Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU

Supplementary Material

This supplementary material provides additional technical details, extended analyses, and supporting evidence for our main paper on machine unlearning in hyperbolic versus Euclidean contrastive learning spaces. We first present formal mathematical descriptions of the CLIP and MERU objectives to establish the geometric foundations that differentiate these approaches (Section 10). Next, we clarify the composition of our forget sets, highlighting the challenges in precisely defining concept boundaries (Section 11). We then provide comprehensive visualizations through confusion matrices that illustrate the different unlearning behaviors between models (Section 12). Additionally, we present complementary linear probing results that further confirm how feature representations are still linearly separable, which already can be observed in latent visualizations (Section 13). Finally, we include extended visualizations of the latent spaces using multiple dimensionality reduction techniques to further support our findings on hyperbolic unlearning (Section 14). These materials provide deeper technical understanding and additional empirical support for the conclusions presented in our main paper.

10. Model Objectives

10.1. CLIP: Contrastive Learning in Euclidean Space

CLIP [30] consists of two encoders, a visual encoder f_{img} and text encoder f_{txt}, mapping images and text into a shared Euclidean space $\mathbb{R}^d$. Given a batch of images and texts $\{(x_i, t_i)\}_{i=1}^N$, we obtain embeddings $x'_i := f_{\text{img}}(x_i)$ and $t'_i := (f_{\text{txt}} t_i)$. CLIP is trained extending 1 to a symmetric cross-entropy loss:

$$L_{\text{CLIP}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(x'_i, t'_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x'_i, t'_j)/\tau)}\right. \tag{18}$$

$$\left. + \log\frac{\exp(\text{sim}(x'_i, t'_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x'_j, t'_i)/\tau)}\right]. \tag{19}$$

where $\text{sim}(x'_i, t'_j) := \cos(\theta_{ij})$ is the cosine similarity between normalized image and text embeddings, θ_{ij} is the angle between them, and τ is a temperature parameter. This contrastive objective places all embeddings on a unit hypersphere, treating all concept relationships uniformly, regardless of their hierarchical nature.

10.2. MERU: Contrastive Learning in Hyperbolic Space

MERU [9] extends contrastive learning to hyperbolic space using Lorentz model. MERU consists of visual and textual encoders, but it projects the image and text embeddings onto a hyperboloid manifold. The distance between two points x, y in the hyperboloid is given by

$$d_{\mathcal{L}}(x, y) = \frac{1}{\sqrt{c}}\cosh^{-1}(-c\langle x, y\rangle_{\mathcal{L}}), \tag{20}$$

where

$$\langle x, y\rangle_{\mathcal{L}} = \langle x_{\text{space}}, y_{\text{space}}\rangle - x_{\text{time}} \cdot y_{\text{time}} \tag{21}$$

is the Lorentzian inner product, $x = (x_{\text{space}}, x_{\text{time}}) \in \mathbb{R}^{n+1}$, $x_{\text{space}} \in \mathbb{R}^n$ and $x_{\text{time}} \in \mathbb{R}$, and $c > 0$ is the curvature of the space. The Lorentzian norm is defined by $||x||_{\mathcal{L}} = \sqrt{|\langle x, x\rangle_{\mathcal{L}}|}$. With this, the Lorentz model of curvature $-c$, $c > 0$, and dimension n is given by the set of vectors:

$$\mathcal{L}^n := \{x \in \mathbb{R}^{n+1} \ : \ \langle x, x\rangle_{\mathcal{L}} = -1/c\}. \tag{22}$$

MERU is trained with a contrastive loss similar to CLIP, but using negative hyperbolic distance as the similarity measure, $\text{sim}_{\mathcal{L}}(x, y) = -d_{\mathcal{L}}(x, y)$. Additionally, MERU incorporates an entailment loss to enforce partial order relationships between text and image embeddings:

$$L_{\text{entail}}(x, t) = \max(0, \text{ext}(x, t) - \text{aper}(t)) \tag{23}$$

where $\text{ext}(x, t)$ is the exterior angle between the text embedding t, given by

$$\text{ext}(x, t) = \cos^{-1}\left(\frac{x_{\text{time}} + t_{\text{time}}\ c\ \langle x, t\rangle_{\mathcal{L}}}{\|t_{\text{space}}\|\sqrt{(c\ \langle x, t\rangle_{\mathcal{L}})^2 - 1}}\right), \tag{24}$$

and image embedding x, and $\text{aper}(t)$ is the half-aperture of the entailment cone for t,

$$\text{aper}(t) = \sin^{-1}\left(\frac{2K}{\sqrt{c}\|t_{\text{space}}\|}\right). \tag{25}$$

The hyperbolic geometry of MERU naturally accommodates hierarchical relationships, as the volume of the space grows exponentially with distance from the origin. This property allows generic concepts to be placed closer to the origin with more capacity to connect to numerous specific instances, in contrast to the uniform treatment of relationships in Euclidean space.

11. Defining the forget set

We built the different forget set aggregating related subfolders (e.g., *dog* = {*bordercollie, bostonterrier, etc.*}, see Table 6). However, other subfolders in the retain set, such as *alltheanimals*, may contain image-text pairs related to dogs, creating conflicting signals during unlearning. This highlights a broader challenge in the machine unlearning field: properly defining concept boundaries for removal remains an open problem [7].

12. Confusion Matrices from Zero-Shot Classification

Figures 3 and 4 present confusion matrices for zero-shot classification before and after unlearning for CLIP and MERU, respectively. These visualizations provide detailed insights into how concept removal affects classification behavior across different classes.

For CLIP (Figure 3), we observe partial concept removal, with the "dog" classification accuracy reduced but not eliminated. Most misclassified dog images are assigned to the "cat" category, indicating that CLIP maintains some understanding of semantic similarity even when attempting to forget. The retain classes show minimal disturbance, maintaining strong diagonal elements in the confusion matrix.

In contrast, MERU (Figure 4) exhibits complete concept removal, with dog images almost entirely reassigned to other categories. The redistribution follows semantic hierarchies, with most dog images classified as "cat" which is a semantically animal category. This pattern supports our hypothesis that hyperbolic geometry leverages hierarchical relationships during unlearning, reassigning forgotten concepts according to their position in the semantic taxonomy. We further observe a decrease in performance on retaining the "horse" concept. However, this can be explained by observing how the original MERU already confuses horses by dogs, and then HAC, treating horses as if they were dogs, also removes them.

13. Linear Probing

Linear probing extracts embeddings from image encoder and trains a linear classifier on these features. This evaluates whether class information remains linearly separable in the latent space after unlearning. Accuracy is reported for both retained and forgotten classes, quantifying how successfully target concepts have been removed while preserving desired knowledge. Testing linear separability of image features provides different insights. While R-acc and F-acc in zero-shot classification measure the unlearning performance in the alignment between images and texts embeddings. Here we measure whether after the damage is done the image features have been mixed between different categories or not.

The linear probe classification results provide complementary insights into how unlearning affects the underlying feature representations. Consistent with prior work on hyperbolic classification [14], Euclidean representations show slightly better linear separability. However, the high forget accuracies in both methods reveal an important distinction between our approach and traditional class-unlearning: alignment calibration specifically targets cross-modal associations rather than altering the fundamental feature structure of either modality in isolation. This explains why images from the class related to the concept to forget remain linearly separable—their visual features are preserved while their association with corresponding text is disrupted. This insight can be also illustrated in a qualitative analysis of the latent spaces Sec. 6.

14. Additional Latent Space Visualizations

To complement visualizations from Section 6 we include more instances in visualizations. Additionally, for the hyperbolic case, we include visualizations from another perspective, using CO-SNE [13]. This method leverages hyperbolic Cauchy distribution (instead of hyperbolic student's t-distribution) and Lorentz distance, to represent global hierarchy and local distances in the same visualization. This allow us to "zoom-in" and see the origin of the hyperboloid from a closer point of view. Figure 5, illustrates the latent space of MERU from three perspectives before and after unlearning. CO-SNE allow us to better see that text embeddings of "dogs" remain closer to the origin, while other instances are pushed further, as discussed in Section 6. Figure 6 illustrates the same idea when scaling the unlearning problem to multiple concept removal. Observe that when "cats" are included in the forget set, the text embeddings for cats also remain close to the origin, and this is not disturb when including "food" and "plants", illustrating the robustness of HAC at scaling the unlearning task.

Concept-class	Subreddits	Image-text samples	% on Redcaps	% on Redcaps2
dogs	dogpictures, bordercollie, bostonterrier, lookatmydog, doggos, bulldogs, australiancattledog, frenchbulldogs, bernesemountaindogs, australianshepherd, beagle, chihuahua, corgi, dobermanpinscher, husky, labrador, pitbulls, pomeranians, pug, pugs, rarepuppers, rottweiler	511585	4.26%	7.33%
cats	cats, blackcats, supermodelcats, catpictures, siamesecats, bengalcats, siberiancats	532640	4.43%	7.63%
food	food, foodporn, veganfoodporn, healthyfood, breakfastfood, chinesefood, tastyfood, budgetfood, baking, bento, breadit, breakfastfood, breakfast, burgers, chefit, pizza, sushi, tacos, veganrecipes, vegetarian	630971	5.25%	9.04%
plants	houseplants, plants, plantedtank, airplants, plantbaseddiet, plantsandpots, carnivorousplants, flowers, bonsai, botanicalporn, cactus, microgreens, monstera, orchids, permaculture, roses, succulents, vegetablegardening, gardening	587798	4.89%	8.42%
Total	68 subreddits	2262994	18.85%	32.43%

Table 6. Grouping of subreddits to higher-order concepts.

Table 7. Linear probing accuracy in retain set (R-acc) and forget set (F-acc), across different tasks, after unlearning: (A) "dog"; (B) "dog" and "cat"; (C) "dog", "cat", "food" and "plant". We report results for both CLIP and MERU after alignment calibration using the optimal configuration from Section 4.3. Values in **bold** indicate whether AC or HAC performed better at retaining or unlearning across A, B and C.

Task	Method	Unlearn Set	CIFAR-10		CIFAR-100		STL-10		O-IIIT Pets		Food101		Flowers102	
			R-acc	F-acc	R-acc	F-acc	R-acc	F-acc	R-acc	F-acc	R-acc	F-acc	R-acc	F-acc
Linear Probe Classification	AC	A	**89.9**	**85.2**	**71.5**	-	**95.1**	**92.4**	**86.1**	87.5	**84.5**	-	**95.4**	-
		B	**95.8**	**91.3**	**71.6**	-	**95.8**	**91.3**	-	87.3	**84.6**	-	**95.7**	-
		C	**95.9**	**91.4**	**71.0**	84.3	**95.9**	**91.4**	-	87.0	-	84.3	-	95.4
	HAC-reg	A	89.3	85.5	69.8	-	94.9	93.6	84.8	**86.7**	83.8	-	93.8	-
		B	95.5	92.1	69.7	-	95.5	92.1	-	**85.0**	83.9	-	93.7	-
		C	95.4	92.4	68.6	**83.1**	95.4	92.4	-	**85.6**	-	**83.0**	-	**92.6**

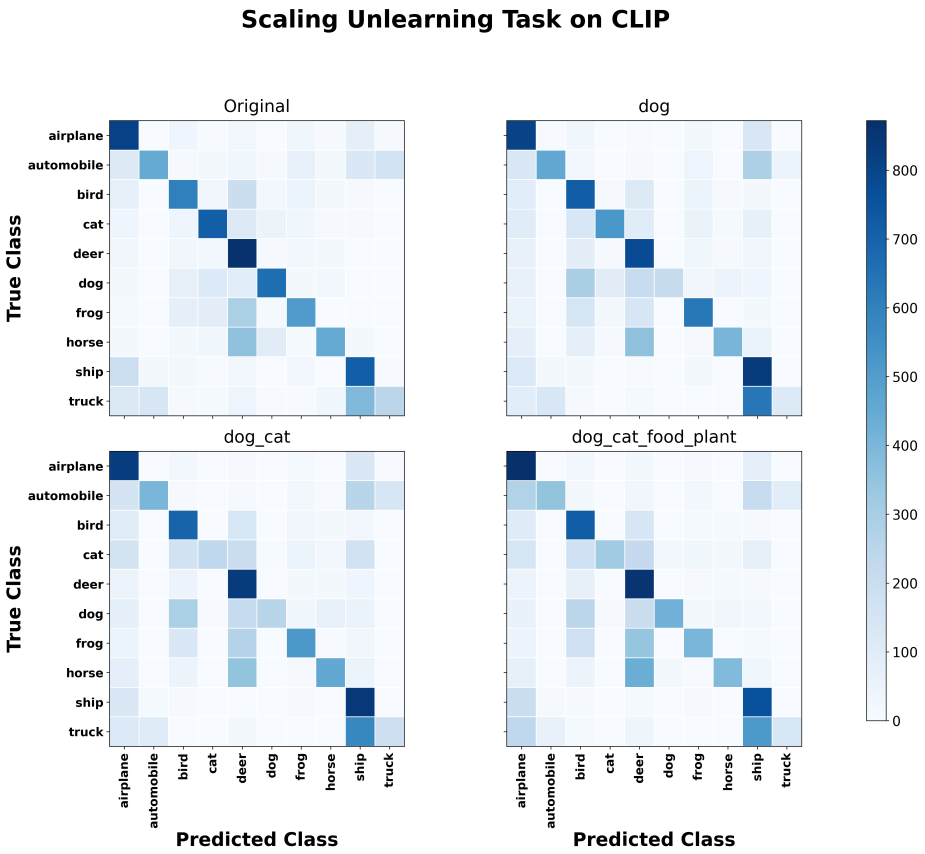


Figure 3. Confusion matrices for CLIP zero-shot classification at different scales of the unlearning task. After unlearning, CLIP shows moderate confusion, with dog images primarily misclassified as cats, but still retaining some dog classification capability.

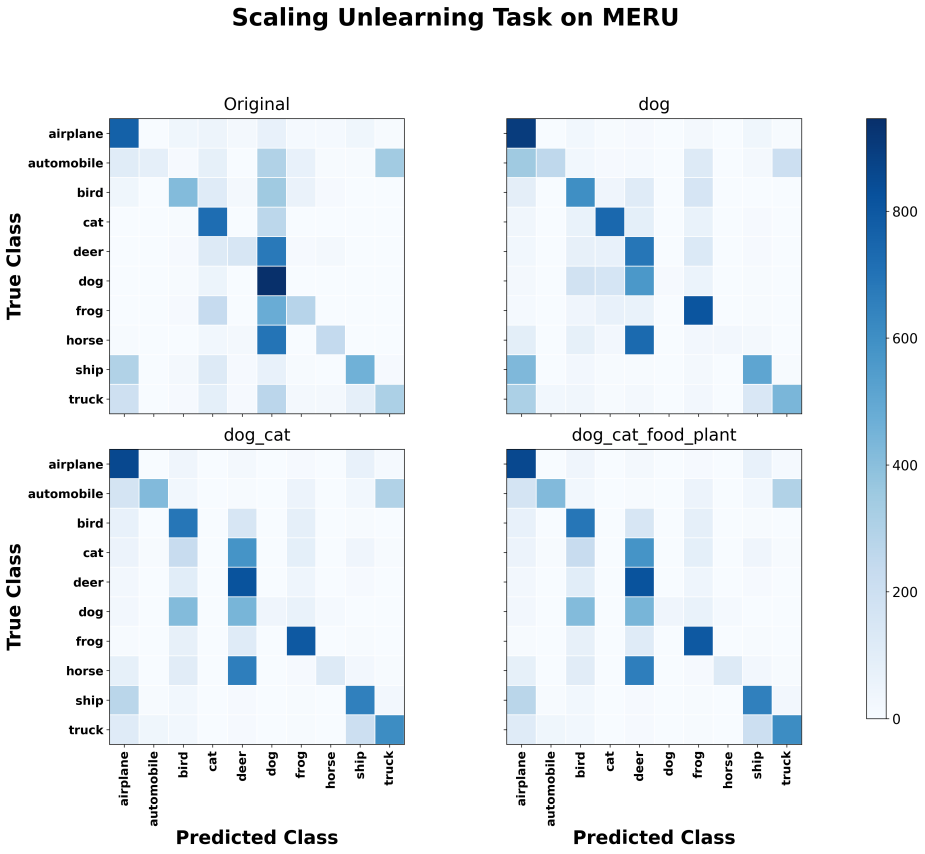
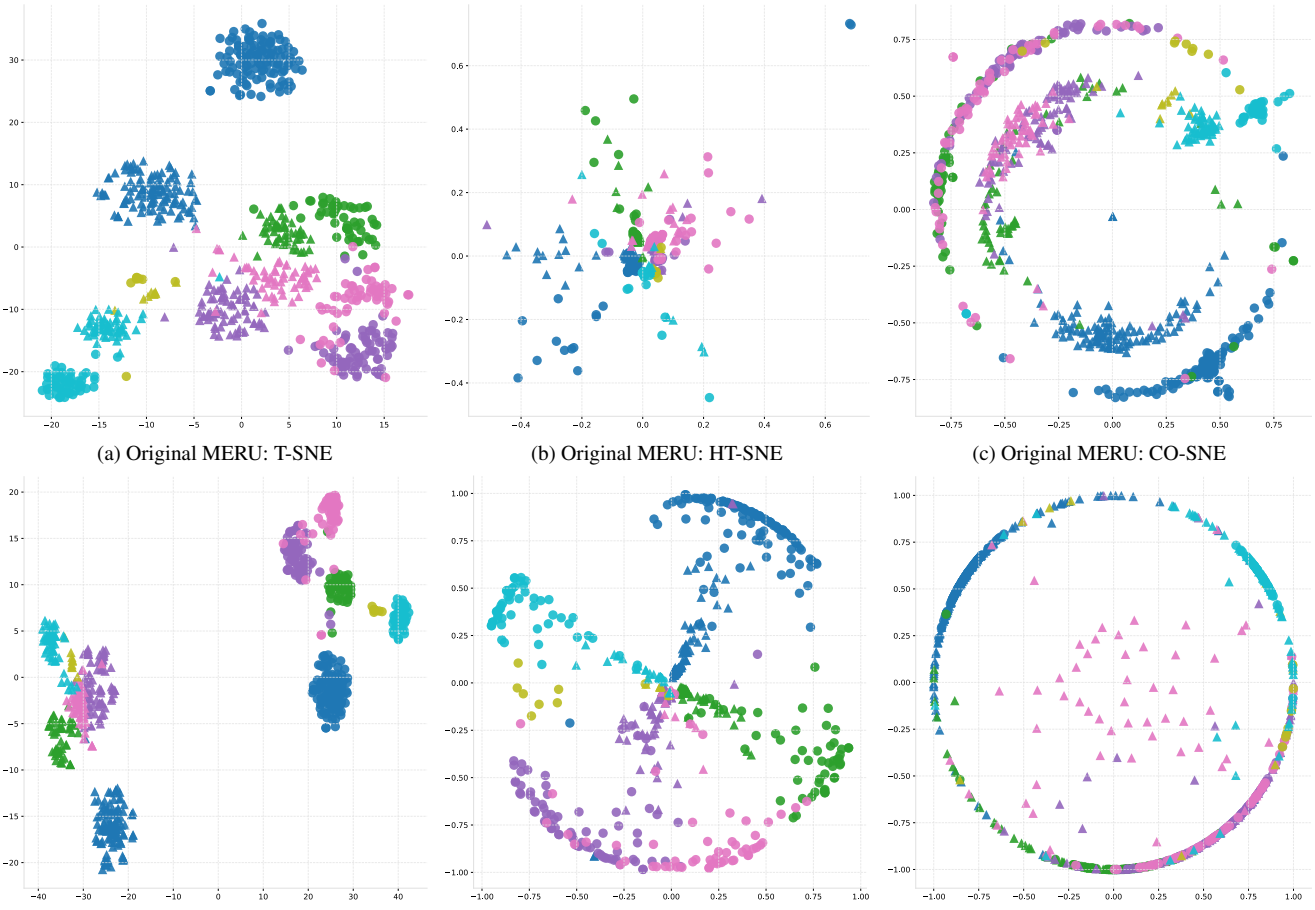


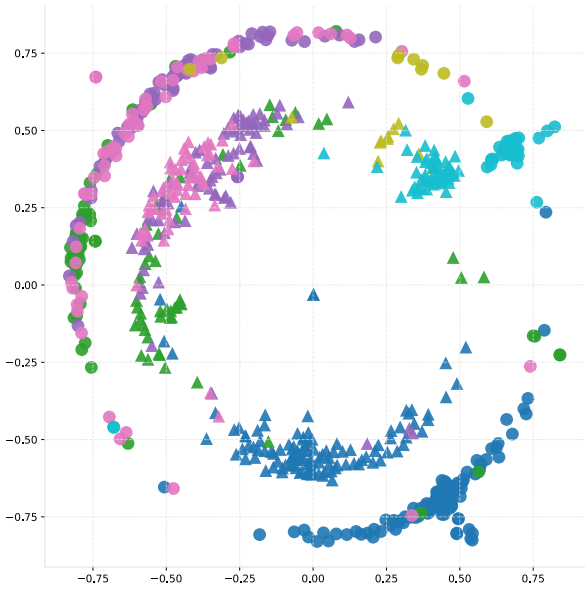
Figure 4. Confusion matrices for MERU zero-shot classification at different scales of the unlearning task. After unlearning, MERU demonstrates complete forgetting of the dog class, with dog images redistributed primarily to cat and horse categories according to semantic similarity.



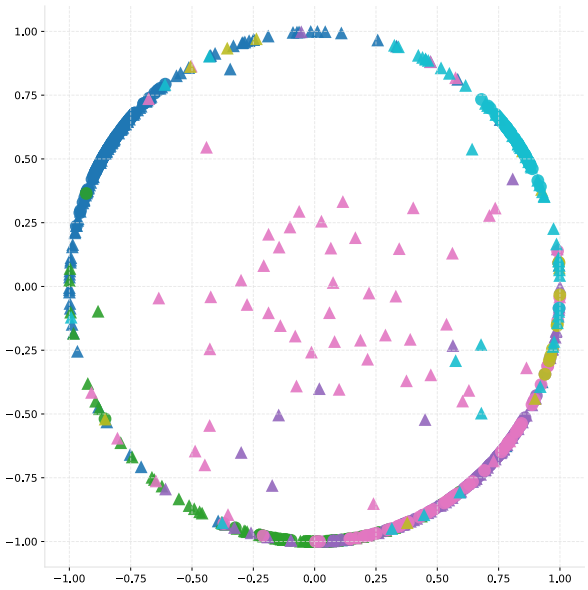
(a) Original MERU: T-SNE (b) Original MERU: HT-SNE (c) Original MERU: CO-SNE

(d) Unlearned MERU: T-SNE (e) Unlearned MERU: HT-SNE (f) Unlearned MERU: CO-SNE

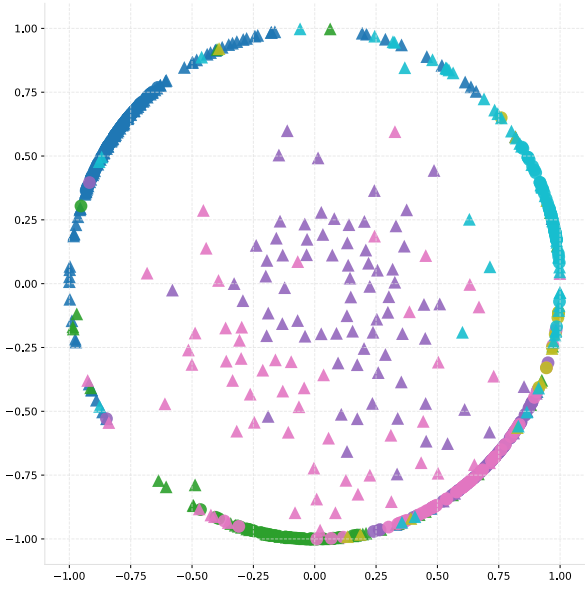
Figure 5. Latent space visualizations with T-SNE, hyperbolic T-SNE and CO-SNE of MERU before and after removing the concept-class "dog". △ refer to text embeddings, ○ to image embeddings, and colors to **dogs**, **cats**, **pizzas**, **buses**, **birds**, and **apples**.



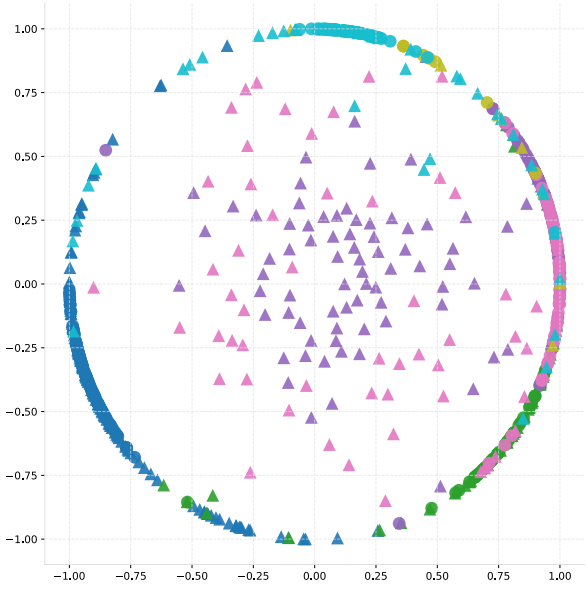
(a) Original MERU



(b) Unlearned ”dog”



(c) Unlearned ”dog”,”cat”



(d) Unlearned ”dog”, ”cat”, ”food”, ”plant”

Figure 6. Latent space visualizations with CO-SNE of MERU at different unlearning tasks. △ refer to text embeddings, ○ to image embeddings, and colors to **dogs**, **cats**, **pizzas**, **buses**, **birds**, and **apples**.