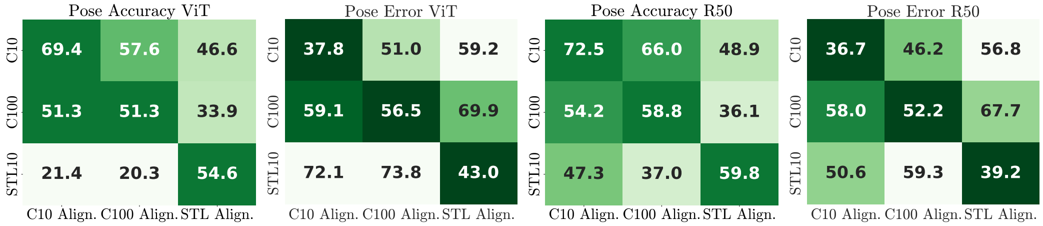
ImageNet	ResNet50 (Vanilla-Trained)		ViT (Vanilla-Trained)		CLIP (Vanilla-Trained)	
Canon.	**Acc**	**Rand Rot. (C8)**	**Acc**	**Rand Rot. (C8)**	**Acc**	**Rand Rot. (C8)**
None	75.2	50.1	80.4	59.6	77.1	67.0
PRLC*	63.1	59.2	63.7	60.5	72.1	69.6
Ours	**66.3** (+3.2)	**63.5** (+4.3)	**73.6** (+9.9)	**71.9** (+11.4)	**75.4** (+3.3)	**74.0** (+4.4)
Oracle	*75.2*	*71.5*	*80.4*	*78.1*	*77.1*	*75.3*

PRLC R50 Aligner			PRLC ViT Aligner			Ours		
Acc (↑)	**Acc @ 45° (↑)**	**Err (↓)**	**Acc**	**Acc @ 45°**	**Err**	**Acc**	**Acc @ 45°**	**Err**
37.9	55.4	63.1	31.8	56.4	65.6	37.0 (-0.9)	**78.9** (+22.5)	**45.3** (-17.8)

Table 3: **FMC generalizes better to ImageNet and outperforms PRLC's canonicalizers.** We find that Foundation Model Canonicalization outperforms PRLC, without any training, on both upright inputs and randomly rotated inputs. We compare against just upright images in the Acc columns. Oracle refers to a system where the exact angle to upright is known, and thus only measures the change in accuracy due to loss of information due to rotating, cropping, and re-rotating. Rand Rot. ($C8$) applies a random $C8$ transform to the input before passing it to the aligner / model. Best non-oracle rows on rotated performance are bolded. For PRLC, the canonicalizers were the best performing ones from other datasets (STL10 for both ResNet50 and ViT). Thus, they were not trained specifically for ImageNet.

6 Additional Results and Figures



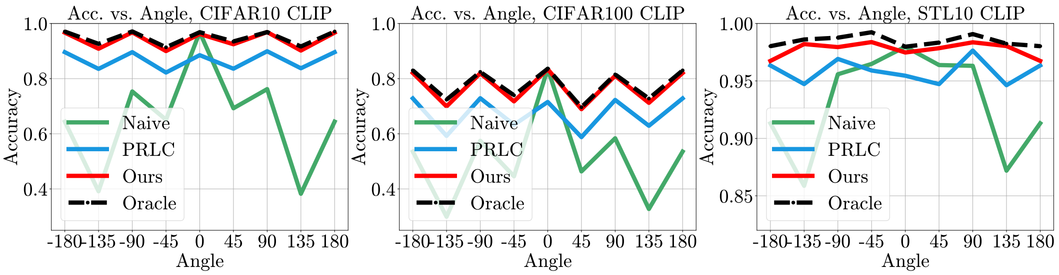
(a) ViT Accuracy (↑) (b) ViT Error (↓) (c) R50 Accuracy (↑) (d) R50 Error (↓)

Figure 7: **FMC generalizes better across datasets when mixing up aligners and downstream models**. PRLC performance on pose estimation drops significantly when using a canonicalizer trained from a different dataset compared to FMC, which applies one technique across all settings. This result highlights the generalizability across datasets of an unsupervised approach.

A Experimental Setup

A.1 Experimental Setup - 3D

For 3D, we first look at the CO3D (Reizenstein et al., 2021) dataset to measure how our FMC's energy function correlates to 3D viewpoint quality. We compare the ranking of viewpoint frames by FMC energy compared to that of the probability of the ground truth label. We then look at Objaverse-LVIS (Deitke et al., 2022) to measure the effect of combining FMC with Zero123 (Liu et al., 2023) as the transformation generation function to simulate new 3D viewpoints from a single image. For Zero123 experiments, we rank the viewpoints by the probability of the ground truth label and then measure the difference in accuracy between the original Objaverse renders and that of the energy minimizing Zero123 render for the respective ranking bins.

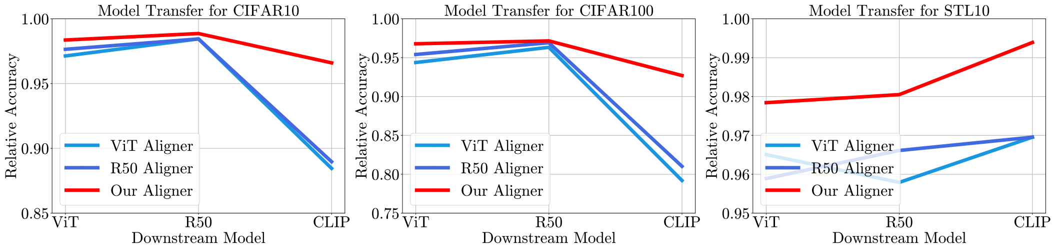


(a) .

(b) .

(c) .

Figure 8: Accuracy vs. $C8$ angle on CLIP. Like on ResNet50, we find that using FMC leads to invariant predictions over angles, outperforming PRLC. The contrast is particularly clear for CLIP on CIFAR10 and CIFAR100, where our accuracy over angle is consistently above PRLC.



(a) CIFAR10.

(b) CIFAR100.

(c) STL10.

Figure 9: Transferring canonicalizers across models. We measure the effects of transferring canonicalizers across different downstream models by plotting the relative accuracy over the naive model. We find that FMC outperforms the PRLC aligners, particularly when transferring to CLIP on CIFAR10 and CIFAR100. These results show the ability of FMC to generalize across downstream models. All ViT and R50 models and aligners are PRLC-trained versions.

The model we use for both experiments is the fine-tuned version of CLIP from OVSEG (Liang et al., 2023) which is designed to work on background removed images, as Zero123 operates on such images. Please see the Appendix for more details on experimental setup.

A.2 Experimental Setup – Color

We define the color shift transformation using the popular von Kries model (KRIES, 1905) where an illuminant vector with the RGB values $L = [\mathrm{L_R}, \mathrm{L_G}, \mathrm{L_B}] \in \mathbb{R}^3$ is multiplied element-wise with every pixel in the image. We then generate this illuminant vector L by sampling in the log-chrominance space (Barron & Tsai, 2017). Specifically,

$$L_u, L_v \sim U[-1, 1] \tag{6}$$

$$[L_R, L_G, L_B] = [\frac{\exp(-L_u)}{z}, \frac{1}{z}, \frac{\exp(-L_v)}{z}] \tag{7}$$

where $z = \sqrt{\exp(-L_u)^2 + \exp(-L_v^2) + 1}$ is a normalizing constant and L_u, L_v are the log-chroma values sampled from the uniform distribution with range $[-1, 1]$. Intuitively, the log-chroma space defines the R/G and B/G ratios in log-space. A range of $[-1, 1]$ corresponds roughly to a 7× change in the ratio between the minimum and maximum points of the range.

A.3 Hyperparameters for the energy functions

All hyperparameters were found using Bayesian Optimization with the same kernel and acquisition function mentioned in Section 4 and performed using the Bayesian Optimization Toolbox (Nogueira, 2014) for 300 time steps. Each energy hyperparameter was tuned on a small training or validation set by recording logits and finding the combination of energy functions that maximized accuracy.

For experiments on ImageNet, CIFAR10, CIFAR100, and STL10, we only used the classification energy for computational efficiency. This setting can be reduced to a single free parameter, which we denote α_{logit}. The coefficient for mean logit is thus α_{logit} and the coefficient for max logit is $(1 - \alpha_{\text{logit}})$.

Specifically, the α_{logit} coefficients we found were: 0.59 for CIFAR10, CIFAR100, and ImageNet, 0.73 for STL10,and 0.64 for our method applied with PRLC's classifiers. For segmentation, the α_{logit} is 0.74 with a diffusion energy factor of 0.94 and a segmentation energy factor of 1.12. For ImageNet, we also found it helpful to include the mean of top-5 logits with a factor of 0.08.

For diffusion energy, we subsample the time steps to range from 500 to 1000 with a stride of 20. This is primarily for computational efficiency.

A.4 3D

For our CO3D experiments, we take 10 random videos from each class, and sample 50 random frames from each video. We crop and preprocess the view following the pipeline in (Liu et al., 2023). Like Appendix A.3, we tune the energy hyperparameter with Bayesian Optimization using a 10% subset of the data. The metric optimized is the difference in mean accuracy of the best five ranks and the worst five ranks. We sort the frames by energy and bin them by their respective video frame ranks, and then compute the accuracy of each rank over the videos. The α_{logit} coefficient found was 1.31.

For Objaverse-LVIS (Deitke et al., 2022), we render 400 objects at 36 views, corresponding to the upper hemisphere of azimuth and elevation angles at an interval of 30 degrees. This generates the test set of images with different viewpoints to evaluate on. Then, to evaluate FMC, we start at each Objaverse render, simulate Zero123 (Liu et al., 2023) generated images at azimuth circles of an interval of 30 degrees at elevation angles of [-60, -30, 0, 30, 60], taking the minimum energy (best) Zero123 generation as the canonical form. Like for CO3D, we rank the frames, sort and bin them, and compute the accuracy for each rank. However, to isolate the effects of Zero123 on FMC, we rank both the baseline and FMC curves by the probability of the ground truth mask as a proxy for the true ranking of viewpoint. We use the same Bayesian Optimization setting as CO3D. The α_{logit} coefficient found was 0.79.