We thank all reviewers for their valuable comments, such as the novelty, well-motivated objective and promising results. The code "ICP-pytorch" has been anonymously released on GitHub. We rebut key issues point-by-point as below.

Response to R#2:

We sincerely hope R#2 to raise the score. R#2 mainly criticized the missing of baseline and ablation study (indeed we did have both), and misunderstood our flowchart to multi-view learning (indeed our flowchart has essential difference to multi-view learning). We rebut these concerns below, and will clarify related issues in our paper.

Q1: Similar idea with multi-view learning. We rebut, with respect, that our flowchart is totally different from multi-view learning. Our idea is to *distill* diverse representations with different constraints, while multi-view learning focuses on the fusion of *predefined* features. Clearly, they reverse in procedures and diverse in mechanisms. To the best of our knowledge, our work is the first attempt to explicitly model diversity in deep representation learning.

Q2: Necessity of information bottleneck. The information bottleneck objective (*i.e.*, Term ② in Eq.3r) is essential for learning diversified representations as an information minimizing part, which results in representations that carry general information. The necessity is also supported by the ablation study in Tab.1. To clarify, we reorder the objective in Eq.3 as Eq.3r below. Term ① cannot be removed for ablation study solely, as it is the final target of our objective which uses the constrained representation parts to complete the downstream task. Term ② (*i.e.*, information bottleneck objective) and Term ③ (*i.e.*, information maximization objective) are different constraints for part z and part y, which force these parts to contain as more/less as possible information for the target task. $\mathcal{I}(z,t)$ and $\mathcal{I}(y,t)$ are used to prevent any one of these features from dominating the task, and thus reduce the diversity of representations. Term ④ is used to make z and y independent of each other. For qualitative evidence, please refer to our responses to Q3&Q4.

$$\max\Big[\ \underbrace{\mathcal{I}(r,t)}_{\text{① Inference}}\ +\underbrace{\mathcal{I}(z,t)-\beta\mathcal{I}(z,x)}_{\text{② Minimize Information}}+\underbrace{\mathcal{I}(y,t)+\alpha\mathcal{I}(y,x)}_{\text{③ Maximize Information}}-\ \underbrace{\gamma\mathcal{I}(z,y)}_{\text{④ Independent}}\ \Big]. \tag{3r}$$

Q3: The missing baseline. Incorrect! Indeed, we did implement the mentioned baseline in Tab.1. Specifically, ICP-ALL refers to the optimization using only Term ① of Eq.3r, which ensembles two models with different initializations as introduced in L186. Overfitting occurs due to the large model capacity without constraints, making the performance of this strategy sub-optimal for learning discriminative representations as shown in Tab.1.

Q4: The lack of ablation study. Incorrect! Indeed, we did provide necessary ablation studies in Tab.1. Specifically, ICP-COM refers to optimization without $\mathcal{I}(z,t)$, $\mathcal{I}(y,t)$ and $\mathcal{I}(z,y)$, which means optimization without the competing process. VIB refers to optimization that removes Term ③ of Eq.3r (*i.e.*, without information maximization constraint), and DIM* refers to optimization that removes Term ② of Eq.3r (*i.e.*, without information bottleneck constraint), all these achieve sub-optimal results due to the limited diversification of representations. VIB$_{\times 2}$/DIM*$_{\times 2}$ ensembles two feature parts with the same constraints. These ablation studies show that expanding models with sole constraint or removing one constraint from the objective decreases the performance.

Response to R#1:

Q1: Meaning of diversified representations. We apologize for the unclear definition in L23-27. Diversified representation learns representations with different mutual information constraints, which results in more powerful representation. In contrast, disentangled representation, as a special case of our diversified representation, lacks diversity due to the single mutual information minimization constraint.

Q2: Confusions of Eq.8 and discriminator. As KL divergence (Eq.7) has no upper bound, we use JS divergence (Eq.8) as a substitution (not the bound). In Eq.8, the discriminator D is used to maximize the JS divergence. D only distinguishes the positive pair (x, y) and negative pair $(x, \hat{y})$, which is different from that of GAN.

Q3: Explanation of Eq.3. We reorder Eq.3 as Eq.3r. Term ① aims to fuse the feature parts for downstream tasks. Term ② and Term ③ are different constraints together with Term ④ for diversifying the feature parts. All terms are important for learning diversified representations and the ablation studies in Tab.1 support our claim.

Q4: Implementation details and error bars. We set small hyper-parameters (lr:0.01, γ:0.1, α:0.1, β:0.01 in classification; lr:1e-4, γ:1, α:5, β:5 in reconstruction) to prevent gradient exploding. The error bars are listed below and ICP is competitive. We will add these into our paper. If needed, please refer to our anonymous source code.

CIFAR10	VGG16	GoogLeNet	ResNet20	DenseNet40	**CIFAR100**	VGG16	GoogLeNet	ResNet20	DenseNet40
Baseline	6.67	4.92	7.63	5.83	Baseline	26.41	20.68	31.91	27.55
ICP	6.20 ± 0.06	4.48 ± 0.11	6.17 ± 0.04	5.27 ± 0.13	ICP	24.85 ± 0.19	19.06 ± 0.13	28.48 ± 0.15	25.48 ± 0.22

Response to R#3:

Q1: Quantitative evaluation for disentanglement. Following this suggestion, we have evaluated the MIG score proposed in β-TCVAE with the same settings on dSprites and 3D Faces. The results of β-VAE are 0.21 and 0.47. ICP are 0.22 and 0.49. We will add these results into our paper.

Q2: More thorough explanation for the ablations. We will add more in-depth analysis for the ablation studies in Tab. 1. Due to the page limit, we cannot itemize them in rebuttal. In general, expanding models with sole constraint or removing one constraint from our objective deteriorates the performance.

Q3: Interpretability of deep representations. Thanks for this inspiring suggestion. The interpretability of representations learned by neural networks remains an open problem. Our experiment (Fig.2) provides insights about the interpretability of learned representations. It shows that a small number of dimensions are sufficient for inference if the representations contain general information, while a large number of dimensions are required for inference if the representations contain specific information.