A. Techniques from Representation Learning

The task of Backward Compatible Representation Learning exploits techniques from the field of representation learning [48, 4, 19, 3, 16, 18], where classification [22, 39, 45, 38, 40], metric learning [25, 41, 44, 34], and contrastive learning [7, 14, 13] are some major methods. For simplicity and better alignment with previous works in backward compatible representation learning [6, 24, 47, 29], we adopt the classification loss for training the representation model.

B. Cosine Similarity vs Euclidean Distance

We note that some of the previous works on representation learning and backward compatibility use the Euclidean Distance for retrieval [31, 29] while others use Cosine Similarity [24, 42, 12, 25, 27]. Preliminary experiments that we conducted did not find any clear superiority between the metrics when compared with public results in [31, 29]. We adopt cosine similarity which provides for clearer analysis and better compatibility with our experiments on multi-modality.

C. Proof of Lemmas

For completeness, we provide proofs to the lemmas stated in the main text.

Proof of Lemma 1. We define an image space $\mathcal{X}$, an n-dimensional representation space $\mathbb{R}^n$, and two representation functions $\phi_{old}, \phi_{new} : \mathcal{X} -> \mathbb{R}^n$ that maps images to a unit ball in $\mathbb{R}^n$. We consider the distance metric d being the negative cosine similarity, and $\forall x, \|\phi_{old}(x)\|_2 = \|\phi_{new}(x)\|_2 = 1$

To construct this counterexample, for two images in the gallery, x_1 and x_2 of the same class y where the representations of x_1, x_2 are $\phi_{old}(x_1), \phi_{old}(x_1)$. The third image as a query, $\bar{x}$ of the same class y, has its old representation and new representation as $\phi_{old}(\bar{x}), \phi_{new}(\bar{x})$. We consider a specific case where $\phi_{old}(\bar{x})$ is close to the cone spanned by $\phi_{old}(x_1), \phi_{old}(x_2)$ defined by by:

$$\phi_{old}(\bar{x}) = a\phi_{old}(x_1) + b\phi_{old}(x_2) + \epsilon,$$
$$\text{for some } a, b > 0, \text{ and small } \epsilon$$

Let the projection of $\phi_{old}(\bar{x})$ to the plane of $\phi_{old}(x_1), \phi_{old}(x_2)$ be $P(\phi_{old}(\bar{x}))$. Let the angle between $P(\phi_{old}(\bar{x}))$ and $\phi_{old}(x_1), \phi_{old}(x_2)$ be θ_1, θ_2, and let the angle between $P(\phi_{old}(\bar{x}))$ and $\phi_{old}(\bar{x})$ be $\delta\theta$ so that $\sin\delta\theta = \epsilon$. Similarly, let the projection of $\phi_{new}(\bar{x})$ be $P(\phi_{new}(\bar{x}))$, whose angle with $\phi_{old}(x_1), \phi_{old}(x_2)$ be θ_3, θ_4, and its angle with $\phi_{new}(\bar{x})$ be $\delta\theta'$. $\theta_1, \theta_2, \theta_3, \theta_4 \in [0, \pi], \delta\theta, \delta\theta' \in [0, \frac{\pi}{2}]$.

By the criterion for backward compatibility defined in Definition 1, we have:

$$\begin{aligned} d(\phi_{new}(\bar{x}), \phi_{old}(x_1)) &\leq d(\phi_{old}(\bar{x}), \phi_{old}(x_1)) \\ d(\phi_{new}(\bar{x}), \phi_{old}(x_2)) &\leq d(\phi_{old}(\bar{x}), \phi_{old}(x_2)), \end{aligned} \quad (1)$$

which gives us

$$\begin{aligned} \cos\theta_3 \cos\delta\theta' &\geq \cos\theta_1 \cos\delta\theta \\ \cos\theta_4 \cos\delta\theta' &\geq \cos\theta_2 \cos\delta\theta \end{aligned}$$

To bound $\delta\theta'$, we first notice that $\theta_1 + \theta_2 = \theta(\phi_{old}(x_1), \phi_{old}(x_2)) \leq \pi$, with $\theta(\phi_{old}(x_1), \phi_{old}(x_2))$ being the angle between $\phi_{old}(x_1), \phi_{old}(x_2)$, because $\phi_{old}(\bar{x})$ lies in the cone. Because of the constraint in Equation 1, $\phi_{new}(\bar{x})$ must also lie in the cone. Therefore, $\theta_1 + \theta_2 = \theta_3 + \theta_4 = \theta(\phi_{old}(x_1), \phi_{old}(x_2)) \leq \pi$, which yields

$$\begin{aligned} (\theta_1 - \theta_3)(\theta_2 - \theta_4) &\leq 0 \\ (\cos\theta_1 - \cos\theta_3)(\cos\theta_2 - \cos\theta_4) &\leq 0. \end{aligned} \quad (2)$$

Comparing Equation 1 and Equation 2, we conclude that $\delta\theta' \leq \delta\theta$, so that $\cos\delta\theta' \geq \cos\delta\theta$.

To further bound $\cos(\theta_1 - \theta_3)$, by inspecting Equation 1, in the case of $\theta_1 < \theta_3$ we have:

$$\begin{aligned} \cos\theta_3 \cos\delta\theta' &\geq \cos\theta_1 \cos\delta\theta \\ \cos\theta_3 &\geq \cos\theta_1 \frac{\cos\delta\theta}{\cos\delta\theta'} \\ \cos\theta_3 &\geq \cos\theta_1 \cos\delta\theta \\ \cos\theta_3 &\geq \cos\theta_1 \sqrt{1-\epsilon^2} \\ \cos\theta_1 - \cos\theta_3 &\leq \frac{1-\sqrt{1-\epsilon^2}}{\sqrt{1-\epsilon^2}}, \end{aligned} \quad (3)$$

where the third inequality follows by upperbounding $\cos\delta\theta'$ to be 1, the fourth inequality by substituting $\delta\theta$ with ϵ, the fifth inequality follows by upperbounding $\cos\theta_3$ to be 1. By further expanding Equation 3,

$$\begin{aligned} \cos\theta_1 - \cos\theta_3 &\leq \frac{1-\sqrt{1-\epsilon^2}}{\sqrt{1-\epsilon^2}} \\ 2\sin\frac{\theta_1+\theta_3}{2}\sin\frac{\theta_3-\theta_1}{2} &\leq \frac{1-\sqrt{1-\epsilon^2}}{\sqrt{1-\epsilon^2}} \\ \sin^2\frac{\theta_3-\theta_1}{2} &\leq \frac{1-\sqrt{1-\epsilon^2}}{2\sqrt{1-\epsilon^2}} \\ \cos^2\frac{\theta_3-\theta_1}{2} &\geq 1 - \frac{1-\sqrt{1-\epsilon^2}}{2\sqrt{1-\epsilon^2}} \\ \cos(\theta_3-\theta_1) &\geq 1 - \frac{1-\sqrt{1-\epsilon^2}}{\sqrt{1-\epsilon^2}}, \end{aligned} \quad (4)$$

where the third inequality follows from lowerbounding $\sin\frac{\theta_1+\theta_3}{2}$ by $\sin\frac{\theta_3-\theta_1}{2}$.

Similarly, in the case of $\theta_1 \geq \theta_3$, we have $\theta_2 \leq \theta_4$, so that $\cos(\theta_3-\theta_1) = \cos(\theta_4-\theta_2) \geq 1 - \frac{1-\sqrt{1-\epsilon^2}}{\sqrt{1-\epsilon^2}}$. Therefore, we have in all cases, $\cos(\theta_3-\theta_1) \geq 1 - \frac{1-\sqrt{1-\epsilon^2}}{\sqrt{1-\epsilon^2}}$.

With both $\cos(\theta_3 - \theta_1)$ and $\cos\delta\theta'$, we have the cosine similarity between $\phi_{old}(\bar{x})$ and $\phi_{new}(\bar{x})$, $\cos(\phi_{old}(\bar{x}), \phi_{new}(\bar{x}))$ being bounded by

$$\begin{aligned}
&\cos(\phi_{old}(\bar{x}), \phi_{new}(\bar{x})) \\
\geq &\cos(\phi_{old}(\bar{x}), P(\phi_{old}(\bar{x}))) \cos(\phi_{new}(\bar{x}), P(\phi_{old}(\bar{x}))) \\
\geq &\sqrt{1-\epsilon^2} \cos(P(\phi_{new}(\bar{x})), \phi_{new}(\bar{x})) \\
&\times \cos(P(\phi_{new}(\bar{x})), P(\phi_{old}(\bar{x}))) \\
\geq &(1-\epsilon^2)(1 - \frac{1-\sqrt{1-\epsilon^2}}{\sqrt{1-\epsilon^2}})
\end{aligned}$$

Therefore, we show that in order to be backward compatible with $\phi_{old}(x_1), \phi_{old}(x_2), \phi_{new}(\bar{x})$ is restricted within a small angle from $\phi_{old}(\bar{x})$, with $\cos(\phi_{old}(\bar{x}), \phi_{new}(\bar{x})) \geq (1-\epsilon^2)(1 - \frac{1-\sqrt{1-\epsilon^2}}{\sqrt{1-\epsilon^2}})$. This limits the room of improvement of $\phi_{new}(\bar{x})$ over $\phi_{old}(\bar{x})$, especially when $\phi_{old}(\bar{x})$ is not good.

Proof of Lemma 2 can be found in [11].

Proof of Lemma 3. For any orthonormal matrix P, and representation function ϕ, any images x_1, x_2, we have

$$\begin{aligned}
&(P(\phi(x_1)))^\top P(\phi(x_2)) \\
=&\phi(x_1)^\top P^\top P \phi(x_2) \\
=&\phi(x_1)^\top (P^\top P) \phi(x_2) \\
=&\phi(x_1)^\top \phi(x_2)
\end{aligned}$$

□

D. Sample Captions for Imagenet-1k

We did not find existing dataset that simultaneously supports both evaluation of image-to-image retrieval representations and image-to-text representations. To our purpose of modality fusion, we generate automatic captions for Imagenet-1k with "vit-gpt2-image-captioning" from [43]. We provide sample captions generated in Figure 3. We observe that although automatic image captions capture daily pictures like dogs and benches well, it does not recognize other less common pictures like wild animals and pills. This is an expected behavior because automatic image captioning models might have encountered more daily pictures during the training than less common ones. Learning under such strong noise pose a significant challenge to the robustness of different methods, and it also causes the evaluation of text-to-image retrieval accuracy lower than it should be.

□

E. Confidence Intervals

Because of limited computational resources, we are unable to provide confidence intervals for all of our experiments. To get a sense of the variances of the experiments,

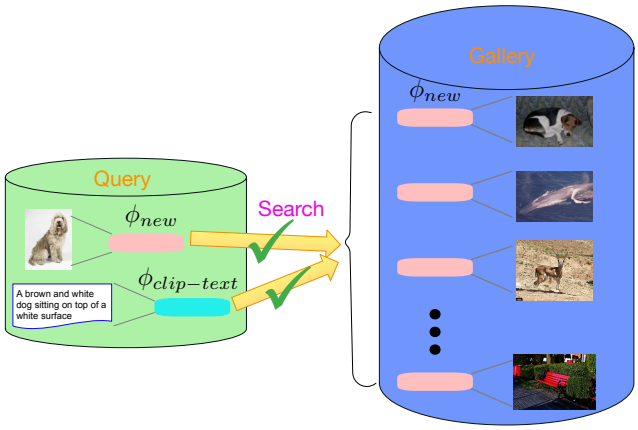


Figure 2. An illustration of the idea of modality fusion. A gallery of images is encoded with a single representation ϕ_{new} but can support query with images encoded by ϕ_{new} and text encoded by $\phi_{clip-text}$ at the same time.

we conduct backward compatible experiments on a subset of Imagenet-1k with 50k images (50 images from each class). We use ResNet50-128 model architecture for both the old model and the new model. Old models are trained using 500 classes of our constructed Imagenet-1k subset while the new models have access to the entire 1000 classes. The independent models (ϕ_{old} and ϕ'_{new}) are only trained once, but we calculate means and standard deviations over 5 random seeds of training the new model.

As shown in Table 8, we found that the standard deviations for both BCT and BT^2 are relatively small with respect to all the metrics (below 0.5%), and the advantage of BT^2 over BCT is indeed statistically significant in both ϕ_{new}/ϕ_{new} and ϕ_{new}/ϕ_{old}. For example, in terms of ϕ_{new}/ϕ_{new} Top-1 accuracy, BT^2 achieves 21.4% while BCT achieves 18.3%. This gain of 3.1% is statistically significant considering the standard deviations of the results are only 0.3% and 0.4% respectively. We hope this supplementary experiment can provide a rough idea of the degree of randomness in our backward compatible experiments.

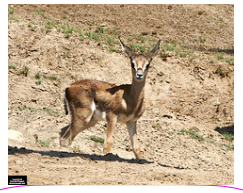
Method	Case	Top1-Top5	mAP
Independent	ϕ_{old}/ϕ_{old}	10.3-25.0	6.2
	ϕ'_{new}/ϕ'_{new}	17.8-37.5	10.5
BCT	$\phi^{bct}_{new}/\phi^{bct}_{old}$	11.5 ± 0.1-29.3 ± 0.3	7.6 ± 0.1
	$\phi^{bct}_{new}/\phi^{bct}_{new}$	18.3 ± 0.4-38.7 ± 0.5	12.7 ± 0.1
BT^2	$\phi^{bt^2}_{new}/\phi^{bt^2}_{old}$	12.6 ± 0.2-31.0 ± 0.3	8.0 ± 0.0
	$\phi^{bt^2}_{new}/\phi^{bt^2}_{new}$	21.4 ± 0.3-42.6 ± 0.3	14.6 ± 0.1

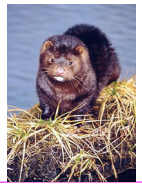
Table 8. Backward compatible experiments on Imagenet-500 to Imagenet-1k (a 50k images subset) with only data change. Both the old model and the new model uses Resnet50-128 architecure.


A brown and white dog laying on top of a couch


A large brown and white dog standing in a field


A large white boat floating on top of a body of water


A lone zebra walking on a dirt road


A brown bear sitting on top of a pile of logs


A brown and white dog sitting on top of a white surface


A white refrigerator filled with lots of food


A red bench sitting in the middle of a park


A bird that is standing on some grass


A large brown and white polar bear sitting on a rock


A black and white photo of a black and white cat


A small bird sitting on top of a pile of leaves


A train crossing a bridge over a river


A small cabin in the middle of a snow covered field


A bed that has a blanket over it


A toy model of a person on a skateboard

Figure 3. Sample captions automatically generated for Imagenet-1k with “vit-gpt2-image-captioning” from [43].