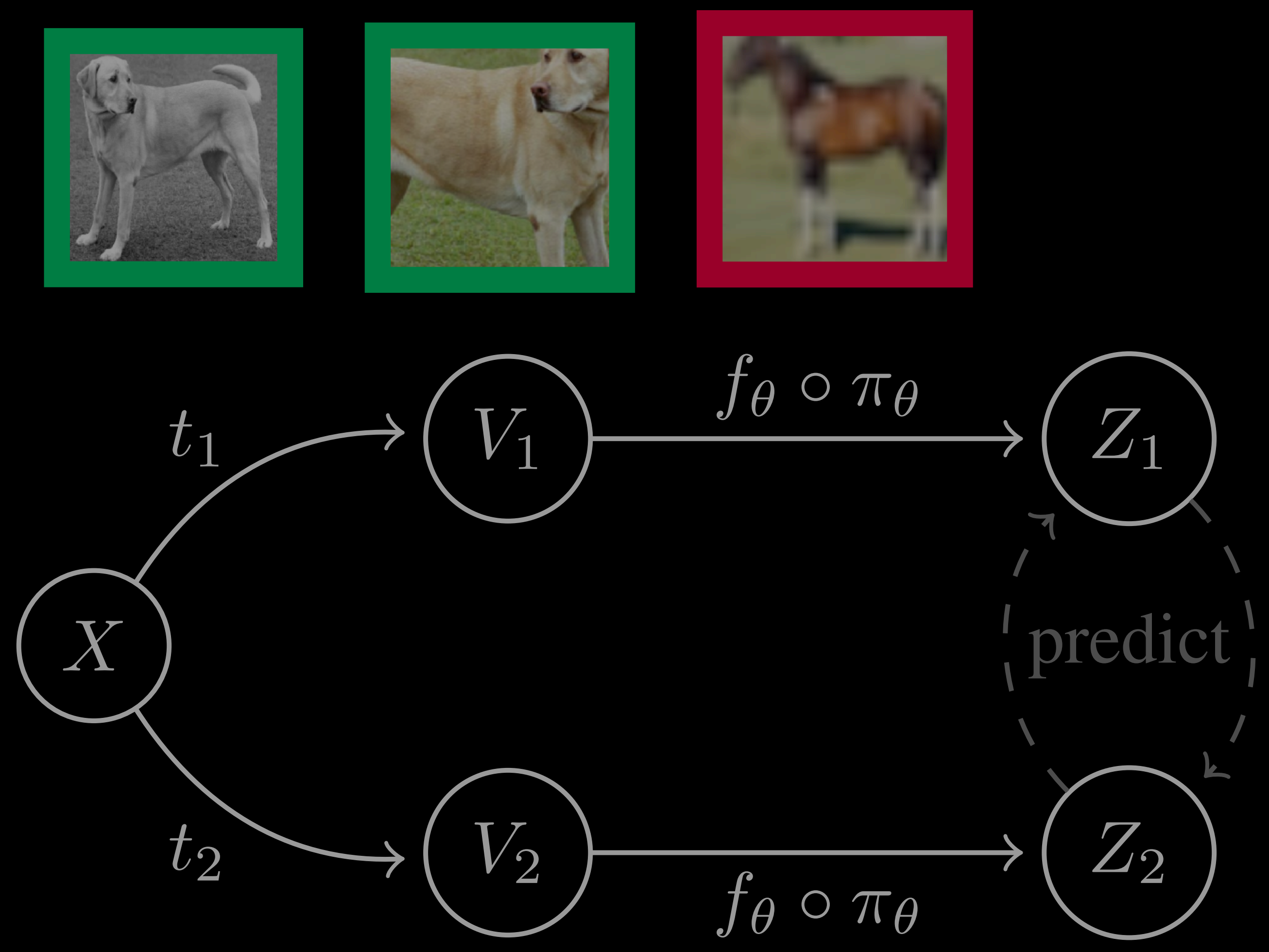
Background: MVSSL families



Contrastive methods

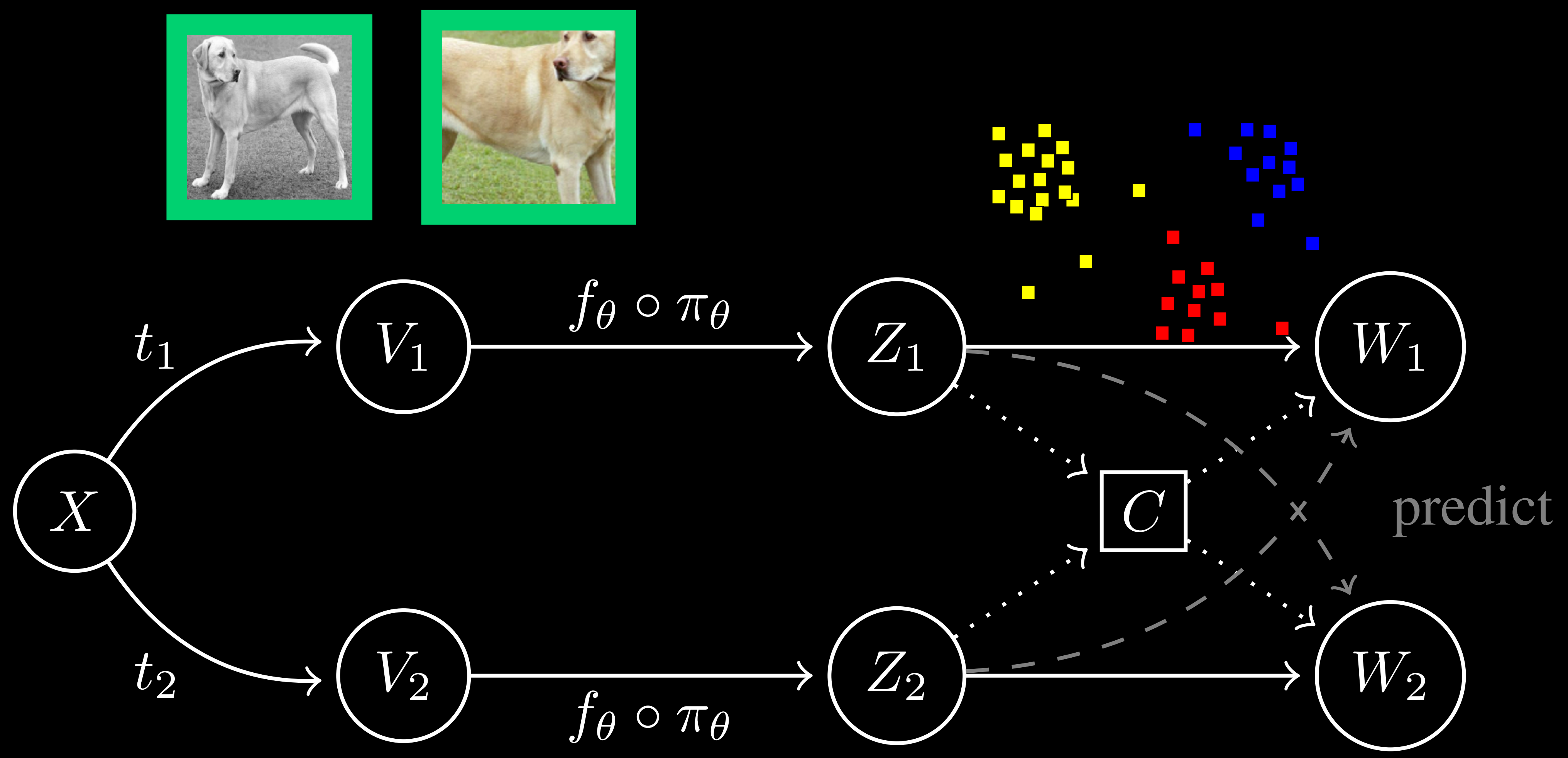
SimCLR, CMC, MoCo

Background: MVSSL families



Contrastive methods

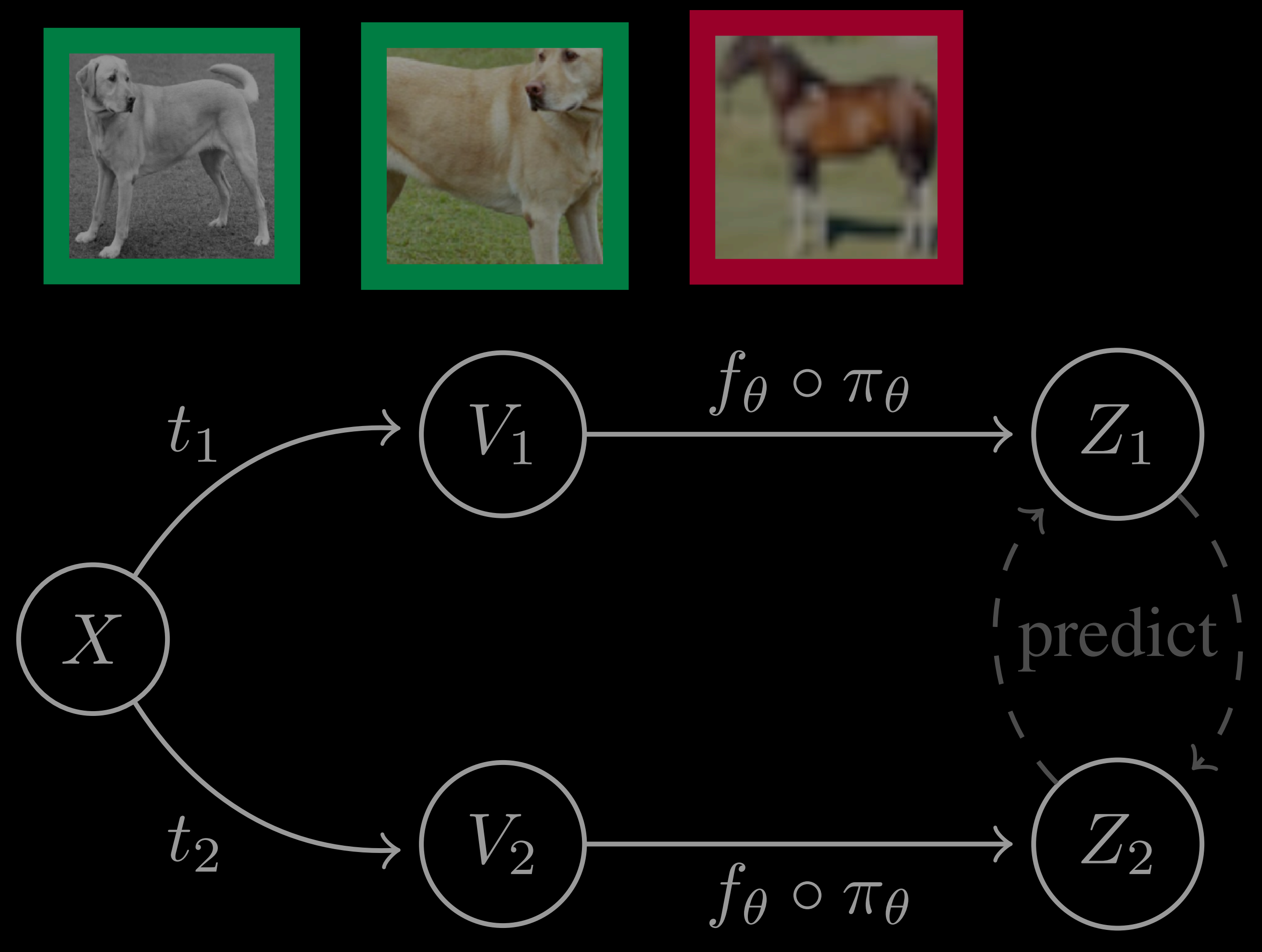
SimCLR, CMC, MoCo



Clustering-based methods

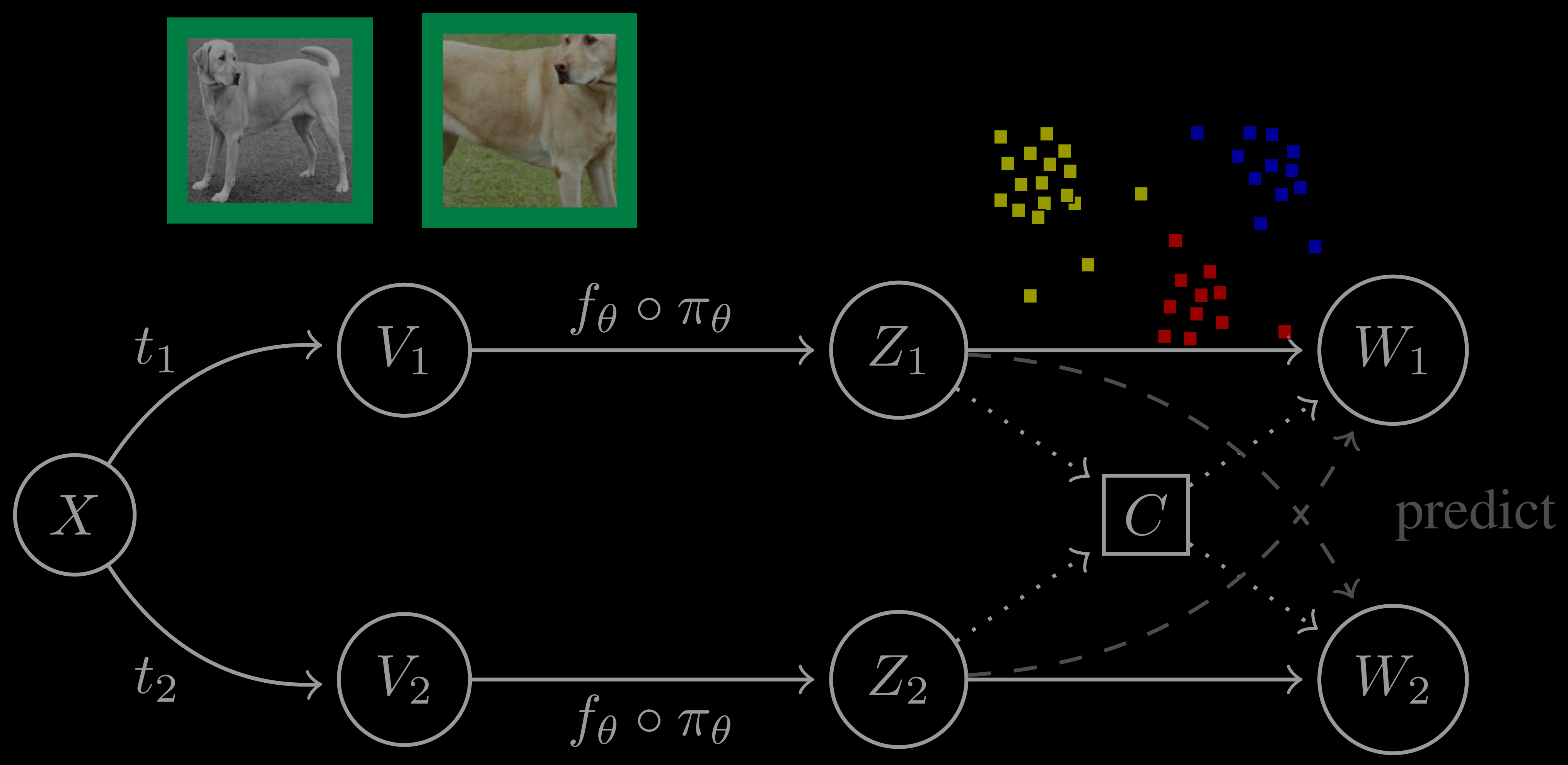
SwAV, DeepCluster

Background: MVSSL families



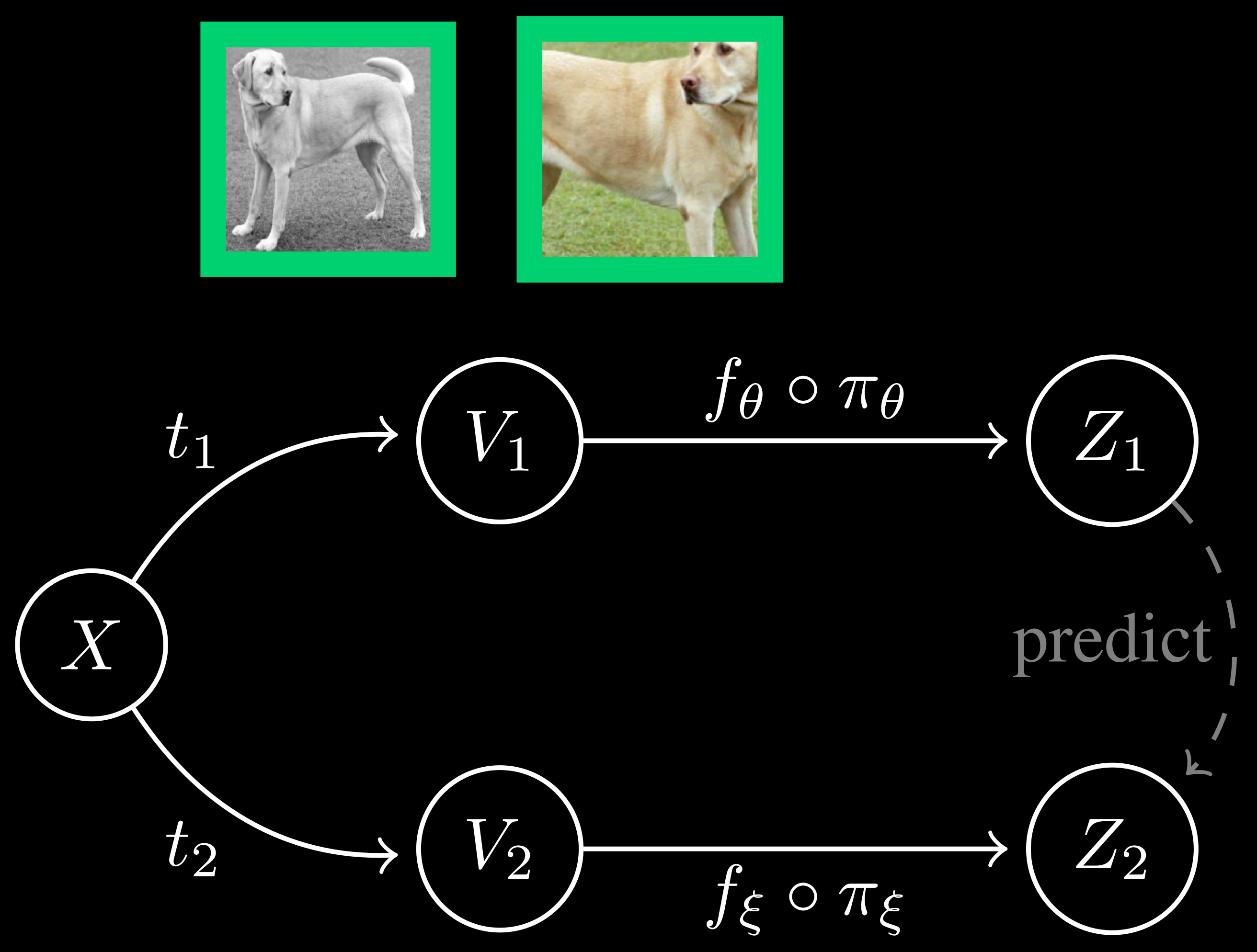
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

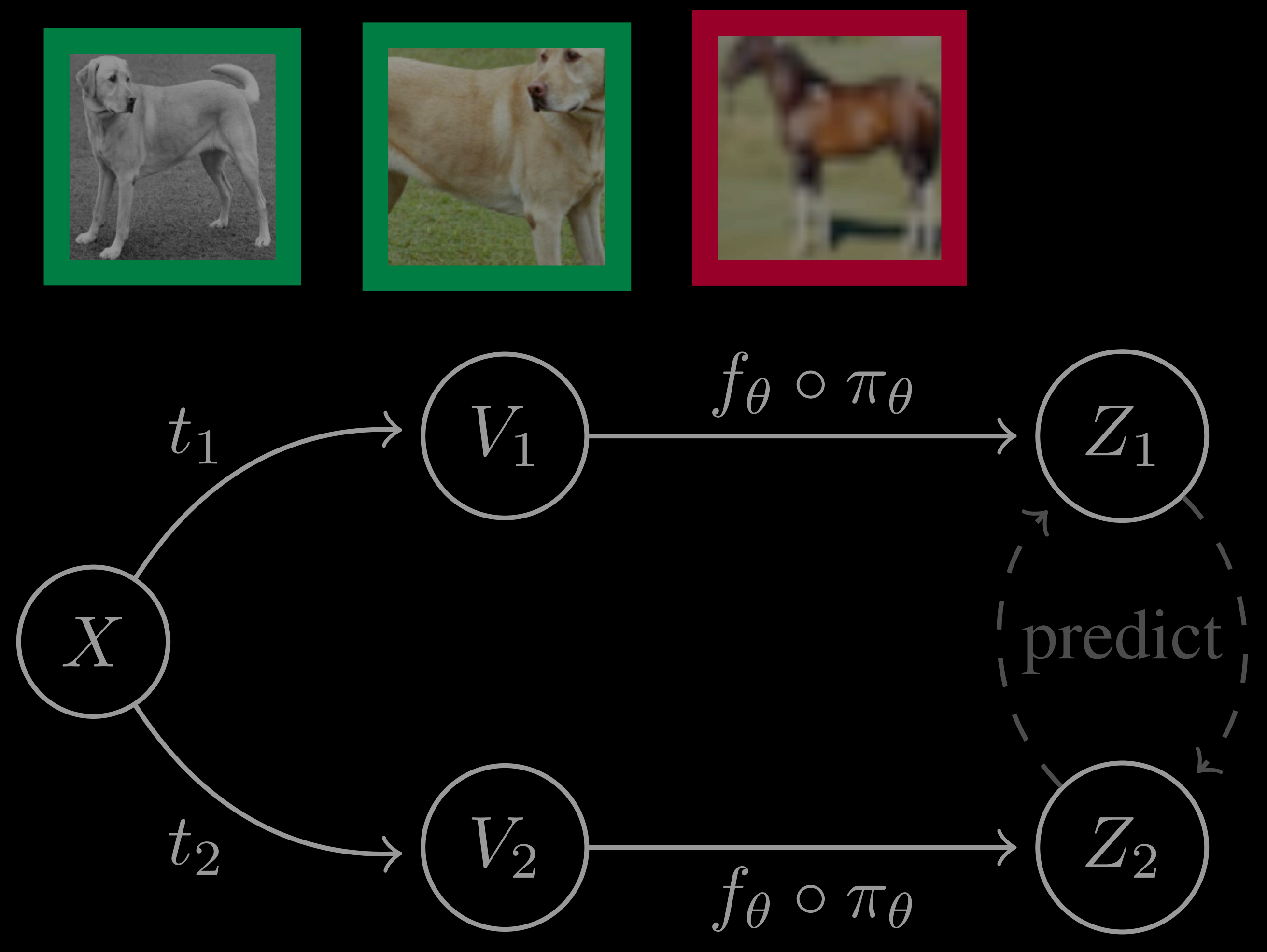
SwAV, DeepCluster



Distillation-based methods

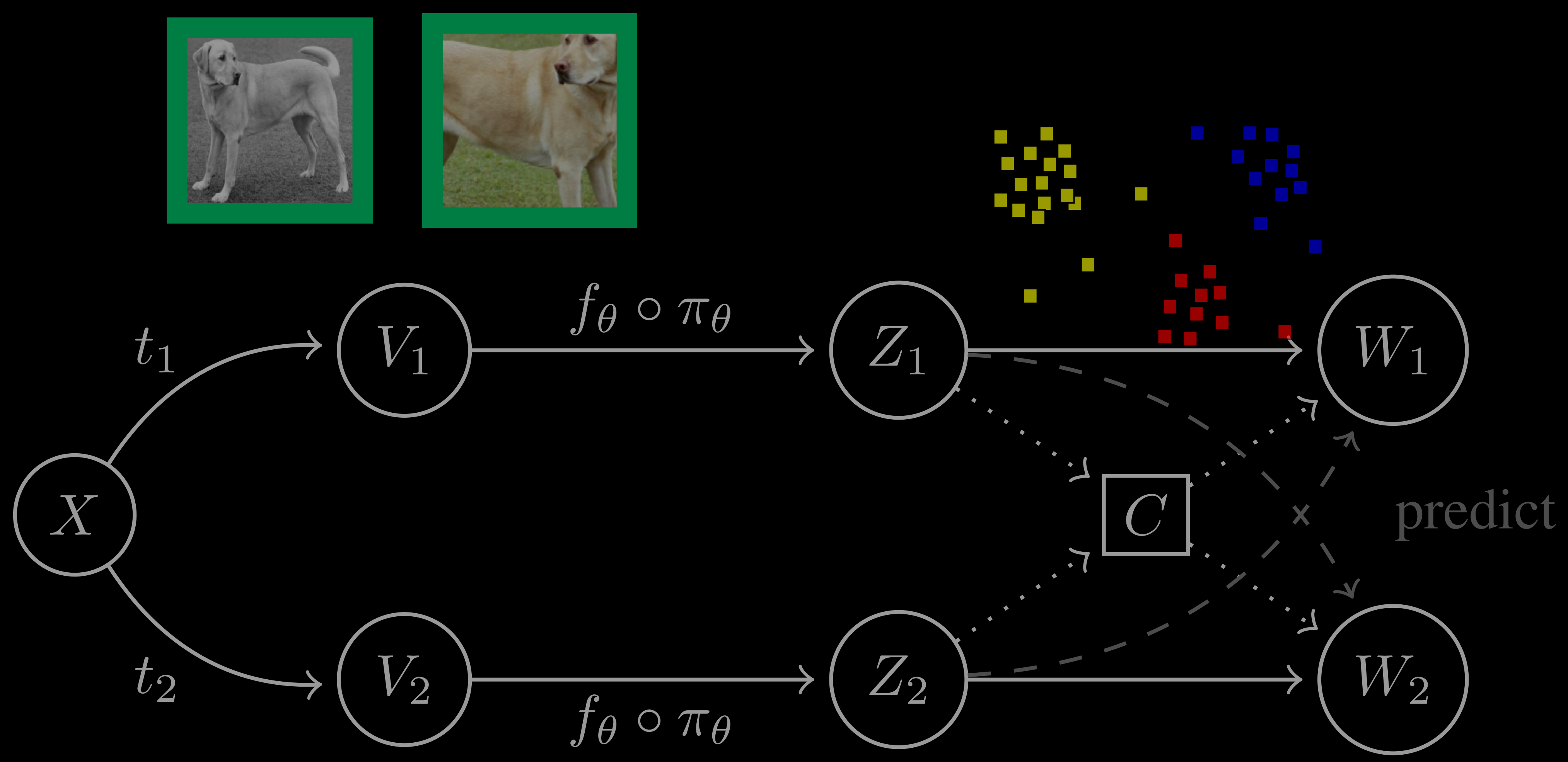
BYOL, DINO

Prior work: What role does MI optimization play in MVSSL?



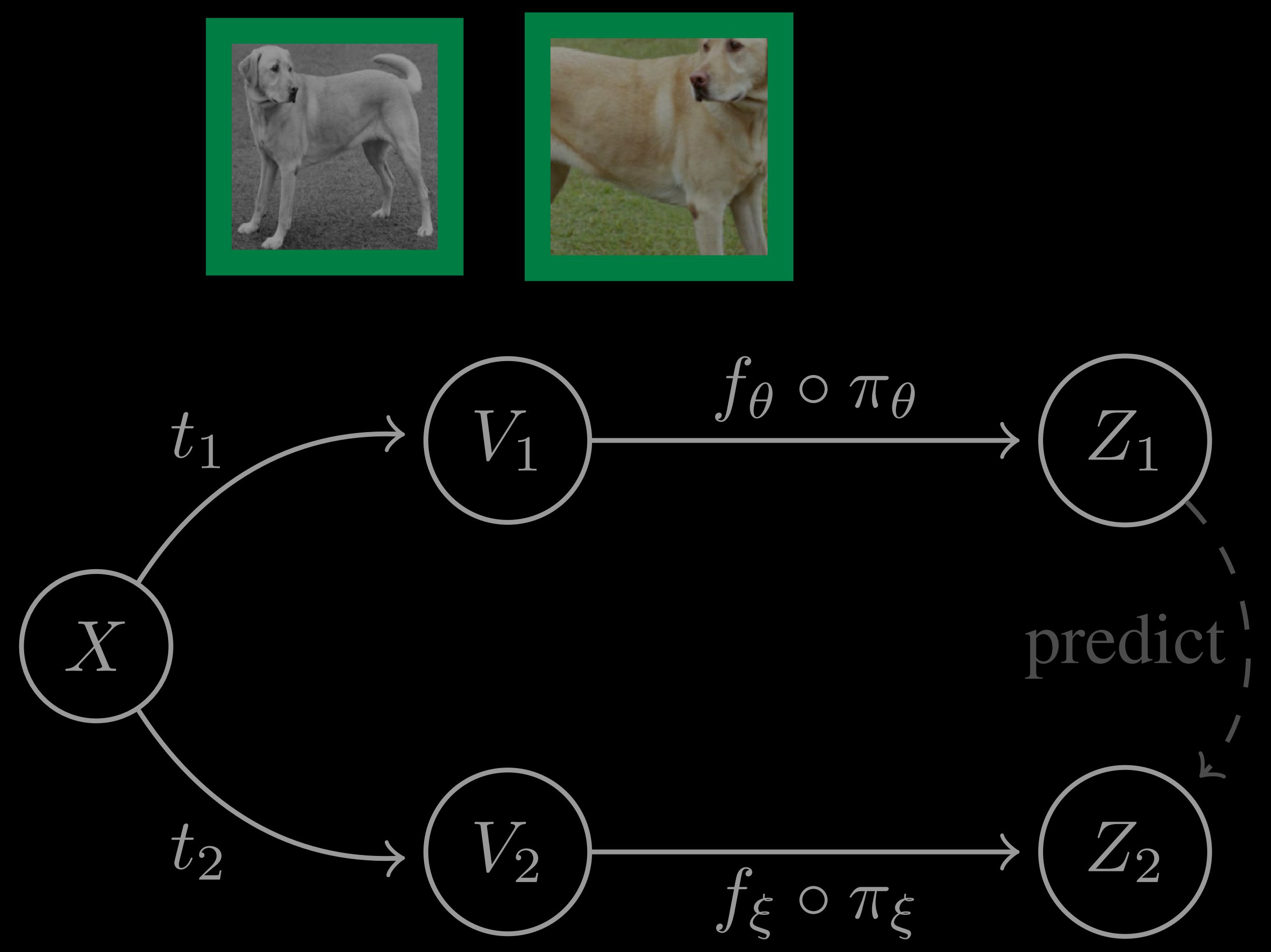
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

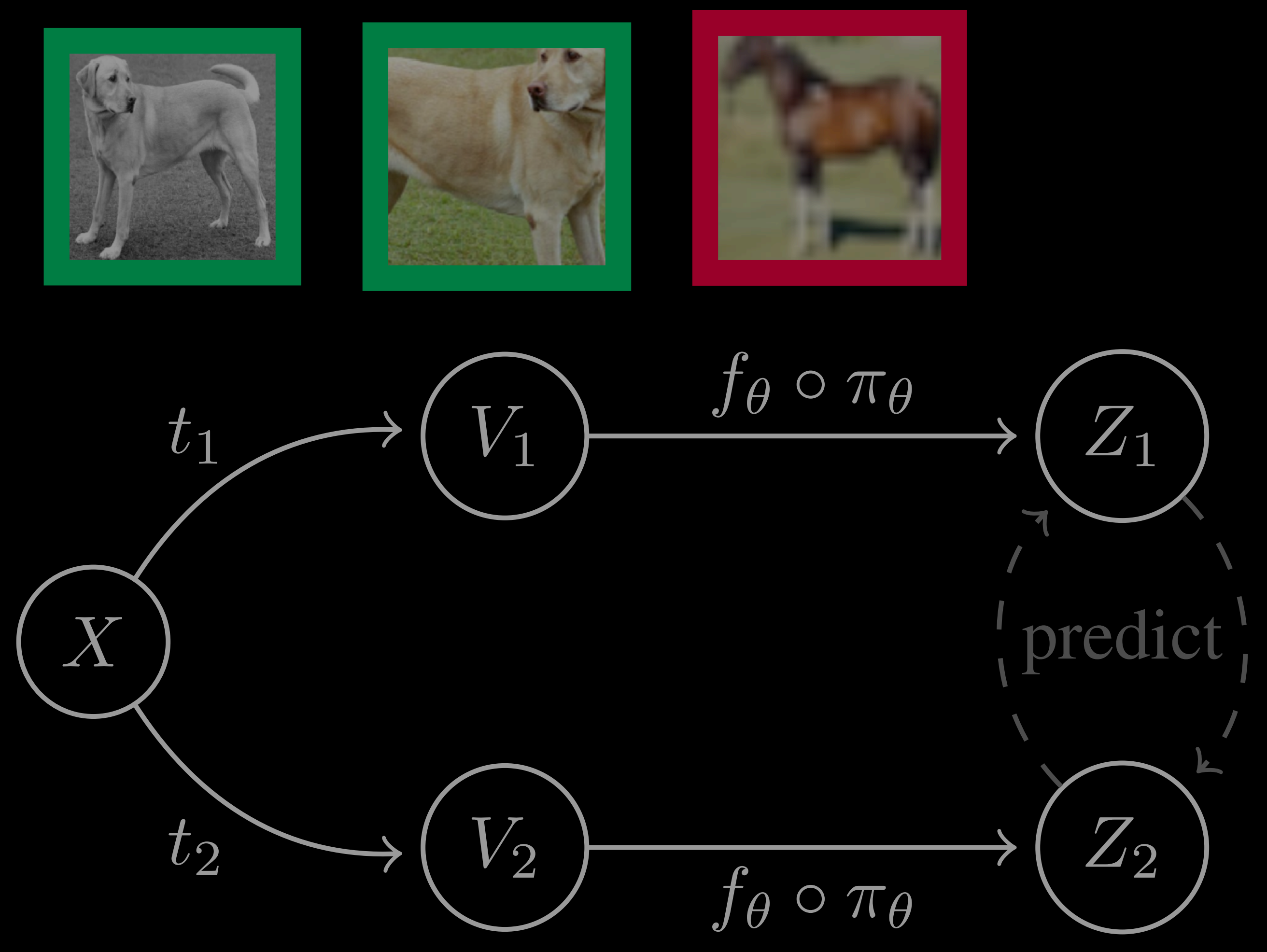
SwAV, DeepCluster



Distillation-based methods

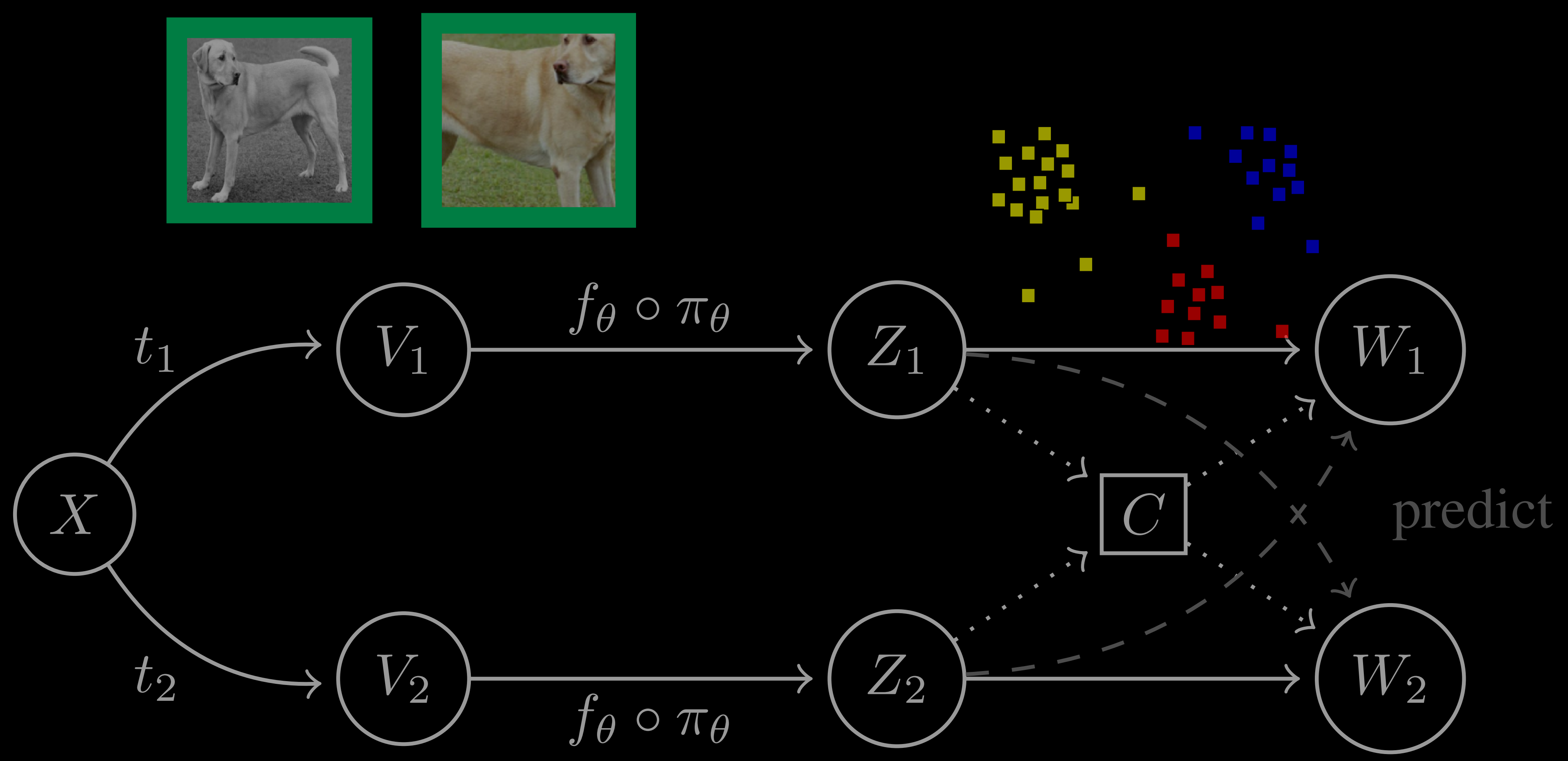
BYOL, DINO

Prior work: What role does MI optimization play in MVSSL?



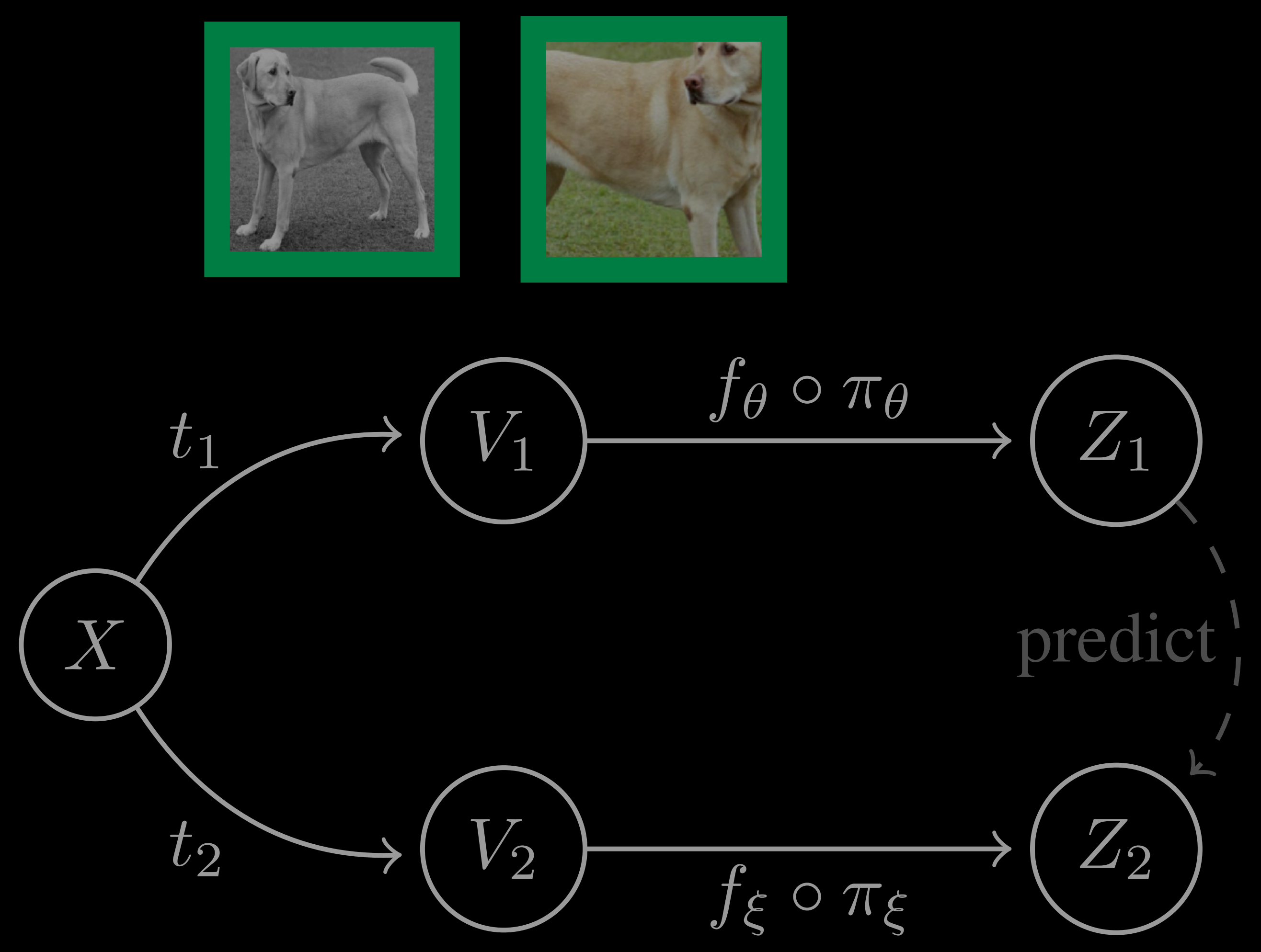
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

SwAV, DeepCluster

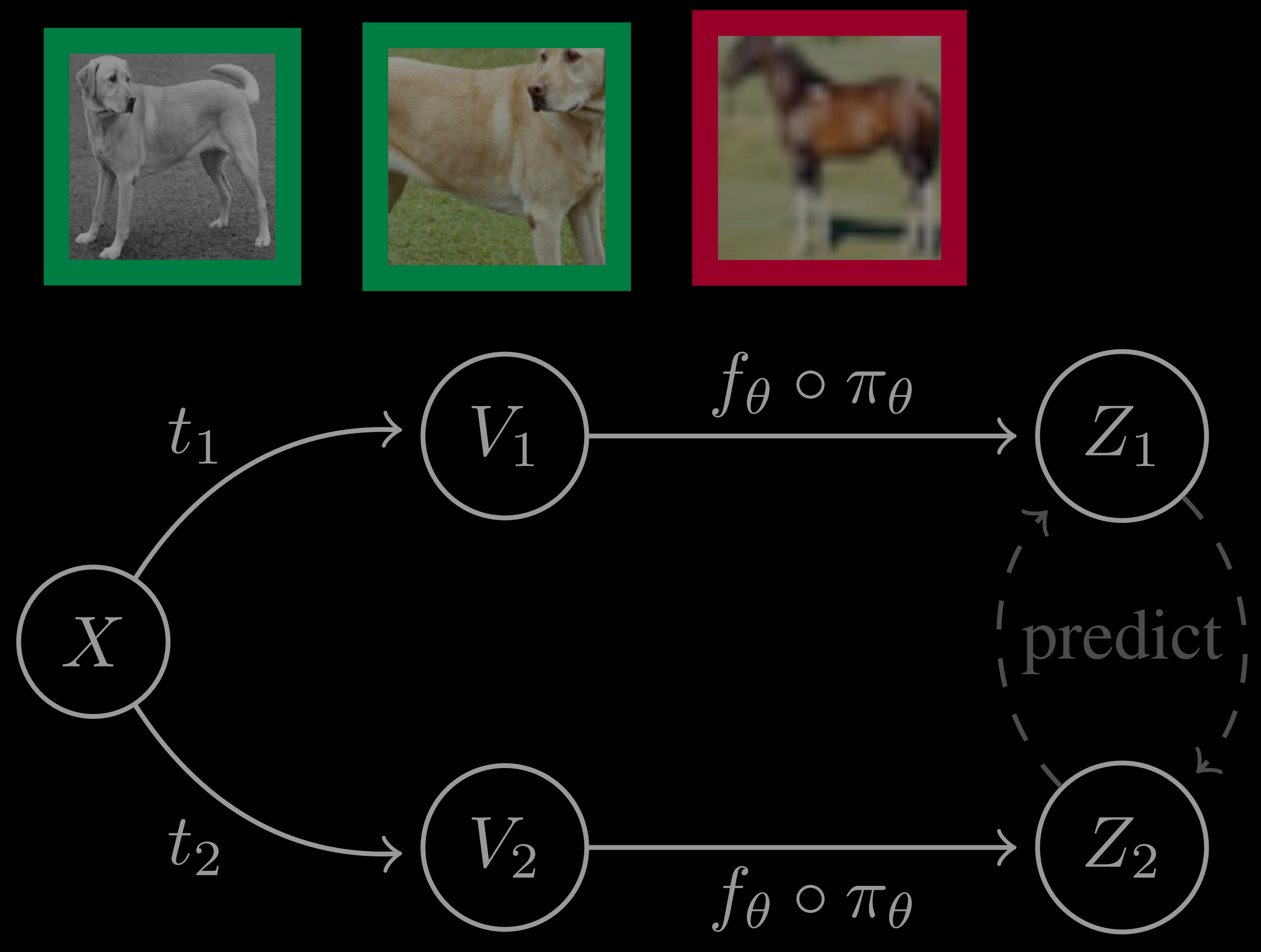


Distillation-based methods

BYOL, DINO

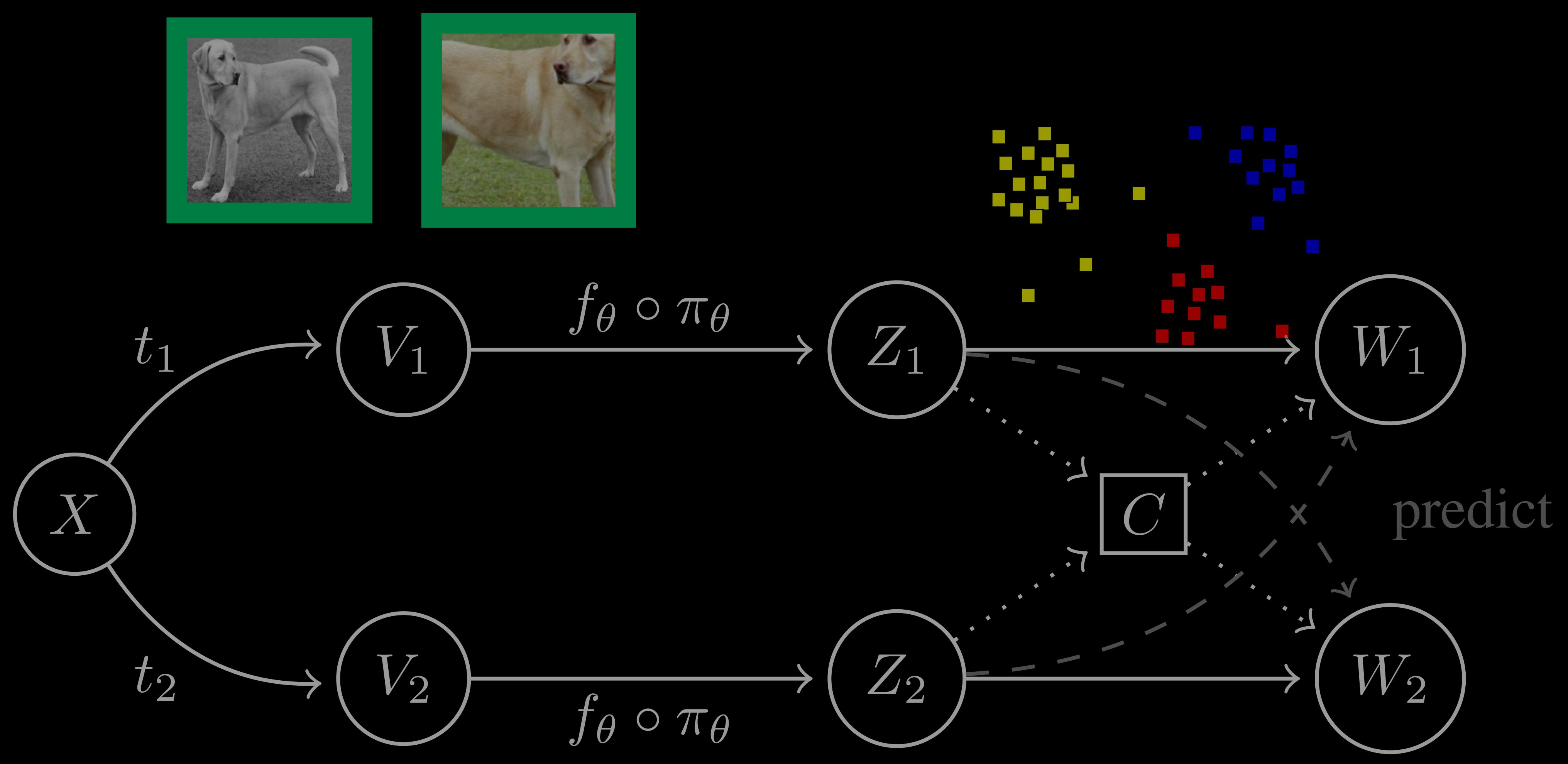
Some optimize $I(Z_1; Z_2)$

Prior work: What role does MI optimization play in MVSSL?



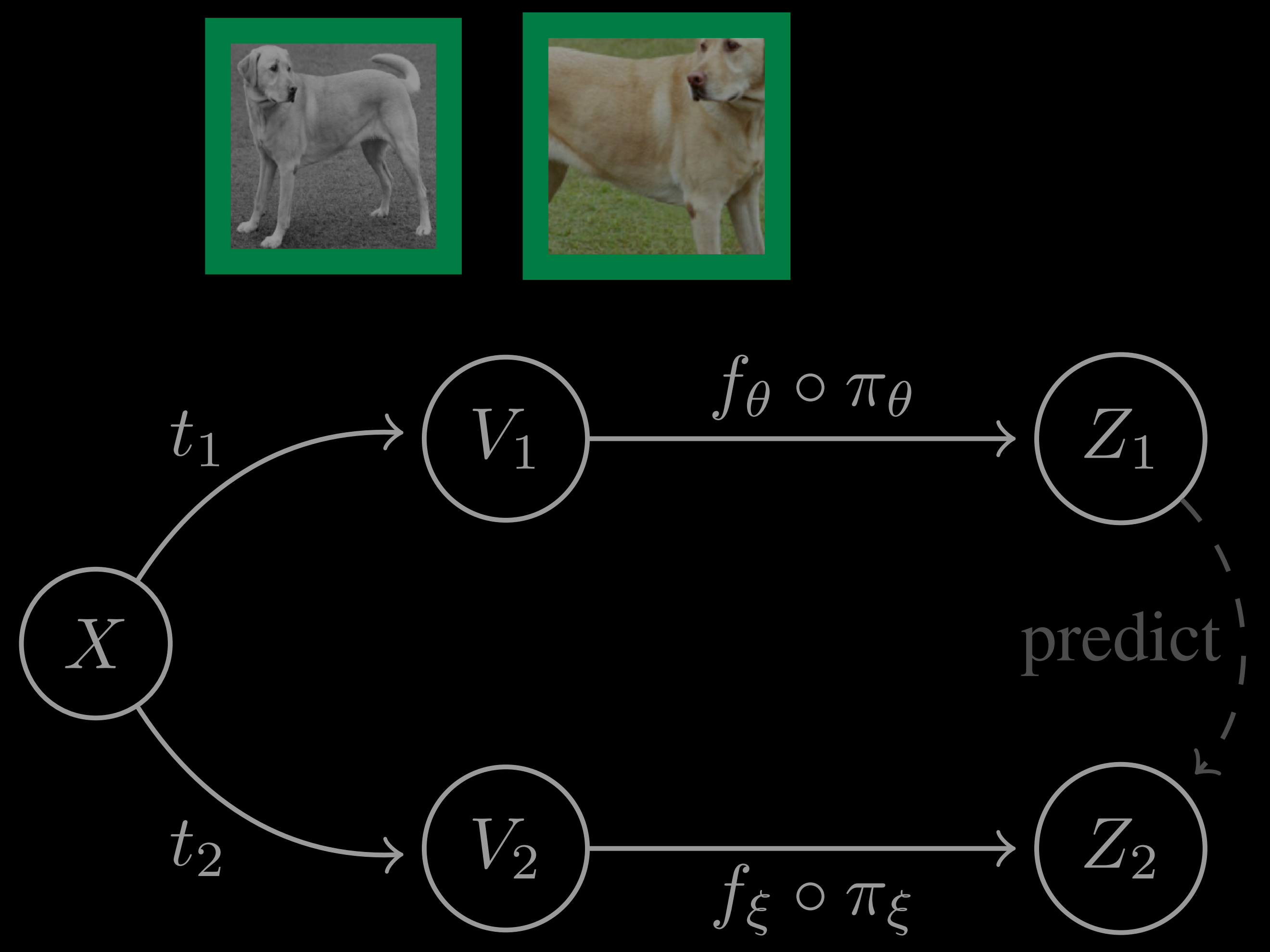
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

SwAV, DeepCluster



Distillation-based methods

BYOL, DINO

Some optimize $I(Z_1; Z_2)$

?

?

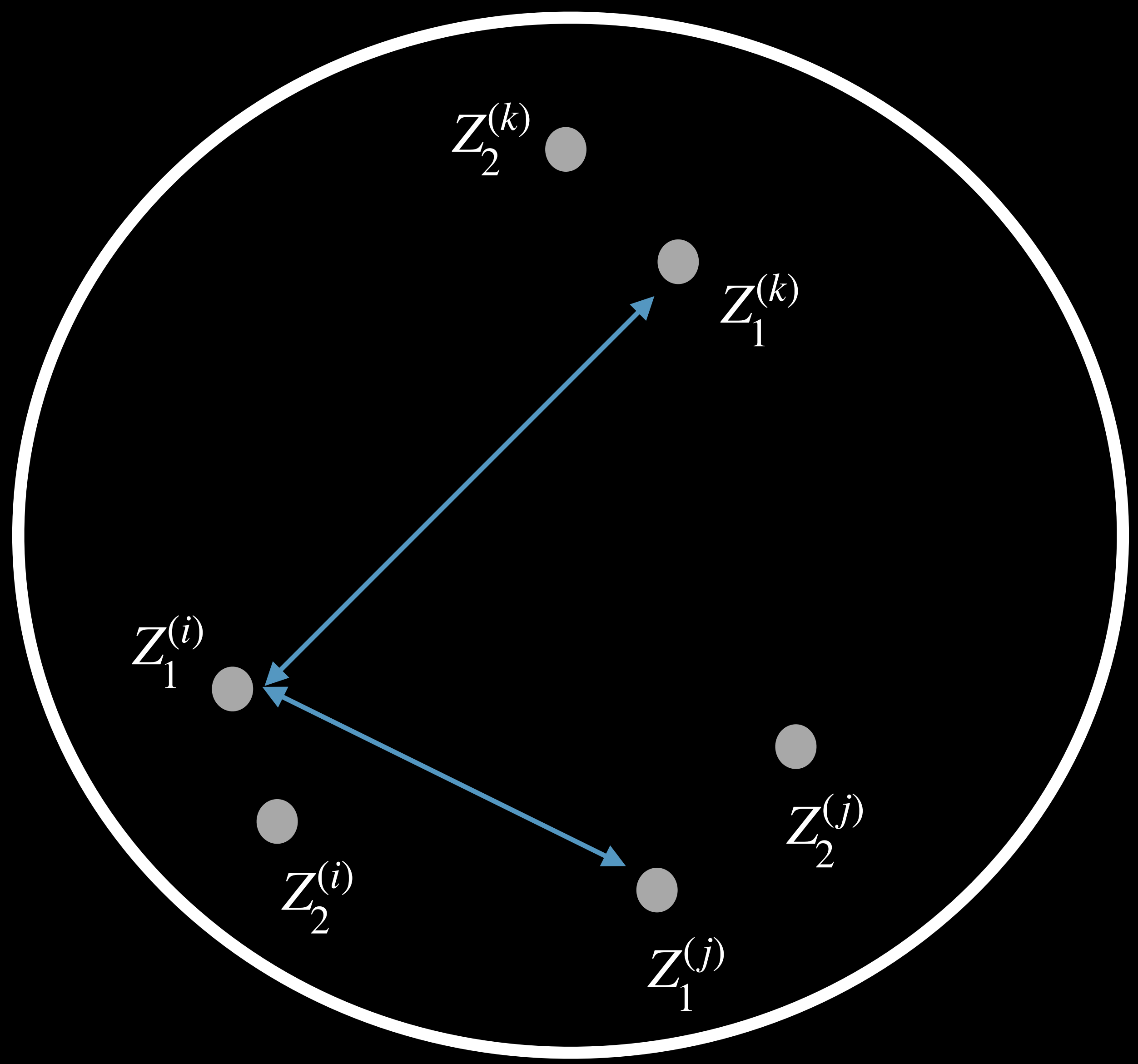
Analysis: Using a different bound on MI w.r.t. prior works

$$
\begin{aligned}
I(Z_1; Z_2) &= H(Z_2) - H(Z_2 \mid Z_1) \\
&\geq H(Z_2) - \mathbb{E}_{Z_1, Z_2}\left[-\log q_{Z_2|Z_1}(Z_2) \right] := I_{\mathsf{ER}}(Z_1; Z_2)
\end{aligned}
$$

Analysis: Using a different bound on MI w.r.t. prior works

$$
\begin{aligned}
I(Z_1;Z_2) &= H(Z_2) - H(Z_2 \mid Z_1) \\
&\geq \boxed{H(Z_2)} - \mathbb{E}_{Z_1,Z_2}\left[-\log q_{Z_2|Z_1}(Z_2)\right] := I_{\mathrm{ER}}(Z_1;Z_2)
\end{aligned}
$$

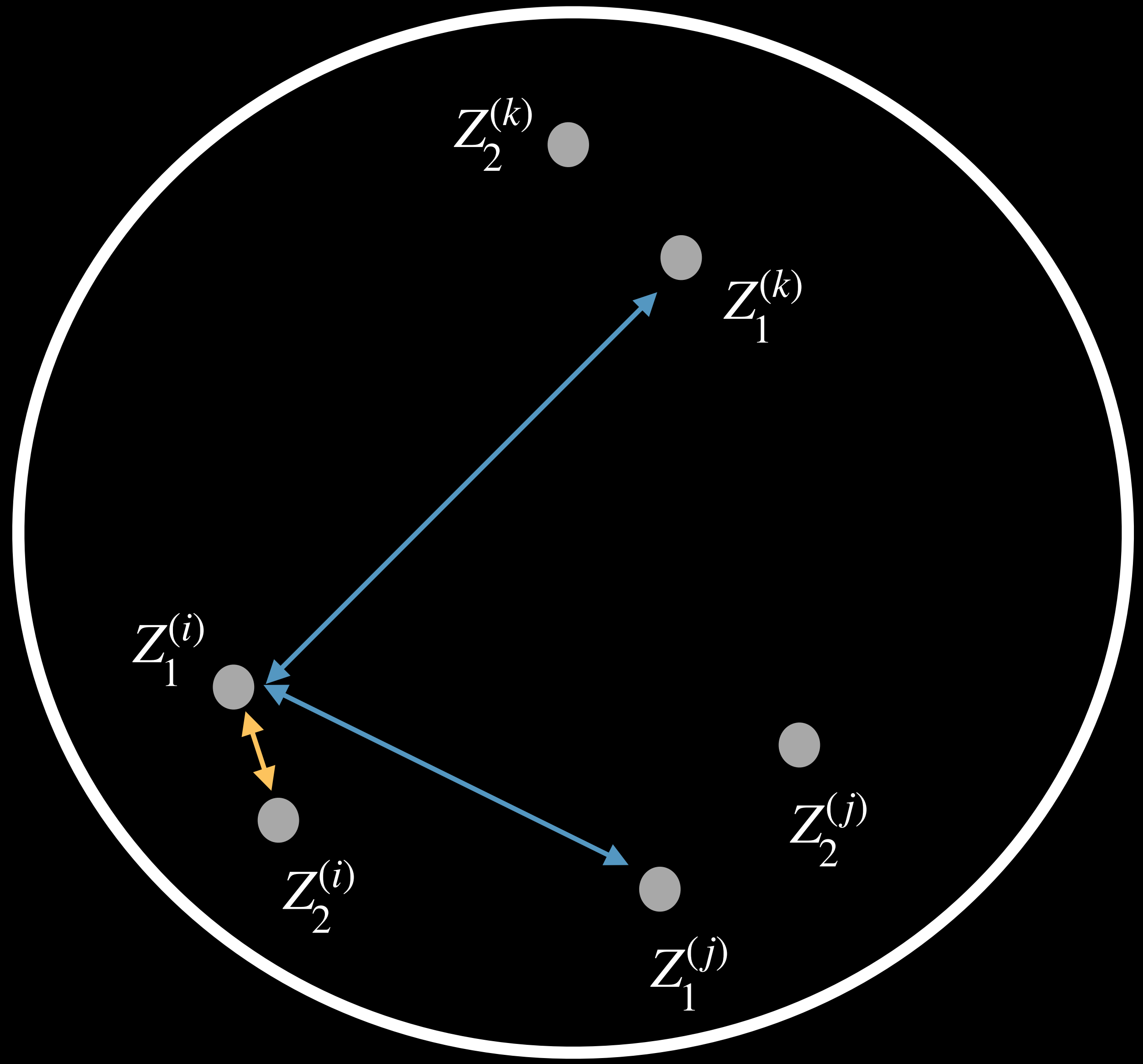
- Entropy: How much information *can* be learnt



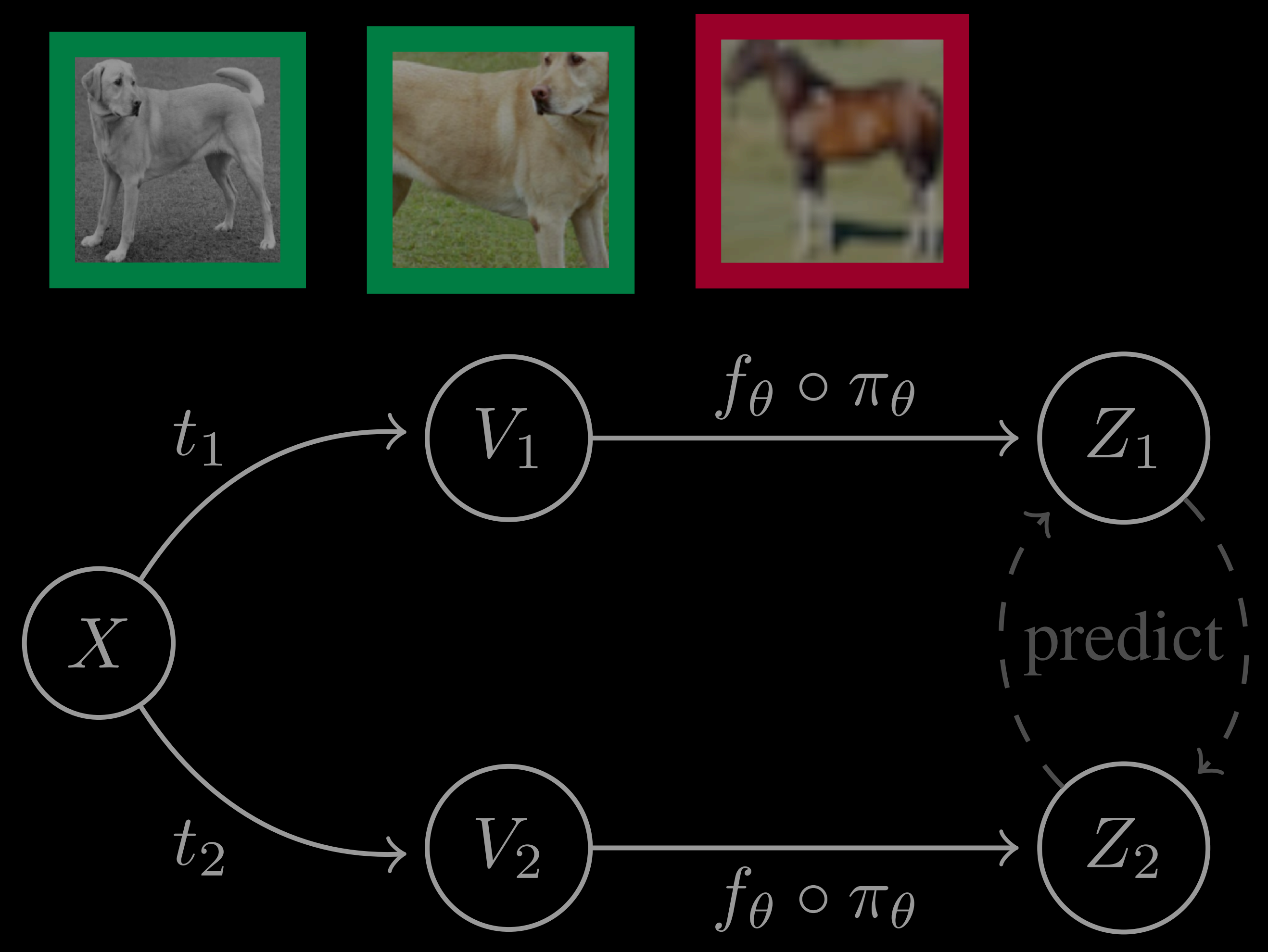
Analysis: Using a different bound on MI w.r.t. prior works

$$
\begin{aligned}
I(Z_1; Z_2) &= H(Z_2) - H(Z_2 \mid Z_1) \\
&\geq \boxed{H(Z_2)} - \boxed{\mathbb{E}_{Z_1, Z_2}\left[-\log q_{Z_2 \mid Z_1}(Z_2) \right]} := I_{\mathrm{ER}}(Z_1; Z_2)
\end{aligned}
$$

- Entropy: How much information *can* be learnt
- Reconstruction: How much information *is* learnt

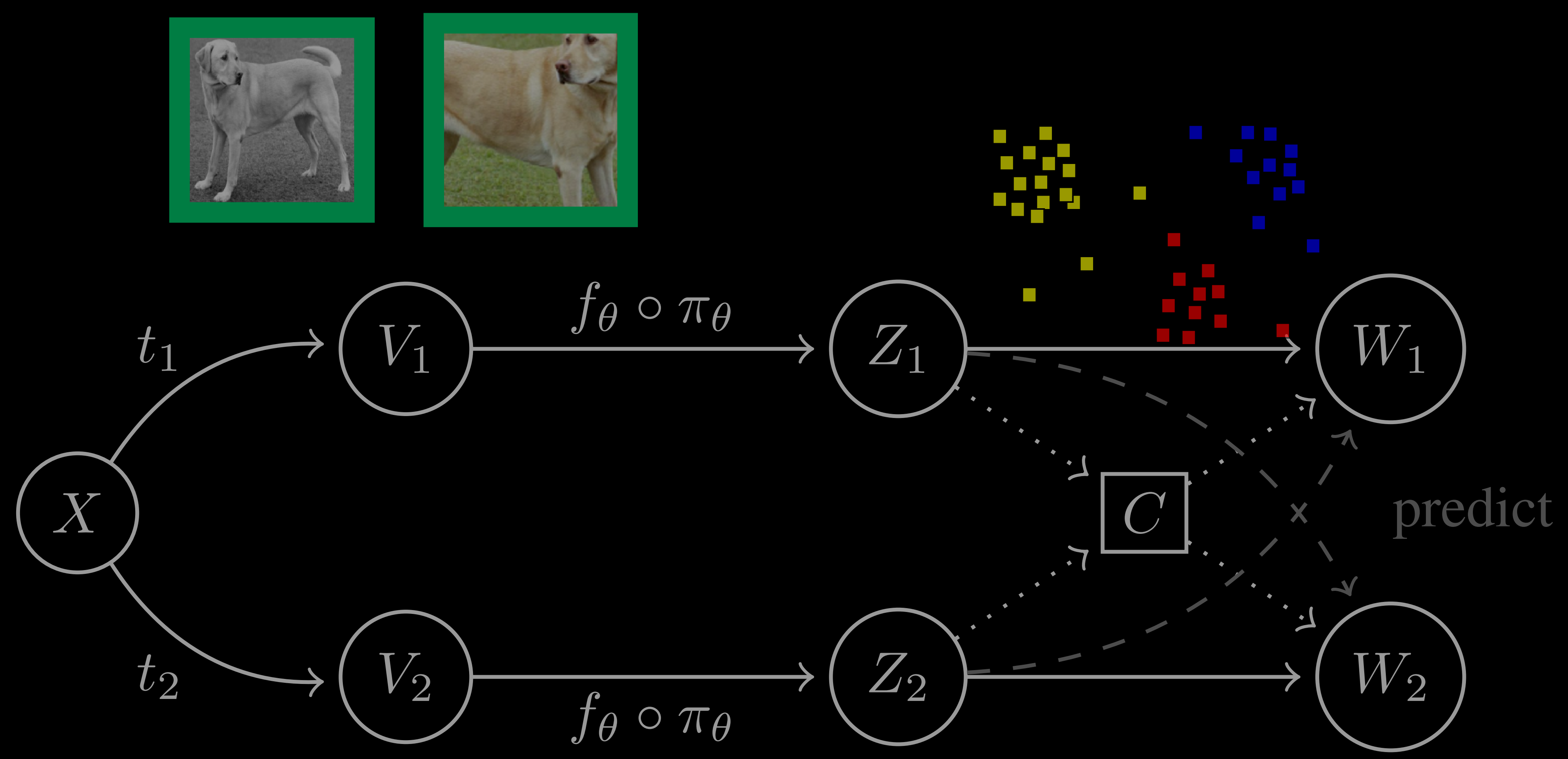


Theoretical analysis



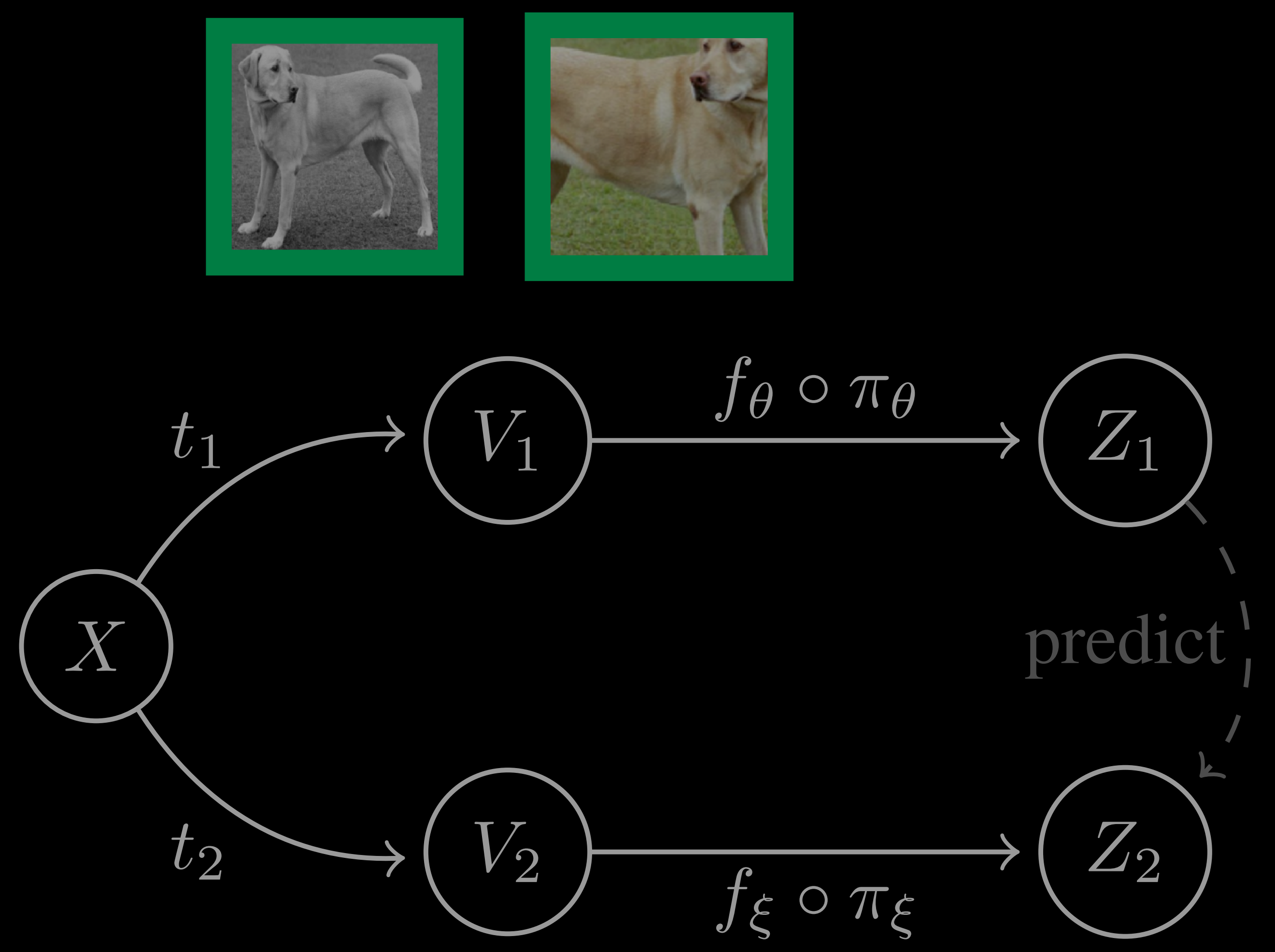
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

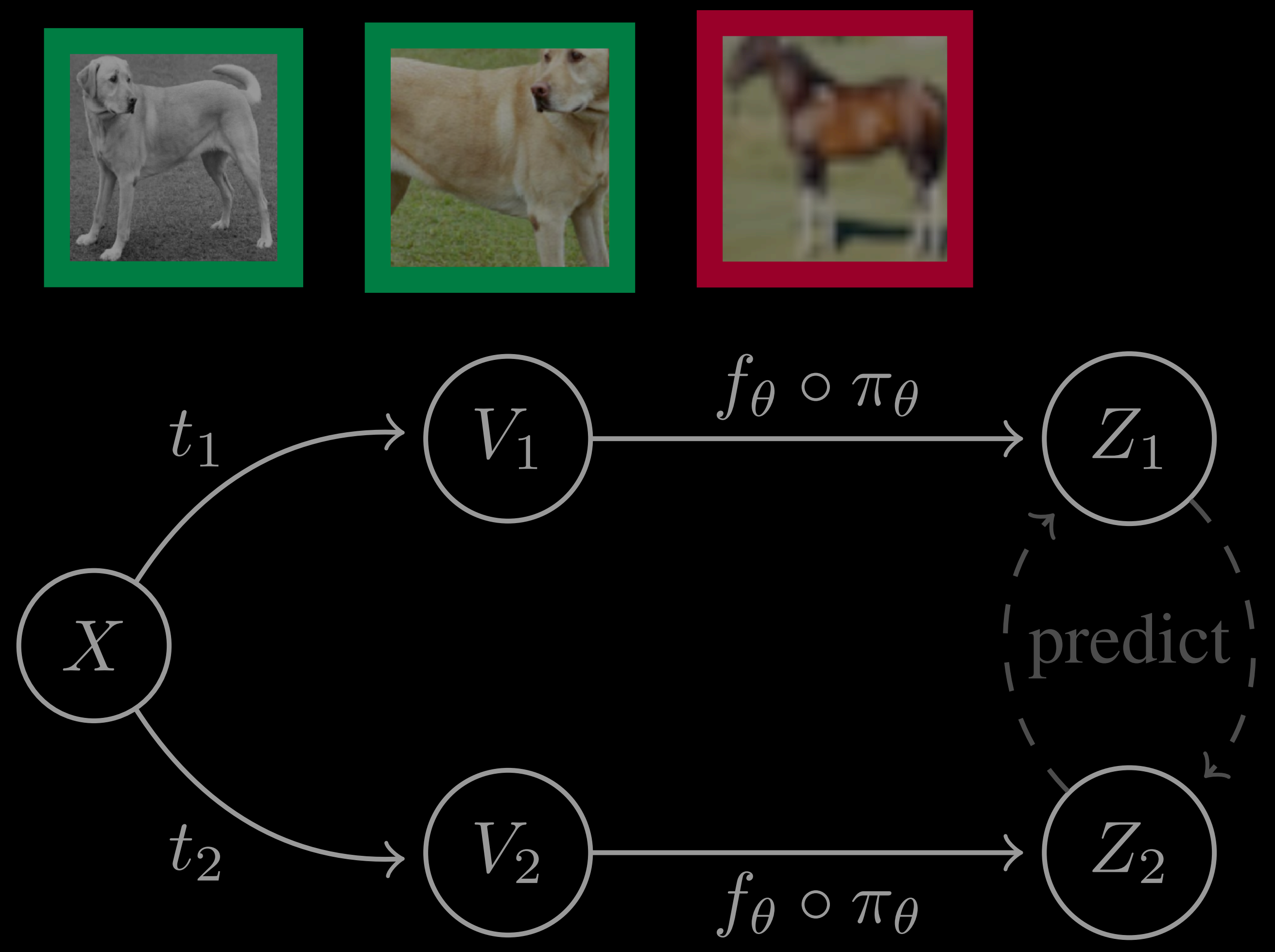
SwAV, DeepCluster



Distillation-based methods

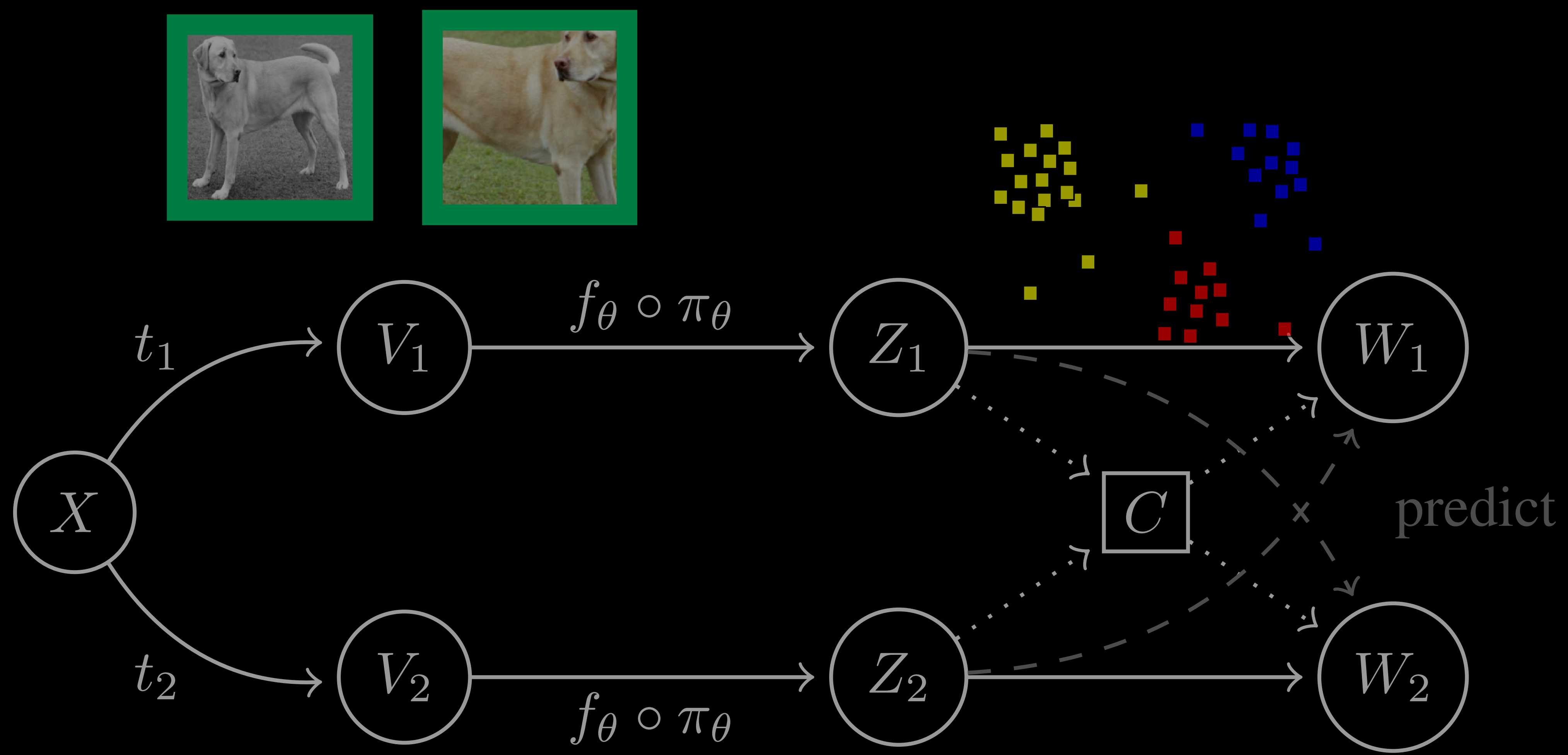
BYOL, DINO

Theoretical analysis



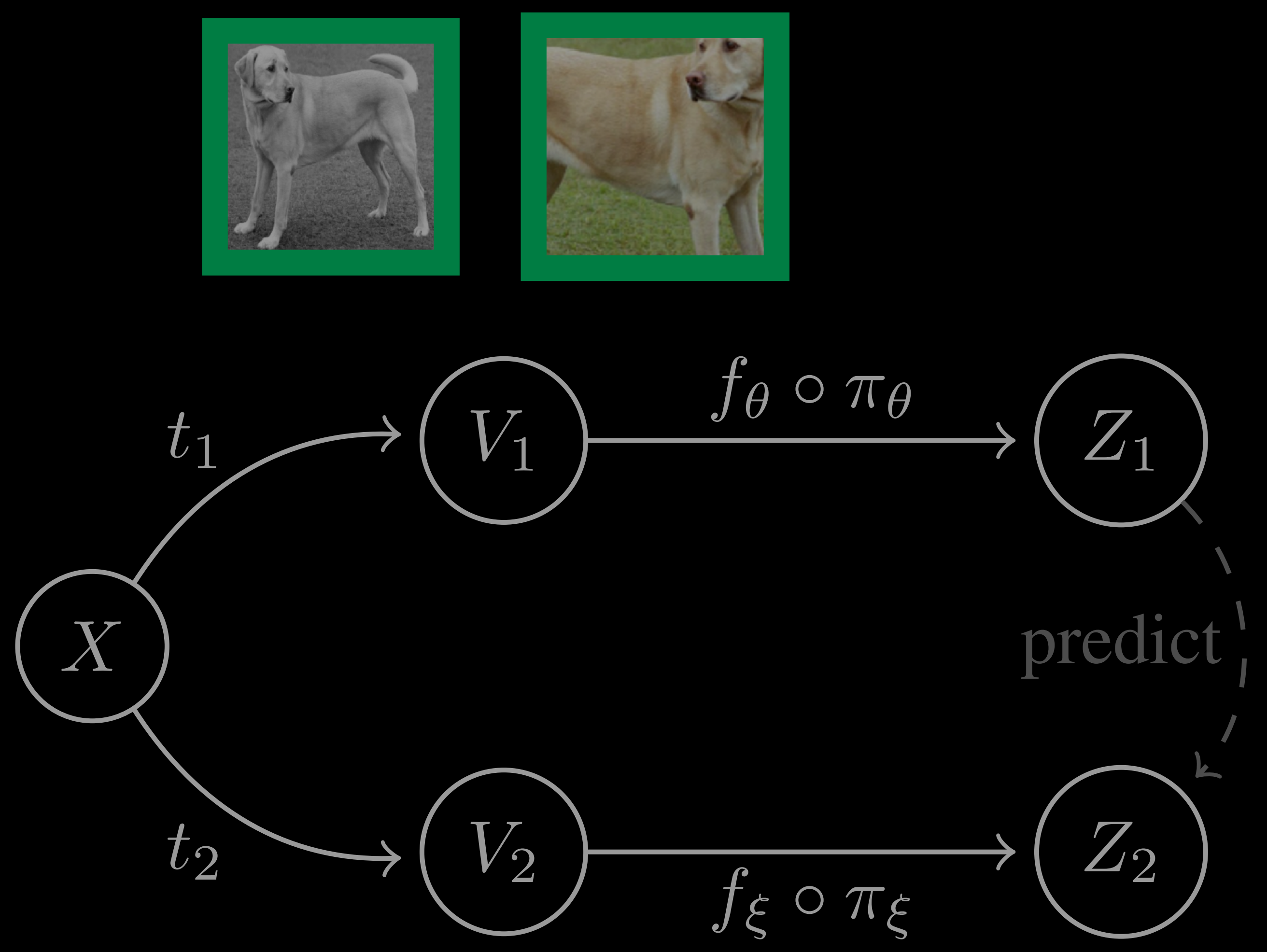
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

SwAV, DeepCluster

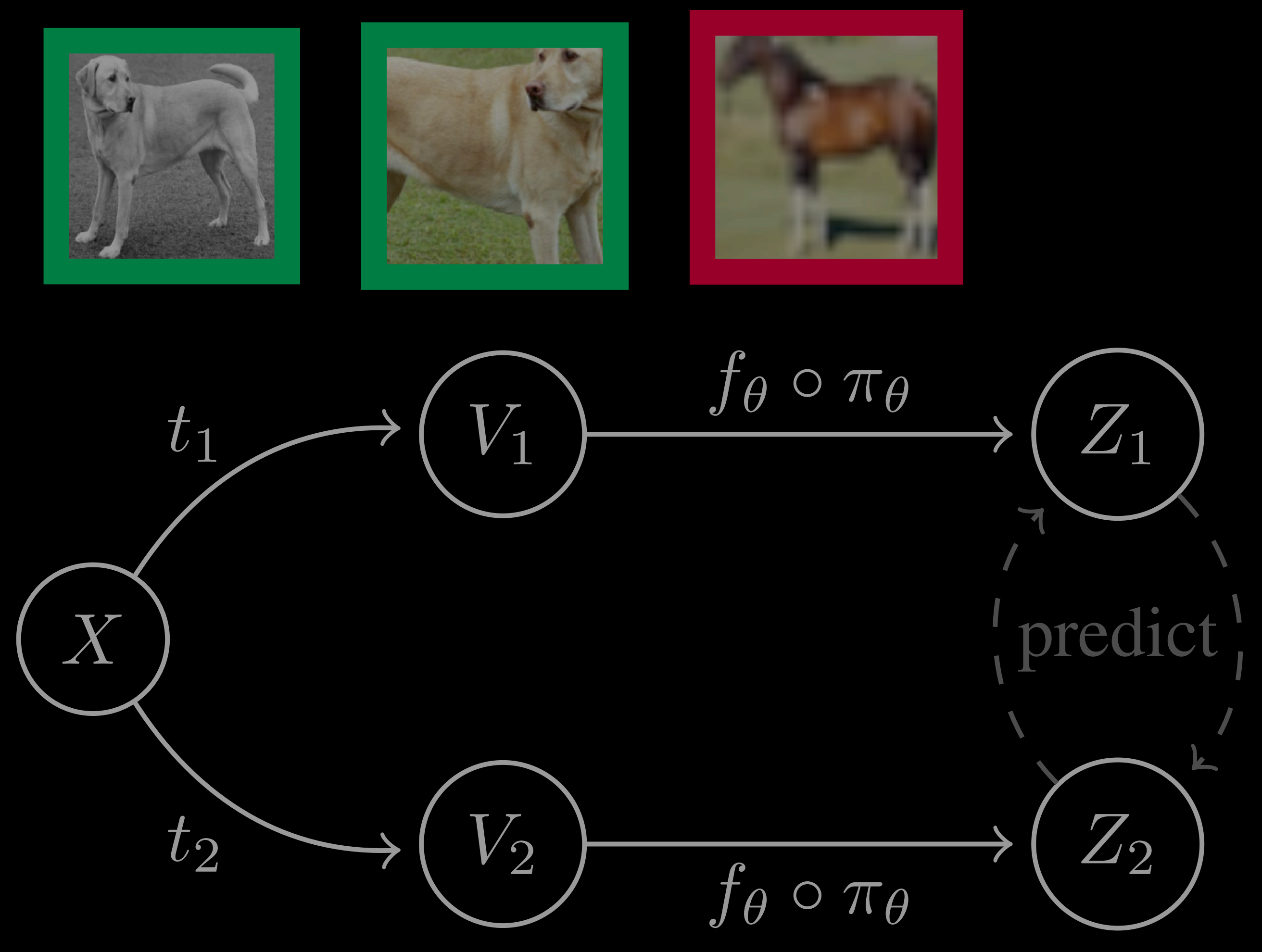


Distillation-based methods

BYOL, DINO

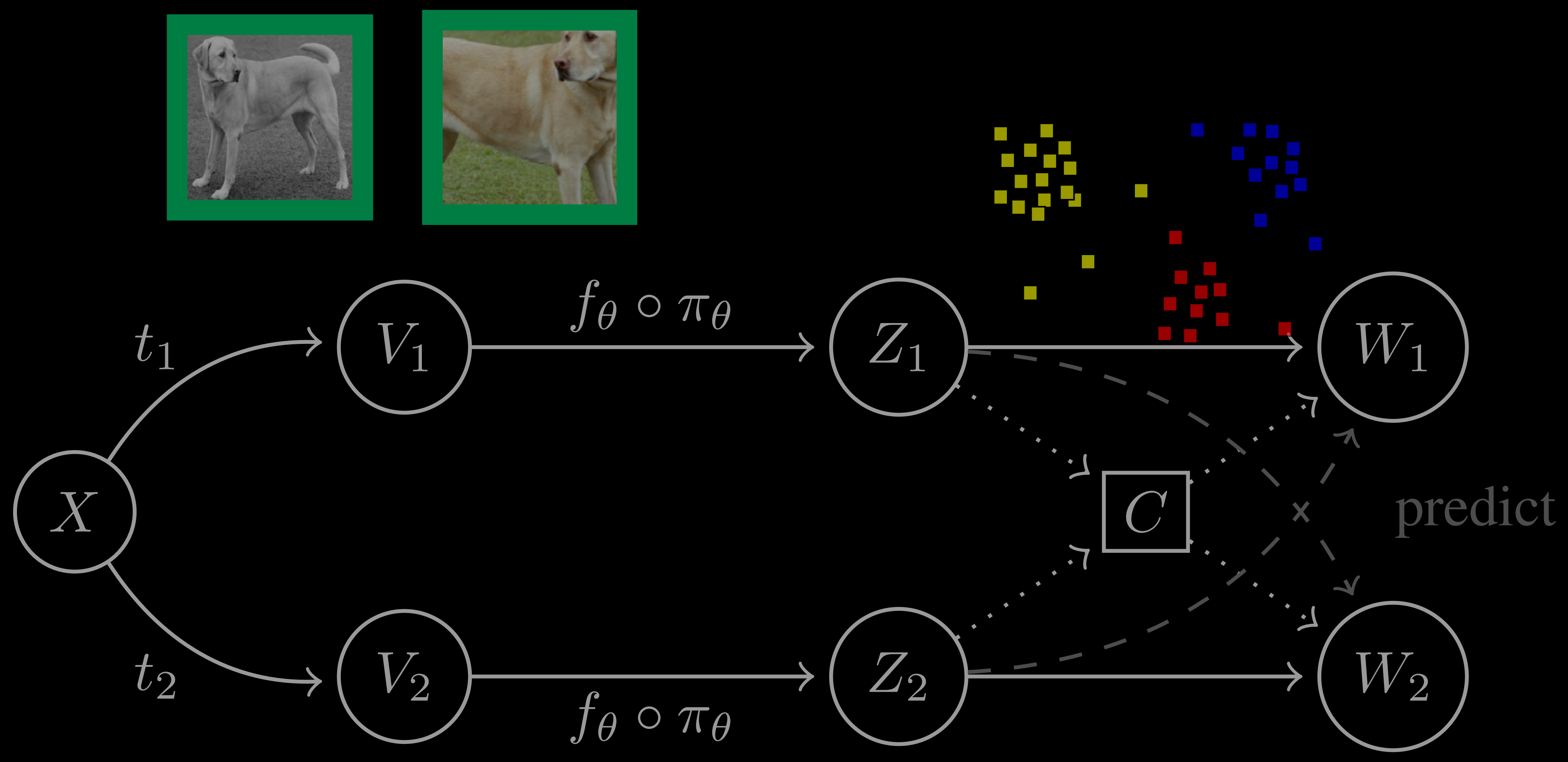
Some optimize $I(Z_1; Z_2)$ exactly, some not exactly

Theoretical analysis



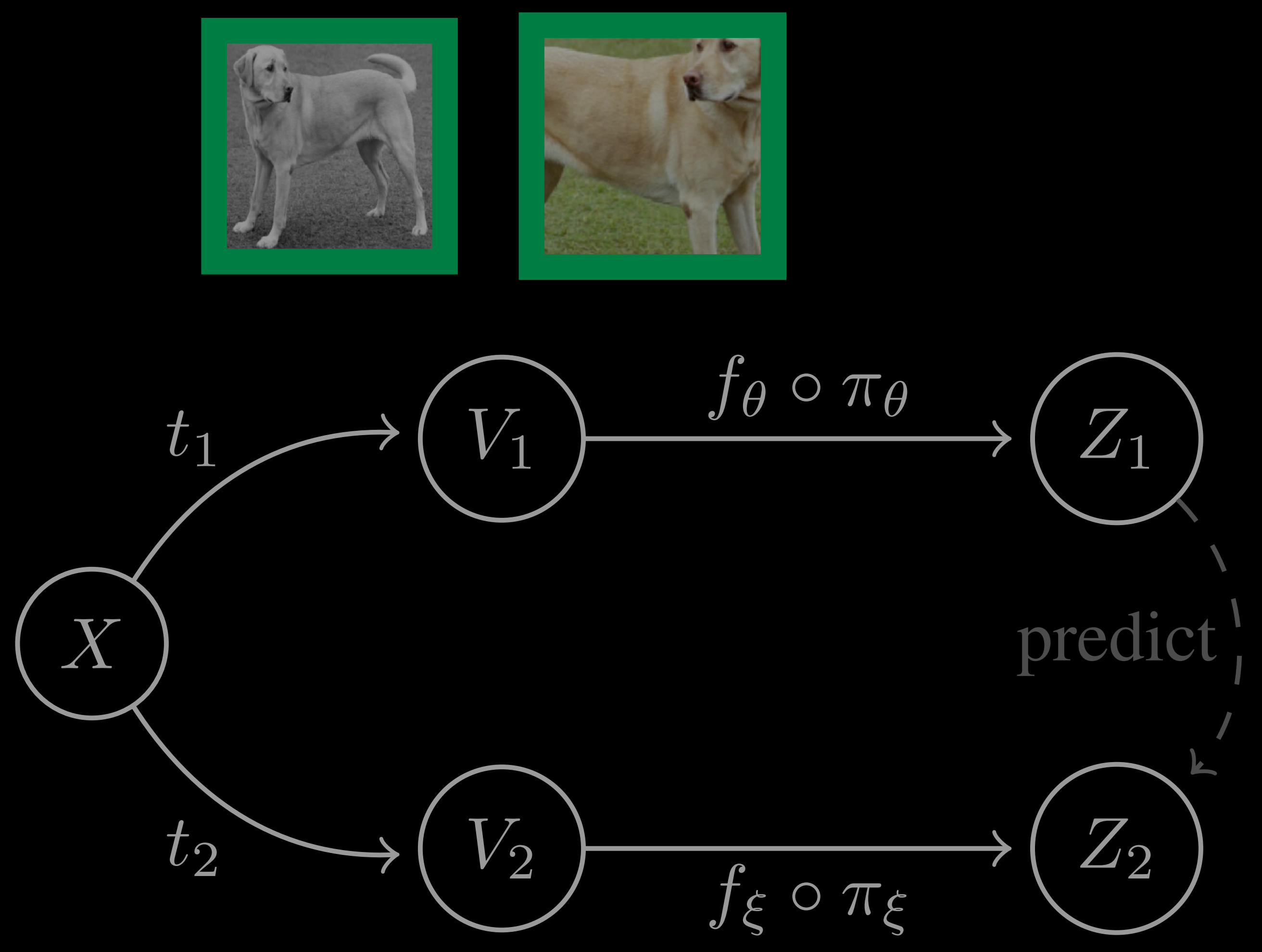
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

SwAV, DeepCluster



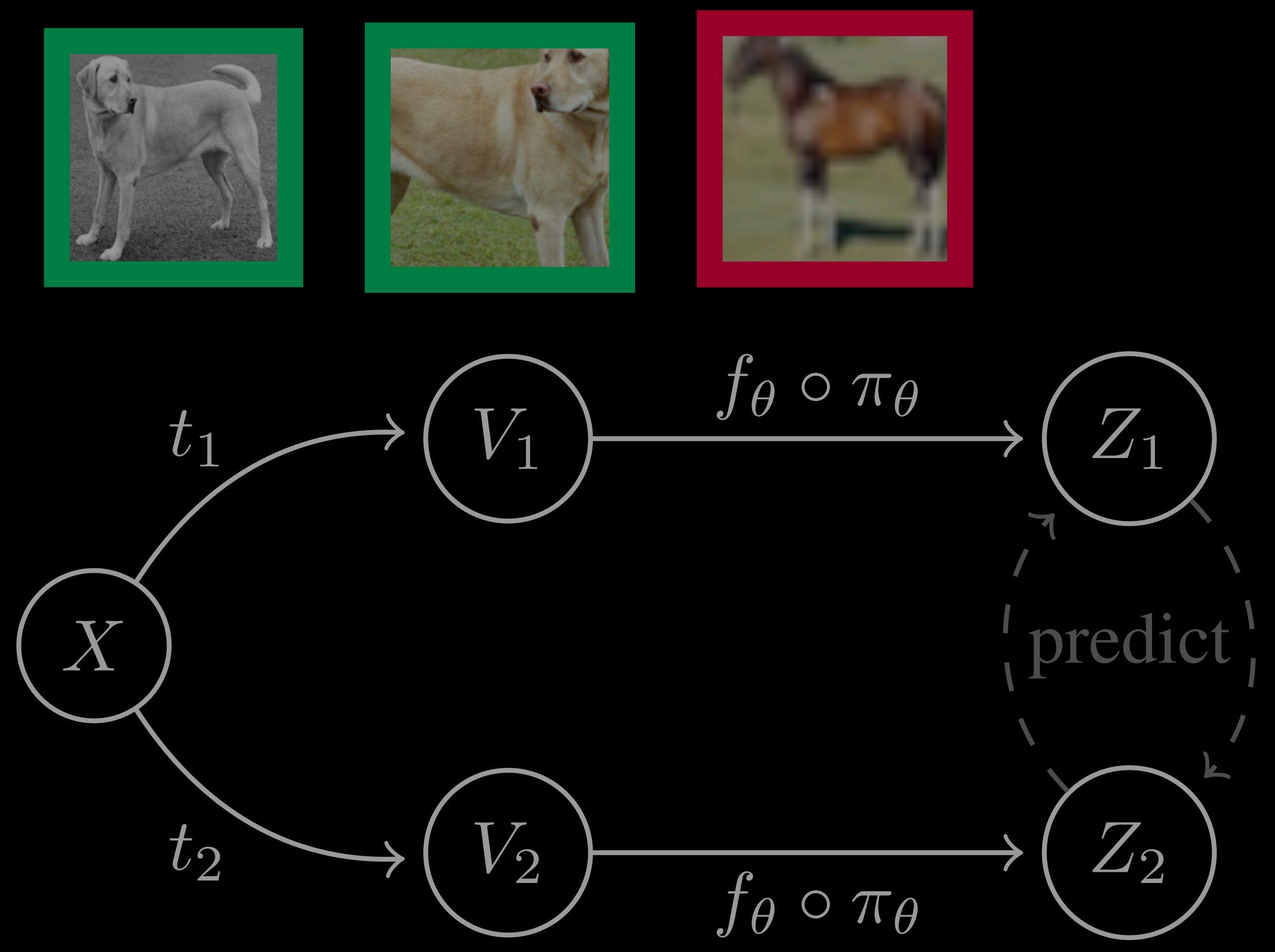
Distillation-based methods

BYOL, DINO

Some optimize $I(Z_1; Z_2)$ exactly, some not exactly

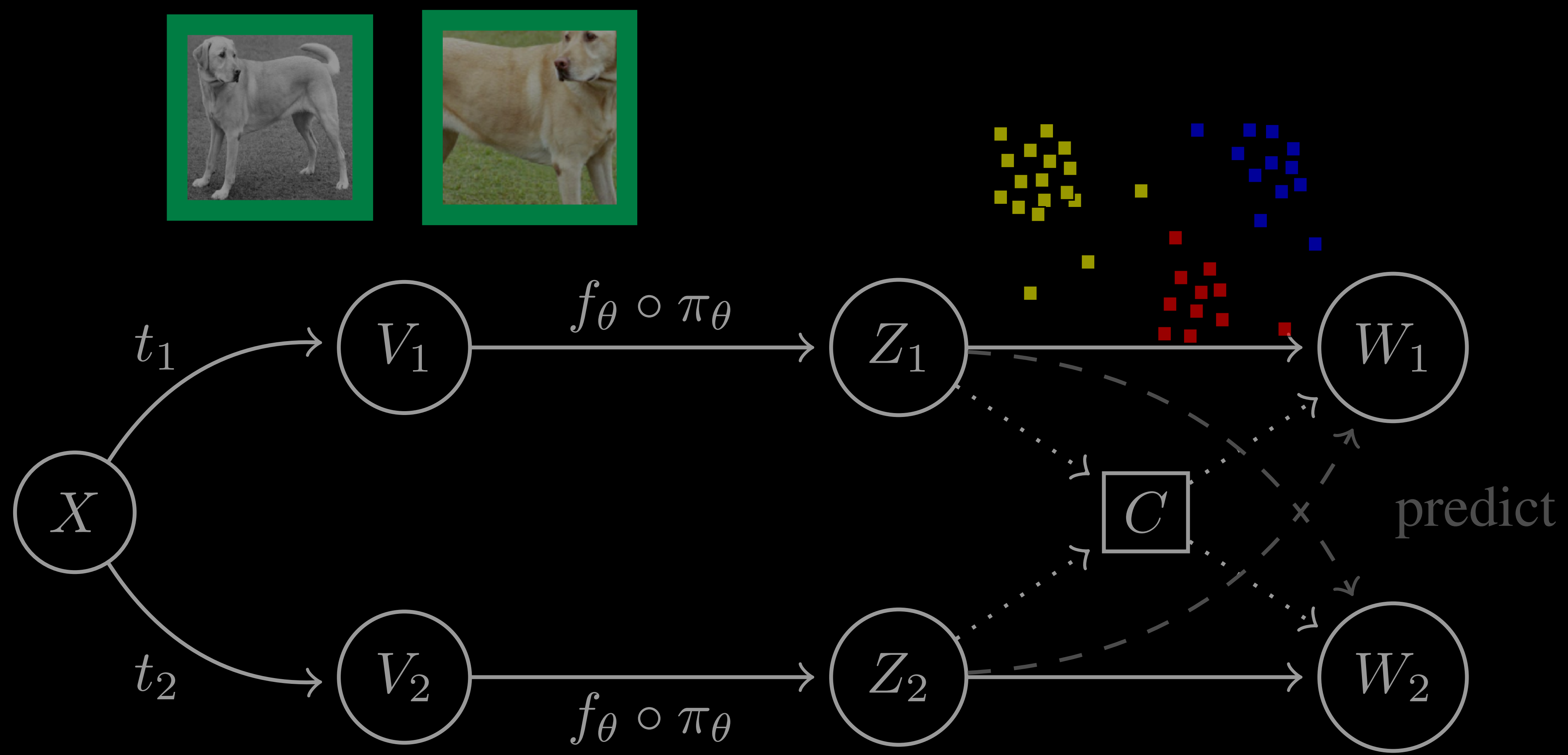
Optimize $I(Z_1; Z_2)$ exactly

Theoretical analysis



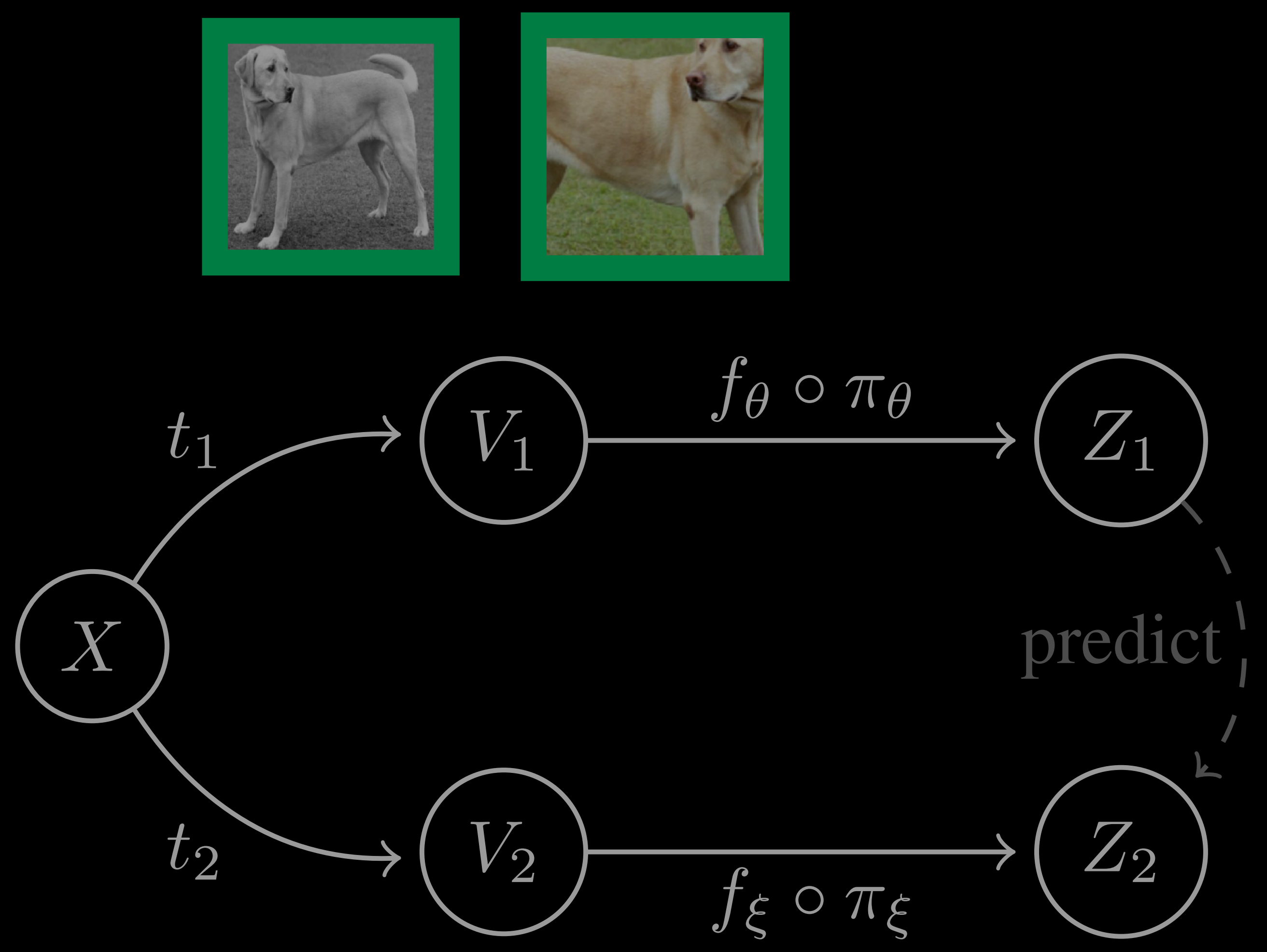
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

SwAV, DeepCluster



Distillation-based methods

BYOL, DINO

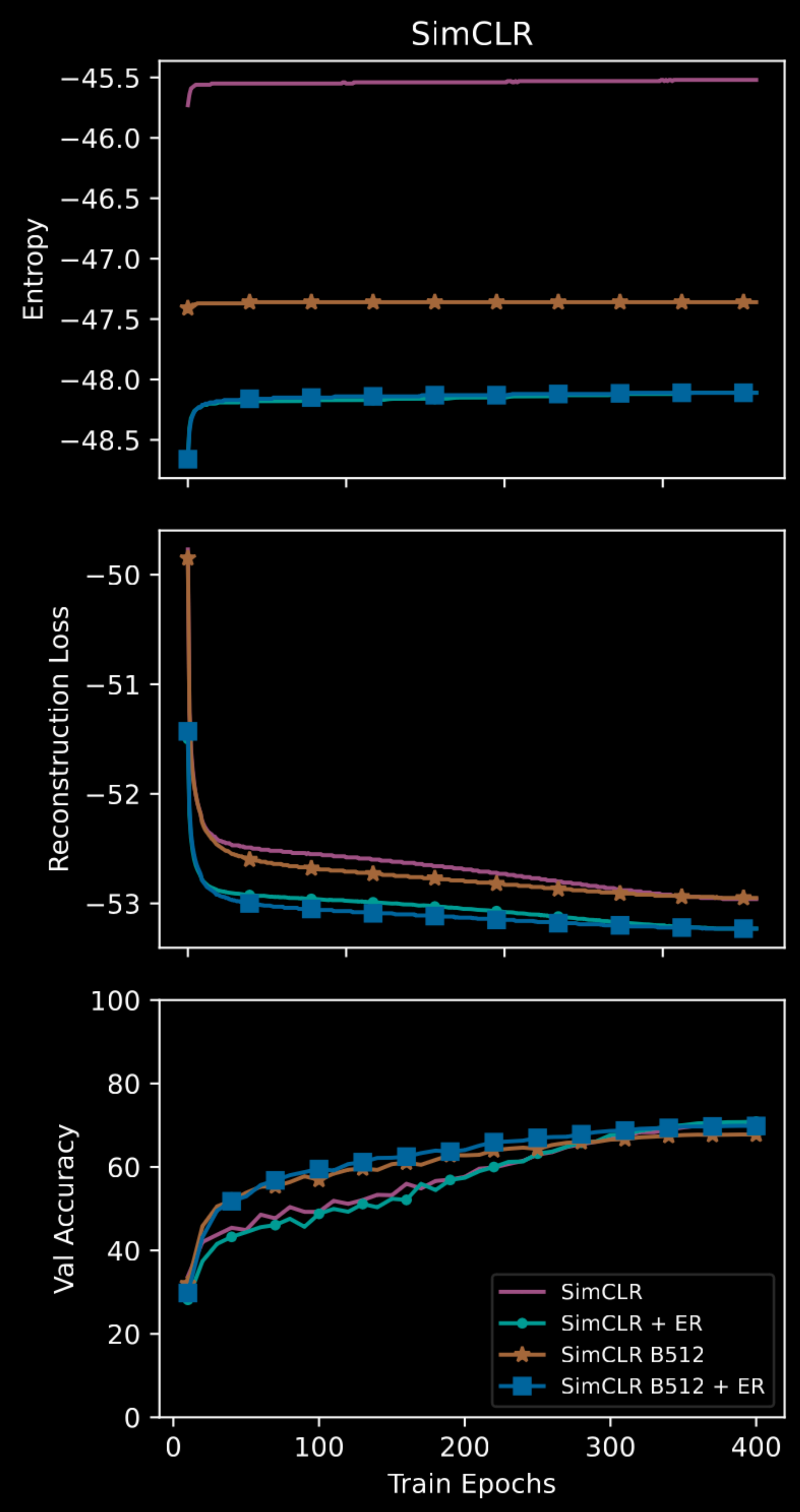
Some optimize $I(Z_1; Z_2)$ exactly, some not exactly

Optimize $I(Z_1; Z_2)$ exactly

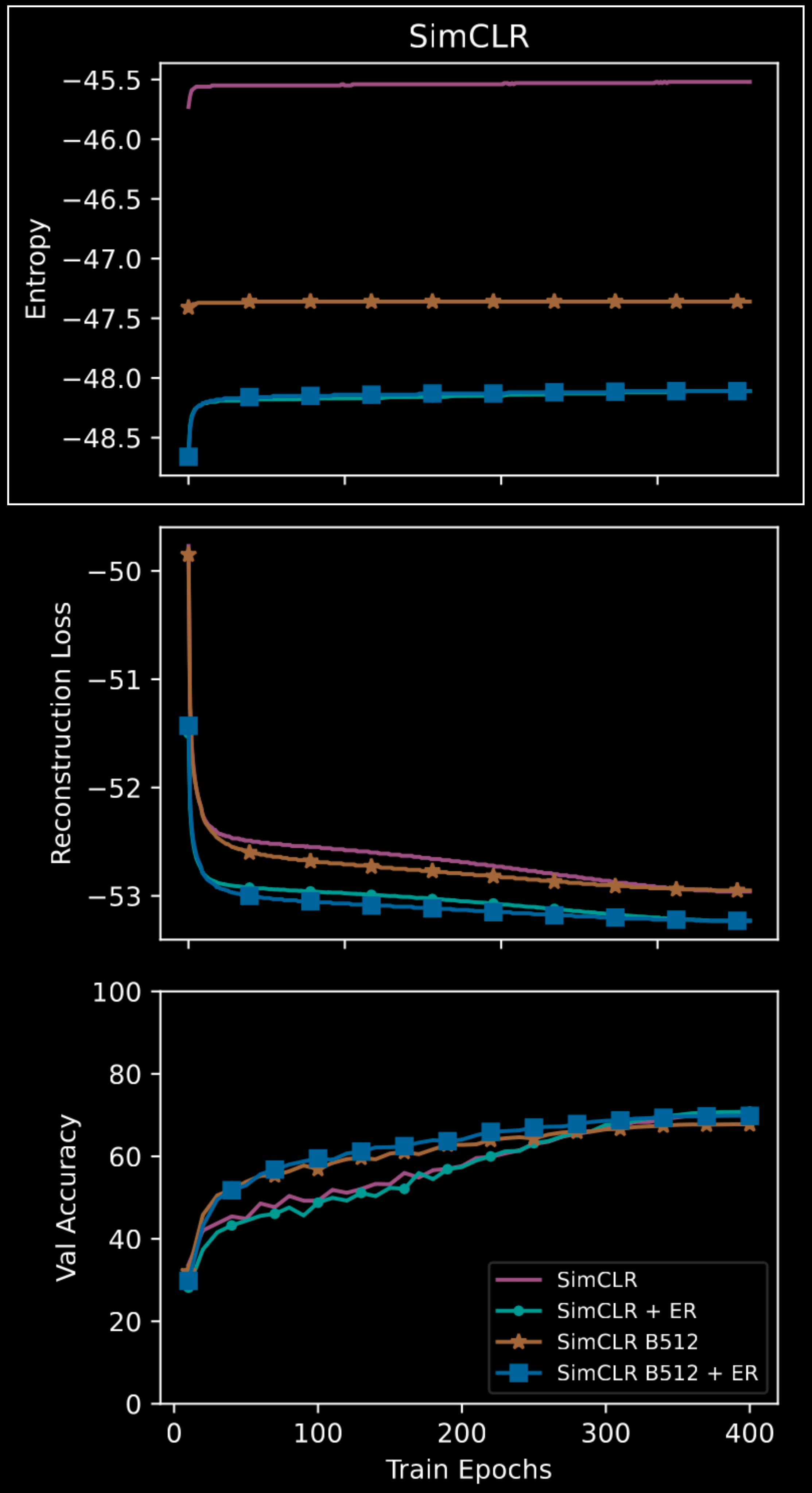
Maximize reconstruction, but the entropy is only maintained stable

Empirical results: We can add entropy optimization explicitly

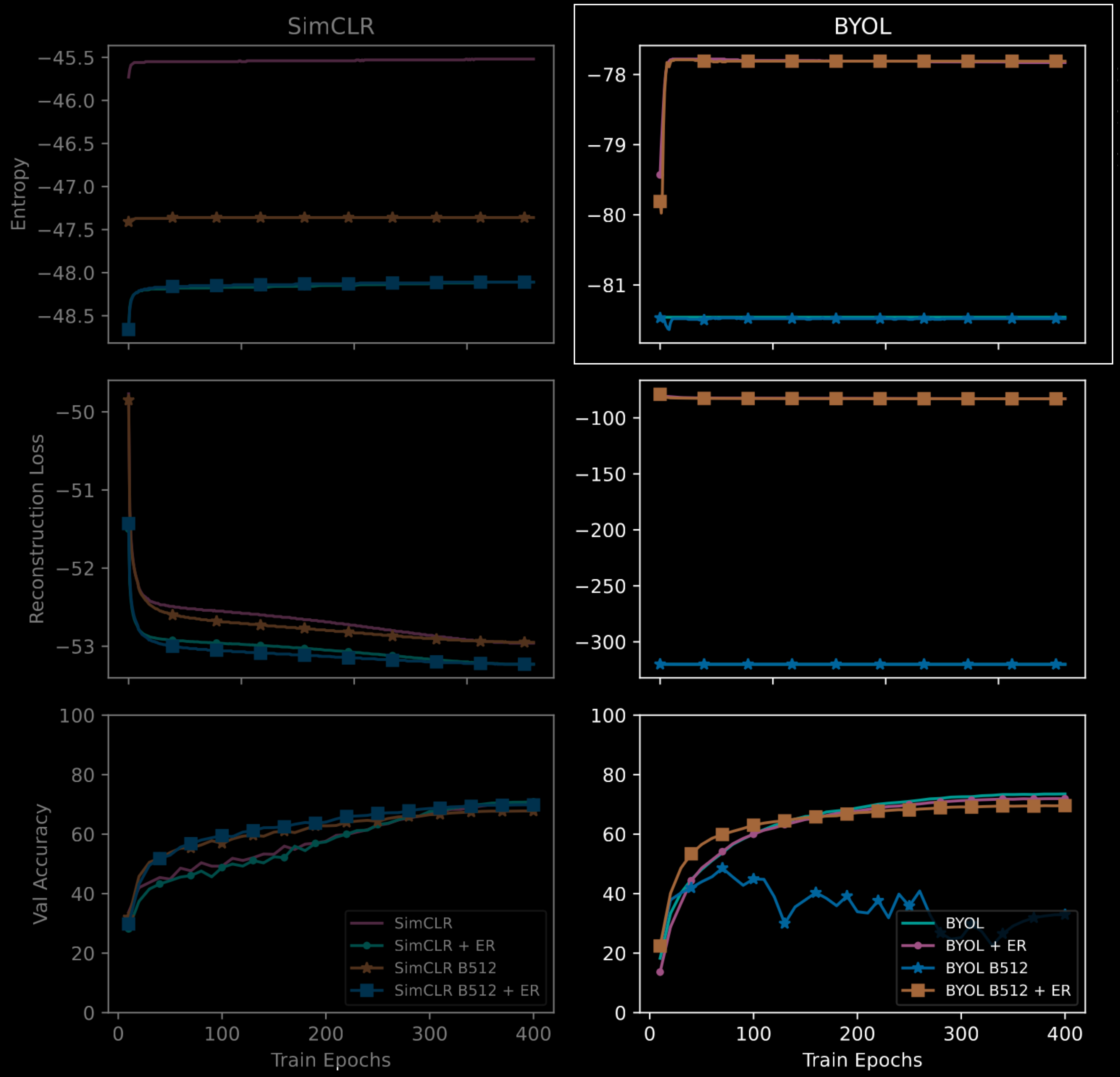
Empirical results: We can add entropy optimization explicitly



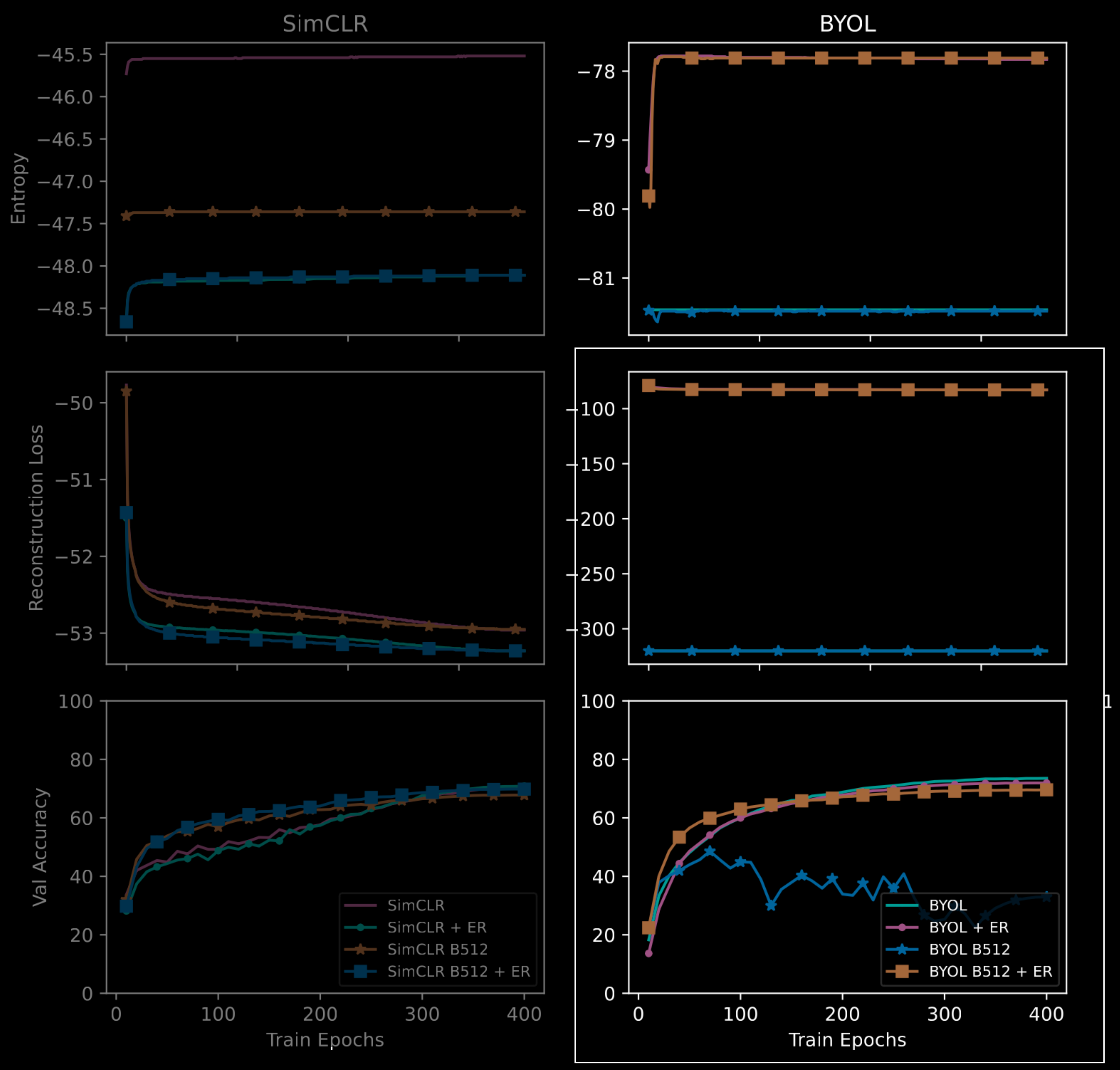
Empirical results: We can add entropy optimization explicitly



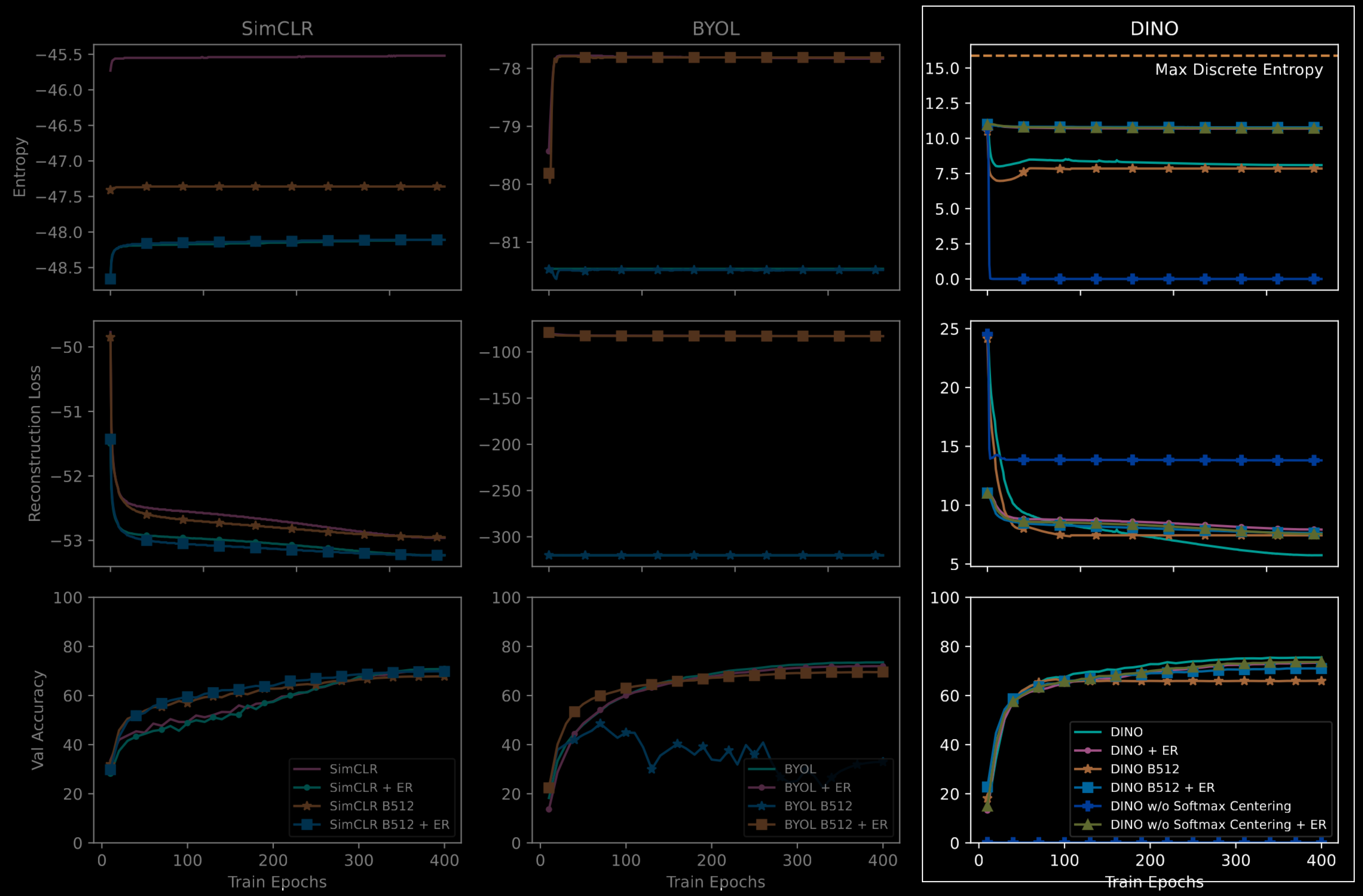
Empirical results: We can add entropy optimization explicitly



Empirical results: We can add entropy optimization explicitly



Empirical results: We can add entropy optimization explicitly



Empirical results: We can add entropy optimization explicitly