Summary of the Appendix

This appendix contains additional details for the paper "*Federated Continuous Category Discovery and Learning*", including implementation details, additional experiment results, discussion of limitations, and broader impact. The implementation code can be found in the attached supplementary materials. This appendix is organized as follows:

- Section A introduces the data splitting strategy (Section A.1), details of the baseline implementation (Section A.2), and detailed settings of the ablation study (Section A.3).
- Section B provides additional results for 1) experiments in the centralized setting (Section B.1); 2) experiments with various degrees of heterogeneity in **FC^{2}DL** (Section B.2); 3) experiments with various data partitions (Section B.3); 4) sensitivity analysis of hyper-parameters (Section B.4); 5) experiments with a large number of participants (Section B.5); 6) experiments of launching data reconstruction attacks for verifying **GPA**'s characteristic of privacy-preserving (Section B.6); 7) results of performance forgetting on known categories (Section B.7); and 8) experiments on pre-trained vision transformers (Section B.8).
- Section C presents the detailed optimization pipeline of **GPA**.
- Section D discusses the limitations of the proposed methods and the possible future directions.
- Section E illustrates the potential broader impact of this work in the real world.

A. Implementation Details

A.1. Data splitting settings

CIFAR-100 contains two separate sets – one is a training set consisting of 500 samples for each class, and the other is used for testing with 100 samples for each class. For Tiny-ImageNet and ImageNet-S, we follow FCIL [10] to select the ending 50 and 100 samples per class for testing, respectively, while the rest 500 samples are used for training. As for the usage of CUB200, StanfordCars, and Herbarium 19, we also follow their default data splitting to prepare the training and testing sets. To simulate the practical non-IID setting, the training samples of each participant are drawn independently with class labels following a categorical distribution over $C = C^{\mathrm{L}} + C^{\mathrm{U}}$ classes, which can be parameterized by a vector $\mathbf{q}$ ($q_i \geq 0, i \in [1, C]$ and $\|\mathbf{q}\|_1 = 1$). And we draw $\mathbf{q} \sim \mathrm{Dir}(\alpha, \mathbf{p})$ from a Dirichlet Distribution [19], where $\mathbf{p}$ is a prior class distribution over C classes and $\alpha > 0$ controls the data heterogeneity among FL participants. A smaller α leads to more heterogeneous data distributions among participants, and we set $\alpha = 0.1$ for all **FC^{2}DL** experiments in the main paper, while the centralized training setting can be regarded as the case when $\alpha \to +\infty$.

Table 5. Performance comparison between **GPA** and other baselines in **FC^{2}DL** with two fine-grained dataset. Only one novel category learning stage is used and the novel category numbers are 20 and 83 for StanfordCars and Herbarium 19, respectively.

Method	StanfordCars			Herbarium 19		
	known	novel	all	known	novel	all
AutoNovel	$45.0_{\pm 0}$	$18.2_{\pm 4.2}$	$42.2_{\pm 1.1}$	$49.2_{\pm 0}$	$21.7_{\pm 1.1}$	$45.4_{\pm 0.2}$
GM	$45.0_{\pm 0}$	$12.5_{\pm 3.1}$	$41.6_{\pm 1.7}$	$49.2_{\pm 0}$	$22.7_{\pm 1.5}$	$45.5_{\pm 0.9}$
iNCD	$43.1_{\pm 1.7}$	$19.0_{\pm 1.5}$	$42.0_{\pm 1.0}$	$48.5_{\pm 0.7}$	$22.4_{\pm 0.9}$	$44.8_{\pm 0}$
Happy	$44.7_{\pm 1.0}$	$24.9_{\pm 0.7}$	$42.6_{\pm 1.1}$	$49.2_{\pm 0}$	$25.5_{\pm 2.1}$	$45.9_{\pm 0.4}$
IIC	$33.3_{\pm 4.7}$	$22.6_{\pm 2.2}$	$32.2_{\pm 2.0}$	$49.2_{\pm 0}$	$17.8_{\pm 2.5}$	$44.9_{\pm 1.6}$
OpenCon	$34.7_{\pm 2.9}$	$14.3_{\pm 7.0}$	$31.5_{\pm 3.7}$	$35.5_{\pm 4.4}$	$13.4_{\pm 8.9}$	$33.2_{\pm 3.9}$
Orchestra	$44.5_{\pm 0.5}$	$18.3_{\pm 2.0}$	$42.3_{\pm 0.4}$	$48.0_{\pm 1.3}$	$22.0_{\pm 0.9}$	$44.4_{\pm 0.5}$
SemiFL	$35.5_{\pm 0.2}$	$19.7_{\pm 1.0}$	$34.4_{\pm 0.4}$	$42.2_{\pm 0.5}$	$23.0_{\pm 0.4}$	$39.5_{\pm 0.2}$
FedoSSL	$42.5_{\pm 0}$	$18.3_{\pm 1.0}$	$40.8_{\pm 0.5}$	$46.2_{\pm 1.1}$	$15.6_{\pm 0.9}$	$41.9_{\pm 0.6}$
AGCL	$40.7_{\pm 0.2}$	$19.3_{\pm 1.0}$	$39.2_{\pm 0.5}$	$46.0_{\pm 0.5}$	$24.1_{\pm 0.2}$	$42.9_{\pm 0.2}$
GPA	$\mathbf{45.0}_{\pm 0}$	$\mathbf{31.5}_{\pm 0.4}$	$\mathbf{43.8}_{\pm 0.1}$	$\mathbf{49.3}_{\pm 0.2}$	$\mathbf{33.7}_{\pm 1.2}$	$\mathbf{47.1}_{\pm 1.0}$

A.2. Baseline methods

To our knowledge, no direct baseline method can be used to compare with our proposed methods in the setting of **FC^{2}DL**. In this case, for a fair comparison, we try our best to integrate certain state-of-the-art baseline methods from related areas, like standard novel category learning and federated self-supervised learning, into FL to solve **FC^{2}DL**. In addition to standard FedAvg [37], we also implement other mainstream FL algorithms including FedProx [31], SCAFFOLD [23], and Moon [29], to see whether our **GPA** can empower them the capability of NCDL. Next, we will provide the details of implementing these state-of-the-art baselines.

First of all, we assume that all baseline methods rely on the same m^{L} trained by **GPA** at the beginning of novel category learning. AutoNovel [16] assumes that both the known and the novel-category data are accessible during novel category learning, and it has separate design and training loss for known and novel data. Considering the unavailability of the known-category data in **FC^{2}DL**, we only apply its pairwise Binary CrossEntropy (BCE) loss without using the labeled known data:

$$\mathcal{L}_{\mathrm{BCE}}^{\mathrm{U},k} = \frac{-1}{{N_B^{\mathrm{U}}}^2} \sum_{i=1}^{N_B^{\mathrm{U}}} \sum_{j=1}^{N_B^{\mathrm{U}}} \Big(s_{ij} \log g_\omega^{\mathrm{U}}(\boldsymbol{z}_i^{\mathrm{U},k})^\top g_\omega^{\mathrm{U}}(\boldsymbol{z}_j^{\mathrm{U},k}) + (1 - s_{ij}) \log(1 - g_\omega^{\mathrm{U}}(\boldsymbol{z}_i^{\mathrm{U},k})^\top g_\omega^{\mathrm{U}}(\boldsymbol{z}_j^{\mathrm{U},k})) \Big), \tag{15}$$

where g_ω^{U} is the novel head of classifier. s_{ij} is obtained by using feature rank statistics, and $s_{ij} = 1$ when the top-k ranked dimensions of two samples in a data pair are identical, otherwise $s_{ij} = 0$. Besides, when we apply AutoNovel in the centralized training setting, we observe that

the model performance drops significantly due to the forgetting of learned knowledge in known categories. To compensate for such forgetting, we also equip AutoNovel with EMA.

GM [62] is proposed to incrementally discover and learn novel categories, and thus its problem settings are similar to **FC^{2}DL** except for whether the model is trained via FL or not. In the design of GM, the model is alternatively updated by a growing phase and a merging phase using the novel class data only. In the setting of **FC^{2}DL**, each selected client conducts the growing phase in the first 20 global epochs and then conducts the merging phase in the last 10 global epochs. Like **GPA**, the EMA is applied after each global epoch, and β is set as 0.99. iNCD [44] is also designed for discovering and learning novel categories continually like GM. There is a feature replay loss in iNCD that relies on the statistical information of known-category data, and we provide such information while conducting iNCD in FL, though we believe that it is unreasonable for both the server and clients to be aware of statistics of known categories. The same setup of known-category data is also employed in Happy [36], one of the latest continual NCDL approaches. IIC [32] heavily relies on known data to construct the loss function, and thus we have to randomly pick half of the known data and provide them to IIC in the novel category learning stage. OpenCon [46], as another label-available approach, is provided with the full set of known-category data during the stage of novel class learning in **FC^{2}DL**. As IIC and OpenCon have been provided with known-category data, we don't incorporate forgetting compensation techniques into them. As for Orchestra [35], although it is a federated self-supervised learning method, only a few modifications are needed to adapt it for solving **FC^{2}DL**. Specifically, we first apply Orchestra to tune the feature extractor on unlabeled novel-category data. During the tuning, we equip Orchestra with EMA to preserve the known-category performance. After sufficient tuning (30 global rounds for each novel category learning stage), the feature extractor is expected to produce more distinguishable representations for novel category data. At this moment, we freeze the feature extractor followed by a new classifier head, and train this head with pair-wise BCE loss of AutoNovel. As for SemiFL [9], although it is designed for cases in which the labeled and unlabeled data samples belong to the same category space, we found that it can also work somewhat effectively where there is no labeled data. As a result, we directly apply SemiFL in **FC^{2}DL** except for assuming there is no labeled data in the novel category learning stage. Furthermore, FedoSSL [61] and AGCL [40] do consider the recognition of unseen categories in FL, but their effectiveness relies on labeled known-category data, which may not be available in **FC^{2}DL**. However, for a fair comparison, we also provide them with a certain number of labeled known-category data during experiments (same as what we provide to IIC and OpenCon).

In addition to these NCDL baseline methods, we also provide the implementation details of two state-of-the-art novel-category number estimation approaches (MACC [50] and EMaCS [7]). As mentioned in the main paper, existing novel-category number estimation methods usually require a certain number of labeled data – MACC and EMaCS are no exceptions. Fortunately, MACC and EMaCS operate in a way that doesn't need labeled data from novel categories. Instead, they only require labeled data from known categories. This requirement can be met without the need to directly provide real known-category data. Specifically, we apply prototype augmentations on known-category prototypes $\mathcal{P}^{\mathrm{L}}$ constructed as neuron weights of classifier g_ω to get labeled known-category representations as follows:

$$\{\boldsymbol{z}_{c,i} = \boldsymbol{p}_c^{\mathrm{L}} + n * \min d(\mathcal{Z}^{\mathrm{U}})\}_{i=1,c=1}^{K^{\mathrm{U}},C^{\mathrm{U}}}, \tag{16}$$

where $n \sim \mathcal{N}(0, 0.01)$ is a random vector with the same dimension as prototypes (we have tried our best to find a suitable Gaussian distribution and found the one with a variance of 0.01 is the best), and $d(\cdot)$ computes all sample-pair Euclidean distances. The reason why we use the minimum sample-pair distance in the local prototype pool $\mathcal{Z}^{\mathrm{U}}$ is to ensure that the augmented representations are located in clusters with high density.

Then, we also provide the details of implementing the combination methods between other FL algorithms and Kmeans or our **GPA**. FedProx [31] introduces a regularization term for balancing the difference between local models and the global model at each round. Thus we follow FedProx to add such regularization to both known category learning and novel category learning stages. As for SCAFFOLD [23], it incorporates a control variate that indicates the stage of each FL participant, and with this control variate, the drift caused by non-IID can be mitigated. In the problem of our interest, we maintain the updating policy of this control variate in SCAFFOLD for both $\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$. Moon [29] leverages the principle of contrastive learning and proposes to conduct model parameter contrastive comparison against non-IID. We also include such parameter contrastive loss in the training of both $\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$. As mentioned in the main paper, unsupervised clustering, such as Kmeans, can be applied to these FL algorithms to discover and learn novel categories. But note that dedicated modifications are still needed as unsupervised clustering is impacted by non-IID, e.g., Kmeans is non-parametric thus it is hard to develop a global Kmeans mechanism that is effective to all participants without cluster merging. To apply Kmeans, we leverage **PPM** to merge clusters of participants and build global prototypes. Then we share these global prototypes with all participants, and they can calculate the distance between their local samples and these prototypes

Table 6. Experiments without using novel-category data filtering when employing **GPA** in **FC^{2}DL**.

Method	CIFAR-100 known	CIFAR-100 novel	CIFAR-100 all	Tiny-ImageNet known	Tiny-ImageNet novel	Tiny-ImageNet all	ImageNet-S known	ImageNet-S novel	ImageNet-S all
GPA-mixture	$71.8_{\pm 0}$	$52.4_{\pm 1.0}$	$67.9_{\pm 0.4}$	$57.5_{\pm 0}$	$39.5_{\pm 1.0}$	$55.7_{\pm 0.3}$	$55.8_{\pm 0}$	$38.0_{\pm 1.2}$	$52.2_{\pm 0.3}$
GPA	$\mathbf{71.9}_{\pm 0}$	$\mathbf{52.3}_{\pm 1.0}$	$\mathbf{68.0}_{\pm 0.4}$	$\mathbf{57.6}_{\pm 0}$	$\mathbf{39.1}_{\pm 1.9}$	$\mathbf{55.8}_{\pm 0.7}$	$\mathbf{55.8}_{\pm 0}$	$\mathbf{37.4}_{\pm 1.0}$	$\mathbf{52.1}_{\pm 0.4}$

Table 7. Performance comparison between **GPA** and other NCDL methods in centralized training setting. The ending 20 categories are the novel categories and there is only one novel category learning stage.

Method	CIFAR-100 known	CIFAR-100 novel	CIFAR-100 all	Tiny-ImageNet known	Tiny-ImageNet novel	Tiny-ImageNet all	ImageNet-S known	ImageNet-S novel	ImageNet-S all
AutoNovel	$34.6_{\pm 1.4}$	$42.0_{\pm 8.4}$	$36.1_{\pm 2.0}$	$27.2_{\pm 1.5}$	$30.7_{\pm 0.9}$	$27.5_{\pm 1.5}$	$41.5_{\pm 0.3}$	$31.9_{\pm 7.3}$	$39.5_{\pm 0.2}$
GM	$71.5_{\pm 0}$	$32.0_{\pm 3.0}$	$63.6_{\pm 0.1}$	$57.3_{\pm 0}$	$25.7_{\pm 1.6}$	$54.2_{\pm 0}$	$\mathbf{55.8}_{\pm 0}$	$26.9_{\pm 8.4}$	$50.0_{\pm 0.3}$
iNCD	$66.3_{\pm 1.6}$	$40.9_{\pm 7.7}$	$61.2_{\pm 2.5}$	$53.0_{\pm 1.4}$	$29.0_{\pm 0.5}$	$50.6_{\pm 1.2}$	$51.7_{\pm 0.2}$	$31.8_{\pm 2.1}$	$47.8_{\pm 0.2}$
Happy	$72.0_{\pm 0.5}$	$45.6_{\pm 1.3}$	$66.7_{\pm 0.6}$	$57.2_{\pm 0.3}$	$34.1_{\pm 2.2}$	$54.9_{\pm 0.7}$	$55.5_{\pm 0.8}$	$31.5_{\pm 0.6}$	$50.7_{\pm 0.4}$
IIC	$68.2_{\pm 1.0}$	$45.4_{\pm 3.9}$	$63.6_{\pm 1.1}$	$55.5_{\pm 0.3}$	$27.0_{\pm 1.3}$	$52.6_{\pm 0.3}$	$52.1_{\pm 1.1}$	$30.4_{\pm 3.4}$	$47.7_{\pm 0.2}$
OpenCon	$63.3_{\pm 1.0}$	$30.7_{\pm 9.7}$	$56.8_{\pm 0.9}$	$48.4_{\pm 0.3}$	$29.9_{\pm 0}$	$46.2_{\pm 0}$	$47.2_{\pm 4.2}$	$22.6_{\pm 2.4}$	$42.3_{\pm 3.6}$
SemiFL	$39.8_{\pm 4.4}$	$31.5_{\pm 0.9}$	$38.1_{\pm 1.0}$	$45.6_{\pm 0.9}$	$34.3_{\pm 2.0}$	$44.5_{\pm 0.9}$	$42.6_{\pm 1.3}$	$30.0_{\pm 1.4}$	$40.1_{\pm 0.8}$
FedoSSL	$50.3_{\pm 1.1}$	$40.4_{\pm 1.0}$	$48.3_{\pm 0.3}$	$54.0_{\pm 0}$	$28.8_{\pm 0.8}$	$51.5_{\pm 0.2}$	$55.4_{\pm 0.5}$	$26.7_{\pm 0.3}$	$49.7_{\pm 0}$
AGCL	$67.3_{\pm 0.4}$	$44.2_{\pm 1.6}$	$62.7_{\pm 0.5}$	$52.0_{\pm 0}$	$29.6_{\pm 0.7}$	$49.8_{\pm 0.1}$	$55.2_{\pm 1.0}$	$31.1_{\pm 0.4}$	$50.4_{\pm 0.2}$
GPA	$\mathbf{72.0}_{\pm 0.2}$	$\mathbf{56.5}_{\pm 1.3}$	$\mathbf{68.9}_{\pm 0.4}$	$\mathbf{57.5}_{\pm 0}$	$\mathbf{45.7}_{\pm 1.7}$	$\mathbf{56.3}_{\pm 0.4}$	$55.7_{\pm 0}$	$\mathbf{43.1}_{\pm 0.9}$	$\mathbf{53.2}_{\pm 0.2}$

to allocate the cluster labels.

A.3. Ablation study settings

According to the detachability and fungibility of different modules in **GPA**, we conduct a thorough ablation study that has been mentioned in the main paper. Here we provide detailed experiment settings and more results. The modified prototype contrastive loss $\mathcal{L}_{\text{PCL}}$ in the known category learning, the contrastive weighted loss $\mathcal{L}_{\text{CWL}}$ and the data mixup $\mathcal{L}_{\text{MIX}}$ in the novel category learning are easy to be detached from **GPA** for ablation study. To validate the effectiveness of **SWL**, instead of arbitrarily replacing **SWL** with some loss functions that are expected to perform poorly, e.g., pseudo-labeling-based CrossEntropy loss or prototype learning loss, we choose the most effective self-training loss in label-unavailable NCDL – pair-wise BCE loss [16]. Model mixup is used in **GPA** to compensate for the forgetting of known categories during novel category learning. We replace model mixup with the old model logits distillation and known prototype augmentation-based feature replay, which are used in iNCD [44] to alleviate forgetting on known categories, during the ablation study.

Experiments without using novel category data filtering. In the current design of **GPA**, we apply a data filtering mechanism to screen out novel-category data and don't store highly potential unlabeled known-category data during novel category learning to reduce memory cost and comply with privacy or intellectual property regulations. However, it only requires a minor modification if we want to handle the scenario where there is a mixture of both unlabeled known and novel-category data. That is, we detach the usage of data filtering to prepare the novel dataset and, instead, leverage it to build the high-confidence subset for data mixup. We carry out experiments for such cases, and the results are shown in the row '**GPA**-mixture' of Table 6. According to these results, there are even some improvements when incorporating unlabeled known-category data in the training, though we think it is impractical in real-world scenarios.

B. More Experiments

B.1. Experiment results in the centralized setting

To further evaluate the general application potential of **GPA**, we also carry out experiments in a centralized training setting. Specifically, we assume that there is only one participant who owns the entire dataset conducting one novel category learning stage with 20 novel categories, and the results are shown in Table 7. It is clear that ***GPA achieves the best performance on nearly all metrics even in the centralized learning setting.***

B.2. Experiment results with various degrees of heterogeneity

It is worth exploring how robust **GPA** is when faced with varying levels of data heterogeneity. Therefore, we test **GPA** and other baseline methods in a more challenging FL scenario, in which the data distributions among participants are more heterogeneous (α of Dirichlet Distribution is set as 0.001). From the experiment results in Table 8, we can observe that **GPA** still performs the best in all cases on all metrics. Combined with the experiment results shown in Table 1 of the main paper and Table 7 here, which exactly correspond to the settings of $\alpha = 0.1$ and $\alpha \to +\infty$, re-

Table 8. Performance comparison between **GPA** and other baselines for **FC^{2}DL** with more heterogeneous data distributions (α of Dirichlet Distribution is set as 0.001). The ending 20 categories are the novel categories with only one novel category learning stage. The experiment results present that **GPA** can retain the performance superiority when the data heterogeneity becomes much heavier.

	CIFAR-100			Tiny-ImageNet			ImageNet-S		
Method	known	novel	all	known	novel	all	known	novel	all
AutoNovel	$65.8_{\pm 4.6}$	$28.8_{\pm 7.0}$	$58.4_{\pm 4.5}$	$34.7_{\pm 14.0}$	$23.0_{\pm 8.1}$	$33.3_{\pm 11.8}$	$54.6_{\pm 0.1}$	$20.4_{\pm 3.0}$	$47.8_{\pm 0.1}$
GM	$71.5_{\pm 0}$	$27.3_{\pm 0.9}$	$62.7_{\pm 0}$	$57.3_{\pm 0}$	$19.6_{\pm 3.7}$	$53.6_{\pm 0}$	$55.8_{\pm 0}$	$19.4_{\pm 0.7}$	$47.5_{\pm 2.7}$
iNCD	$69.9_{\pm 0.1}$	$31.5_{\pm 3.7}$	$62.2_{\pm 0}$	$56.2_{\pm 0.4}$	$6.5_{\pm 7.1}$	$51.2_{\pm 0.7}$	$56.0_{\pm 0.3}$	$21.3_{\pm 1.1}$	$48.9_{\pm 0}$
Happy	$71.1_{\pm 0.1}$	$32.0_{\pm 1.5}$	$63.3_{\pm 0.6}$	$57.0_{\pm 0}$	$24.5_{\pm 1.7}$	$53.8_{\pm 0.5}$	$55.5_{\pm 0}$	$23.7_{\pm 1.2}$	$49.1_{\pm 0.4}$
IIC	$66.8_{\pm 0.5}$	$20.8_{\pm 6.6}$	$57.6_{\pm 1.2}$	$47.7_{\pm 0.9}$	$15.6_{\pm 2.1}$	$44.5_{\pm 1.1}$	$53.0_{\pm 0}$	$15.3_{\pm 0}$	$45.4_{\pm 0}$
Orchestra	$66.9_{\pm 0.1}$	$30.2_{\pm 0.8}$	$58.2_{\pm 0}$	$53.0_{\pm 0.1}$	$22.4_{\pm 0.5}$	$48.7_{\pm 0.2}$	$52.2_{\pm 0.4}$	$19.7_{\pm 0.1}$	$46.6_{\pm 0}$
OpenCon	$20.3_{\pm 13.6}$	$22.1_{\pm 12.0}$	$20.6_{\pm 12.9}$	$11.9_{\pm 2.7}$	$16.5_{\pm 10.7}$	$12.6_{\pm 4.3}$	$20.9_{\pm 7.1}$	$17.6_{\pm 0.3}$	$20.2_{\pm 5.0}$
SemiFL	$40.9_{\pm 2.0}$	$24.4_{\pm 1.5}$	$37.6_{\pm 0.9}$	$33.2_{\pm 0.5}$	$24.3_{\pm 0.8}$	$32.3_{\pm 0.1}$	$35.2_{\pm 1.2}$	$26.0_{\pm 0.9}$	$33.4_{\pm 0.4}$
FedoSSL	$60.7_{\pm 0.4}$	$18.0_{\pm 0}$	$52.2_{\pm 0}$	$50.8_{\pm 0.6}$	$10.0_{\pm 2.0}$	$46.7_{\pm 0.8}$	$52.6_{\pm 0.7}$	$16.1_{\pm 1.0}$	$45.3_{\pm 0.2}$
AGCL	$60.4_{\pm 1.7}$	$32.0_{\pm 1.4}$	$54.7_{\pm 0.3}$	$43.7_{\pm 0.6}$	$20.9_{\pm 1.4}$	$41.4_{\pm 0.3}$	$45.5_{\pm 0}$	$20.5_{\pm 1.1}$	$40.5_{\pm 0.2}$
GPA	$\mathbf{71.0}_{\pm 0.3}$	$\mathbf{44.7}_{\pm 1.8}$	$\mathbf{65.7}_{\pm 0.6}$	$\mathbf{57.3}_{\pm 0}$	$\mathbf{34.0}_{\pm 1.3}$	$\mathbf{55.0}_{\pm 0.3}$	$\mathbf{55.6}_{\pm 0.2}$	$\mathbf{32.6}_{\pm 1.0}$	$\mathbf{51.0}_{\pm 0.2}$

Table 9. Performance comparison between **GPA** and other baselines for **FC^{2}DL** with different data partitions. 20 categories of CIFAR-100 are randomly selected with different seeds as the novel categories with one novel category learning stage. The experiment results present that both **PPM** and **GPA** are robust to different data partitions.

Seed	2023			2024			2025		
PPM est.#	19			20			20		
Method	known	novel	all	known	novel	all	known	novel	all
AutoNovel	73.1	21.3	62.7	73.1	22.1	63.0	71.5	21.6	61.8
GM	71.1	38.3	64.5	70.4	21.6	60.6	69.7	41.5	64.1
iNCD	73.0	24.7	63.4	72.5	25.5	63.3	71.0	24.4	62.6
Happy	71.0	39.4	64.7	71.5	37.9	64.8	71.9	39.0	65.3
IIC	55.1	36.2	51.3	49.8	33.1	46.4	51.6	36.4	48.6
OpenCon	60.7	20.9	52.8	61.5	18.7	52.9	60.0	19.5	52.2
Orchestra	71.1	22.8	61.5	72.0	25.5	62.8	70.0	20.8	60.8
SemiFL	47.0	30.6	43.7	52.3	29.5	47.7	48.7	30.4	45.0
FedoSSL	65.1	22.4	56.6	64.3	23.8	56.2	63.0	20.3	54.5
AGCL	67.2	36.6	61.1	65.9	35.5	59.8	66.0	36.3	60.1
GPA	**73.0**	**54.4**	**69.3**	**73.3**	**52.0**	**69.0**	**72.5**	**53.2**	**68.6**

spectively, we can conclude that our approach **GPA** is able to consistently perform well and achieve effective novel category learning in **FC^{2}DL** under various levels of data heterogeneity.

B.3. Experiment results in settings with different data partitions

In experiments of the main paper, following regular NCDL studies [44, 62], we choose the same widely-used data partition settings based on ordered label sequence, but this does not mean that the effectiveness of **PPM** relies on the data partition. We conduct additional experiments of various data partitioning settings by randomly selecting 20 categories as the novel ones while the rest are the known categories with three seeds, 2023, 2024, and 2025, respectively. The detailed results are shown in Tables 9, 10 and 11. We can observe that regardless of data partitions, **PPM** always provides accurate novel-category number estimation, and **GPA** performs the best all the time.

B.4. Sensitivity analysis of hyper-parameters

There are several hyper-parameters in our algorithm, some of which are directly adopted from common values. For instance, the τ in Eqs. (4), (9), and (10) in the main paper are set to 0.07 as the standard contrastive loss for fair comparison. As for others, we conduct thorough sensitivity analysis, including η in the known category learning stage, r' in the data mixup, and β in the model mixup on CIFAR-100 with one novel category learning stage. The experiment results are shown in Table 12. We can observe that **GPA** is robust to different values of η and r', and performs the best when $\eta = 0.10$ and $r' = 0.95$. Different values of β directly associate with the preservation of feature extraction ability learned in the known category learning stage, thus it impacts the performance of **GPA**.

B.5. Experiment results with a large number of participants

In real-world application scenarios, there can be a large amount of participants in the FL system. Thus, we evaluate

Table 10. Performance comparison between **GPA** and other baselines for **FC^{2}DL** with different data partitions. 20 categories of Tiny-ImageNet are randomly selected with different seeds as the novel categories with one novel category learning stage. The experiment results present that both **PPM** and **GPA** are robust to different data partitions.

Seed	2023			2024			2025		
PPM est.#	21			19			20		
Method	known	novel	all	known	novel	all	known	novel	all
AutoNovel	56.5	24.9	53.5	56.2	20.6	53.1	57.1	19.7	53.4
GM	56.6	22.0	53.1	56.7	24.8	53.5	56.2	26.2	53.2
iNCD	56.5	27.8	53.8	56.0	22.7	53.3	56.5	22.4	53.3
Happy	56.5	27.5	53.6	56.2	28.0	53.4	55.9	27.9	53.1
IIC	46.5	23.1	44.2	48.2	21.7	45.6	47.3	18.6	44.4
OpenCon	45.5	16.0	43.2	42.7	18.2	40.1	47.0	15.5	43.8
Orchestra	56.5	25.0	53.5	56.5	22.4	53.4	56.7	22.0	53.3
SemiFL	36.4	25.9	35.4	40.2	27.8	39.0	36.5	26.0	35.5
FedoSSL	56.0	10.8	51.5	56.2	18.9	52.5	56.7	12.0	52.2
AGCL	47.0	25.5	44.9	46.6	22.9	44.2	48.0	27.5	46.0
GPA	**56.7**	**39.8**	**55.0**	**56.9**	**39.4**	**55.2**	**57.3**	**38.5**	**55.4**

Table 11. Performance comparison between **GPA** and other baselines for **FC^{2}DL** with different data partitions. 20 categories of ImageNet-S are randomly selected with different seeds as the novel categories with one novel category learning stage. The experiment results present that both **PPM** and **GPA** are robust to different data partitions.

Seed	2023			2024			2025		
PPM est.#	19			19			20		
Method	known	novel	all	known	novel	all	known	novel	all
AutoNovel	54.0	19.5	47.1	57.2	22.9	50.4	55.5	17.0	47.8
GM	54.0	29.7	49.3	55.3	25.2	49.3	54.7	24.1	48.6
iNCD	54.0	22.7	47.6	56.7	23.0	50.1	55.0	20.7	48.1
Happy	54.0	29.7	49.1	55.6	29.4	50.4	55.0	29.0	49.8
IIC	54.0	27.7	48.8	54.2	30.6	49.5	54.3	26.6	48.7
OpenCon	47.0	16.6	41.2	46.7	18.3	41.4	47.2	18.0	41.5
Orchestra	53.0	20.7	47.0	56.2	24.0	50.1	55.0	20.5	48.1
SemiFL	42.4	25.0	38.9	44.0	26.9	40.6	41.5	25.0	38.2
FedoSSL	53.0	17.7	45.9	54.0	22.5	47.7	51.5	20.9	45.4
AGCL	44.7	20.9	39.9	46.8	23.2	42.1	45.0	22.1	40.4
GPA	**54.1**	**36.5**	**50.6**	**57.0**	**37.5**	**53.1**	**55.6**	**37.2**	**51.9**

FC^{2}DL under a situation where there are a large number of participants for CIFAR-100 with one novel category learning stage. Specifically, we randomly select 20 clients from all participants every global round to conduct local training. The results are shown in Table 13, which shows that **GPA** can still perform the best when there are many participants.

B.6. Experiment results of launching data reconstruction attacks

In **GPA**, the sharing of local prototypes only occurs once before novel category learning, which is more secure than many FL works [21, 48] of other fields where prototypes are shared every round. The local prototypes are constructed by conducting unsupervised clustering on unlabeled novel data. Each participant maintains an identical count of clusters, and the semantic affiliations of distinct clusters across different clients diverge, as illustrated in Eq. (6) of the main text. Therefore, a direct comparison of cluster labels of local prototypes does not divulge sensitive label information. Label information leakage could arise from comparing the similarities between different participants' local prototypes. However, when local prototypes are shared, the model's capability to extract meaningful features from unlabeled novel data is relatively weak. Consequently, the similarity between prototypes is unreliable, implying that similar prototypes might correspond to different categories, and distinct prototypes could potentially correspond to the same category. This also prevents privacy leakage caused by data reconstruction attacks. The reason is that when constructing prototypes, each cluster contains a mixture of multiple categories due to the weak feature extraction ability, thereby causing the cluster centers to contain information from multiple categories. Moreover, for those clusters only with small data volumes, the high similarity in the representation space may not reliably correspond to a similarly high similarity in the input space. Moreover, works [21, 48]

Table 12. Sensitivity analysis of η in known category learning stage, r' in data mixup, and β in model mixup. Experiments are conducted using CIFAR-100 with one novel category learning stage (20 novel categories).

η	known	novel	all	β	known	novel	all	r'	known	novel	all
0.02	71.6	49.8	67.2	0.10	43.8	27.5	40.5	0.50	71.7	51.7	67.7
0.05	71.9	51.1	67.7	0.50	64.0	38.2	58.8	0.80	71.7	52.0	67.8
0.10	**71.9**	**52.3**	**68.0**	0.80	68.0	41.4	62.7	0.90	71.9	52.1	67.9
0.50	71.1	52.0	67.3	0.95	71.4	51.6	67.4	0.95	**71.9**	**52.3**	**68.0**
1.00	71.0	51.3	67.1	0.99	**71.9**	**52.3**	**68.0**	0.99	71.9	52.0	67.9

Table 13. Performance comparison between **GPA** and other baselines for **FC^{2}DL** with more participants. The ending 20 categories of CIFAR-100 are used as the novel categories and there is only one novel category learning stage in **FC^{2}DL**. The experiment results present that **GPA** can retain the performance superiority where there are a large number of participants in **FC^{2}DL**.

Participant #	20			50			100		
Method	known	novel	all	known	novel	all	known	novel	all
AutoNovel	68.2	30.1	60.5	68.0	27.9	59.6	68.3	27.0	59.5
GM	71.1	36.1	64.1	71.1	31.3	63.2	71.1	30.8	63.1
iNCD	70.5	31.7	62.7	70.5	29.6	62.3	71.1	25.2	62.0
Happy	71.1	39.0	64.7	71.3	37.9	64.6	71.0	38.2	64.4
IIC	61.8	21.7	53.8	62.3	22.4	54.6	61.1	16.2	52.1
Orchestra	67.7	31.1	60.6	68.5	25.6	59.3	67.8	26.0	59.2
OpenCon	61.7	21.0	53.6	71.9	17.8	61.1	71.7	19.3	61.2
SemiFL	46.5	27.0	42.6	46.1	24.8	41.8	45.4	26.3	41.6
FedoSSL	61.8	18.0	53.0	61.9	17.6	53.0	60.7	16.5	51.9
AGCL	64.8	34.2	58.7	63.6	32.5	57.4	61.9	30.6	55.6
GPA	**72.0**	**52.0**	**68.0**	**71.7**	**51.4**	**67.6**	**71.5**	**51.2**	**67.4**

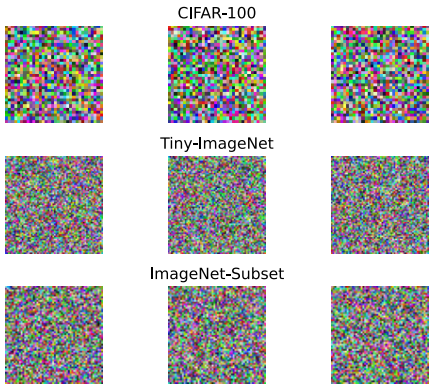

Figure 3. Recovered images of conducting data reconstruction attack on three local prototypes of three datasets. The data reconstruction attack optimizes the recovered images in the input space and tries to make their representations as close to local prototypes as possible. We apply Adam to minimize the mean square errors until the error cannot decrease further.

highlight that prototypes are formed in a low-dimensional space by averaging data representations, and mainstream model structures include numerous operations of dropout, pooling, and ReLU activations. These two factors are irreversible, which further strengthens the privacy preservation of our work.

Given the absence of data reconstruction attacks against representations in the literature, we speculate that attacks are most likely to follow Deepleakage [65] – optimizing dummy data to align the representation as closely as possible with the target. The results of this attack are shown in Figure 3. It is nearly impossible to capture interpretable semantics from the recovered images, showing that the privacy of prototypes is preserved.

B.7. Experiment results of the forgetting of known categories

To measure the forgetting of known categories after novel category learning, we only need to know the performance of m^{L} right after known category learning. Therefore, here we provide the detailed results before $\mathcal{T}^{\mathrm{U}}$, which are shown as Tables 14 and 15.

B.8. Experiment results on pre-trained vision transformers

In recent NCDL studies, the backbone model is assumed to be pre-trained with a large amount of public data, but in FL, this assumption is not reasonable as the data is private. However, we also test our **GPA** approach with other baseline methods using pre-trained vision transformers (ViT). We follow Ma et. al [36] to adopt ViT-B/16 as the backbone model, and the results are shown in Table 16. What we can observe is that **GPA** can still achieve the best performance in all cases, which validates the general effectiveness of it.

Algorithm 1 Overall Global Prototype Alignment.

Given: A global model $m = f_\theta \circ g_\omega$ consists of a feature extractor f_θ and a classifier g_ω. At the beginning, there are K^{L} participants $\{\mathcal{S}^1, \mathcal{S}^2, ..., \mathcal{S}^{K^{\mathrm{L}}}\}$, and each holds its labeled known category dataset $\{\mathcal{D}^{\mathrm{L},1}, \mathcal{D}^{\mathrm{L},2}, ..., \mathcal{D}^{\mathrm{L},K^{\mathrm{L}}}\}$. FL's total training round is E_g.

Known Category Learning:

for $t = 1, \cdots, E_g$ **do**
- Server randomly selects $\mathcal{K}^{\mathrm{L}}$ clients;
- **for** $\mathcal{S}^{s,k}$ *in* $\{\mathcal{S}^{s,1}, \mathcal{S}^{s,2}, \cdots, \mathcal{S}^{s,\mathcal{K}^{\mathrm{L}}}\}$ **do**
  - Apply $\mathcal{L}_{\mathcal{T}^{\mathrm{L}}} = \mathcal{L}_{\mathrm{CE}} + \eta\mathcal{L}_{\mathrm{PCL}}$;
  - Upload $\nabla_{\theta,\omega}\mathcal{L}^k_{\mathcal{T}^{\mathrm{L}}}$ to the Server;
- Server applies FedAvg to calculate aggregated gradients $\nabla_{\theta,\omega}\mathcal{L}^{\mathrm{Avg}}_{\mathcal{T}^{\mathrm{L}}}$;
- Server distributes $\nabla_{\theta,\omega}\mathcal{L}^{\mathrm{Avg}}_{\mathcal{T}^{\mathrm{L}}}$ to all participants.

Novel Category Data Filtering:

for $\mathcal{S}^k$ *in* $\{\mathcal{S}^1, ..., \mathcal{S}^{\mathrm{U}}\}$ **do**
- **while** $\mathcal{D}^{\mathrm{U},k}$ *is not full* **do**
  - Apply Eq. (5) of the main paper to filter out known category data;
  - Store the remaining data in the data memory and form the novel category dataset $\mathcal{D}^{\mathrm{U},k}$;

Novel Category Learning:

Conduct Algorithm 2;

Algorithm 2 Novel Category Learning.

Given: After known category learning, a model $m = f_\theta \circ g_\omega$ is given. There are K^{U} active participants $\{\mathcal{S}^1, \mathcal{S}^2, ..., \mathcal{S}^{K^{\mathrm{U}}}\}$, and each holds unlabeled novel category dataset $\{\mathcal{D}^{\mathrm{U},1}, \mathcal{D}^{\mathrm{U},2}, ..., \mathcal{D}^{\mathrm{U},K^{\mathrm{U}}}\}$. **PPM** ascending iteration is E. FL's training round is E_g.

All Active Participants:

for $\mathcal{S}^k$ *in* $\{\mathcal{S}^1, \mathcal{S}^2, \cdots, \mathcal{S}^{K^{\mathrm{U}}}\}$ **do**
- Apply Kmeans on $f_\theta(\mathcal{D}^{\mathrm{U},k})$;
- Return Kmeans cluster centers $\mathcal{Z}^{\mathrm{U},k}$;

Server:

// Conduct **PPM**

Receive and shuffle $\mathcal{Z}^{\mathrm{U}} = \cup_{k=1}^{K^{\mathrm{U}}} \mathcal{Z}^{\mathrm{U},k}$;

for $e = 0, \cdots, E$ **do**
- $D = \{d(\boldsymbol{z}^{\mathrm{U}}_i, \boldsymbol{z}^{\mathrm{U}}_j)\}_{\boldsymbol{z}^{\mathrm{U}}_i, \boldsymbol{z}^{\mathrm{U}}_j \in \mathcal{Z}^{\mathrm{U}}}$;
- $\epsilon_e = \min D + \frac{e}{E} \cdot (\max D - \min D)$;
- Apply DBSCAN with $n_{\mathrm{size}} = 2$ and ϵ_e;
- Record unique cluster number $\tilde{c}^{\mathrm{U}}_e$;

Estimate the novel category number as $\widetilde{C}^{\mathrm{U}} = \max\{\tilde{c}^{\mathrm{U}}_e\}_{e=0}^{E}$;

Apply Kmeans with $\widetilde{C}^{\mathrm{U}}$ to construct global prototypes $\mathcal{P}^{\mathrm{U}} = \{\hat{\boldsymbol{z}}^{\mathrm{U}}_c\}_{c=1}^{\widetilde{C}^{\mathrm{U}}}$ and randomly initialize g_ω;

Selected Clients:

for $t = 1, \cdots, E_g$ **do**
- Server randomly selects $\mathcal{K}^{\mathrm{U}}$ clients;
- **for** $\mathcal{S}^{s,k}$ *in* $\{\mathcal{S}^{s,1}, \mathcal{S}^{s,2}, \cdots, \mathcal{S}^{s,\mathcal{K}^{\mathrm{U}}}\}$ **do**
  - Apply **SWL**, **CWL**, and data mixup;
  - Update f_θ with model mixup and g_ω in backpropagation;
  - Upload gradients to the Server.
- Server applies FedAvg;
- Server distributes aggregated gradients;

Table 14. Model performance before novel category learning for CIFAR-100, Tiny-ImageNet, and ImageNet-S. All baseline methods rely on the same model m^{L}. Performance forgetting of known categories can be calculated by subtracting the known accuracy after novel category learning.

	CIFAR-100			Tiny-ImageNet			ImageNet-S		
Metric	known	novel	all	known	novel	all	known	novel	all
m^{L}	$72.7_{\pm 0.1}$	$13.4_{\pm 4.2}$	$59.9_{\pm 1.5}$	$57.6_{\pm 0}$	$15.7_{\pm 3.1}$	$53.4_{\pm 1.2}$	$55.8_{\pm 0.1}$	$12.8_{\pm 2.0}$	$47.2_{\pm 1.1}$

C. Overall Optimization Pipeline

GPA is proposed to enable FL systems to discover and learn unseen novel categories. We assume any FL system can periodically leverage **GPA** to incorporate novel categories or functionalities after the training on labeled known-category data. Specifically, when the novel category learning stage starts, all available FL participants at that moment first need to apply unsupervised clustering on their local data to find local potential prototypes. These local prototypes will be sent to the server only once and then **GPA** will apply **PPM** to merge these prototypes to estimate the global novel-category number and construct the global novel prototypes. Then at each round, each selected client can leverage **SWL**, **CWL**, and data mixup to conduct local training. The backbone local feature extractor is updated by our model mixup, while the classifier head is optimized via back-propagation. This pipeline is shown in Algorithm 2, and the overall **GPA** workflow is Algorithm 1.

D. Limitations and Future Work

This paper mainly focuses on achieving continuous category discovery and learning for FL from the algorithmic perspective. Although the proposed **GPA** can be empirically demonstrated effective, there is no rigorous theoretical proof that this will always be the case. Therefore, in future work, establishing a theoretical framework for **FC^{2}DL** and providing rigorous analysis of **GPA** could be the first step. Moreover, considering **FC^{2}DL** in more advanced scenarios

Table 15. Model performance before novel category learning for CUB200, StanfordCars, and Herbarium 19. All baseline methods rely on the same model m^{L}. Performance forgetting of known categories can be calculated by subtracting the known accuracy after novel category learning.

	CUB200			StanfordCars			Herbarium 19		
Metric	known	novel	all	known	novel	all	known	novel	all
m^{L}	$40.7_{\pm 0}$	$10.1_{\pm 1.2}$	$38.3_{\pm 0.5}$	$45.0_{\pm 0}$	$11.2_{\pm 0.1}$	$42.6_{\pm 0}$	$49.5_{\pm 0.1}$	$10.0_{\pm 1.0}$	$44.0_{\pm 0.2}$

Table 16. Performance comparison between **GPA** and other NCDL methods on pre-trained ViT-B/16. The ending 20 categories are the novel categories and there is only one novel category learning stage.

	CIFAR-100			Tiny-ImageNet			ImageNet-S		
Method	known	novel	all	known	novel	all	known	novel	all
AutoNovel	57.3	45.0	54.8	47.8	45.7	47.6	72.3	47.8	67.4
GM	86.0	40.2	76.8	77.9	43.3	74.4	81.5	47.6	74.7
iNCD	78.3	47.5	72.1	72.7	49.1	70.3	80.4	50.2	74.4
Happy	86.2	48.7	78.7	78.8	52.2	76.1	82.9	55.0	77.3
IIC	73.7	47.3	68.4	75.5	47.0	72.7	79.4	47.9	73.1
OpenCon	72.5	38.8	65.8	65.9	42.6	63.6	74.7	42.5	68.3
SemiFL	60.2	40.7	56.3	67.9	47.0	65.8	66.6	45.8	62.4
FedoSSL	75.5	46.2	69.6	75.0	48.8	72.4	76.7	47.0	70.8
AGCL	80.3	50.2	74.3	73.5	51.6	71.3	78.9	53.4	73.8
GPA	86.5	55.6	80.3	79.0	54.3	76.5	83.5	59.2	78.6

and domains, e.g., object detection in autonomous driving, medical semantic segmentation, and some natural language processing cases, is also worthy of exploring in the future.

E. Broader Impact

The **FC^{2}DL** study outlined in this paper presents substantial societal implications and potential benefits without an apparent negative impact. As privacy concerns become increasingly important, the need for efficient methods to handle dynamic data distributions without compromising privacy is critical. Our **GPA** framework is designed to address the challenges of **FC^{2}DL**, specifically in merging and aligning novel categories identified and learned by different clients while preserving privacy. This, in essence, supports a more sustainable and adaptable machine learning system that can evolve with changing data scenarios. **GPA**'s impressive results, even in non-FL or centralized training scenarios, indicate its potential for wide-reaching application in various real-world scenarios. These scenarios include but are not limited to, healthcare, financial services, telecommunications, and social networking platforms, where preserving user privacy while continually adapting to new information is paramount. Moreover, the improved model performance achieved with **GPA** can contribute to more reliable and efficient systems, enhancing user experiences and outcomes. We believe that our research in **FC^{2}DL** and the development of **GPA** pave the way for advancements in privacy-conscious, dynamic learning systems, fostering a more secure and adaptable digital landscape.