UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Mahmoud Said Hosny Elsayed Karamalla

FedCAPE: Federated Concept Alignment for Privacy-Preserving Explanations

Master's Thesis (30 ECTS)

Supervisor:
Radwa El Shawi, PhD

Tartu 2025

FedCAPE: Federated Concept Alignment for Privacy-Preserving Explanations

Abstract:

Traditional machine learning pipelines are limited by their dependence on centralized data, making them unsuitable for privacy-sensitive domains and distributed real-world settings. Furthermore, these methods often lack concept-level interpretability and require labor-intensive manual annotation to identify and explain meaningful concepts in complex datasets. While recent automated approaches, such as [1] proposed by Ghorbani et al.,(NeurIPS 2019), have advanced the automation of concept discovery and explanation, they do not address the challenges of privacy preservation, data decentralization, or collaborative concept alignment across multiple participants.

In this thesis, we propose **Federated Concept Alignment for Privacy-Preserving Explanations (FedCAPE)**, a novel Framework designed to enable scalable, privacy-preserving, and fully decentralized concept discovery, alignment, and interpretability. FedCAPE leverages self-supervised learning (DINO) and the multimodal capabilities of OpenAI CLIP to automatically assign semantic meaning to image segments, thereby eliminating the need for manual annotation and introducing a semantic layer over the extracted features. Critically, FedCAPE employs federated K-means clustering to collaboratively align and refine discovered concepts across clients, ensuring that shared conceptual knowledge emerges without the need to exchange raw data.

Through this federated approach, FedCAPE achieves end-to-end interpretability, improved concept alignment, and enhanced transparency of model predictions—surpassing both traditional and state-of-the-art automated approaches in terms of privacy preservation, scalability, and explainability. Experimental evaluation across multiple distributed clients demonstrated strong cross-client semantic consistency, with human evaluators preferring FedCAPE clusters over random baselines in 80–100% of cases. Quantitatively, FedCAPE achieved high TCAV scores for salient concepts, in some cases exceeding the centralized baseline while avoiding over-clustering

and improving cluster purity. These results highlight FedCAPE's potential to bridge the gap between interpretable AI and privacy-preserving machine learning in distributed environments.

Keywords: Federated Learning, Explainable AI, TCAV, Concept Alignment, Privacy

CERCS: P170 Computer science, numerical analysis, systems, control

FedCAPE: Mõistepõhine selgitamine privaatsust säilitavas hajusõppes

Lühikokkuvõte:

Traditsioonilised masinõppe töövood on piiratud oma sõltuvusega tsentraliseeritud andmetest, mis muudab need sobimatuks privaatsustundlikes valdkondades ja jaotatud reaalsetes keskkondades. Lisaks puudub neil meetoditel sageli kontseptsioonitasandi interpreteeritavus ning nad nõuavad töömahukat käsitsi märgistamist, et tuvastada ja selgitada keerukates andmestikes tähenduslikke kontseptsioone. Kuigi hiljutised automaatsed lähenemised, nagu [1] (Ghorbani jt., NeurIPS 2019), on edendanud kontseptsioonide avastamise ja selgitamise automatiseerimist, ei käsitle need privaatsuse säilitamise, andmete detsentraliseerimise ega koostööl põhineva kontseptsioonide joondamise väljakutseid mitme osaleja vahel.

Käesolevas töös esitame **Federated Concept Alignment for Privacy-Preserving Explanations (FedCAPE)** — uue raamistiku, mis on loodud võimaldama mastaapset, privaatsust säilitavat ja täielikult detsentraliseeritud kontseptsioonide avastamist, joondamist ja interpreteeritavust. FedCAPE kasutab isejuhitud õppimist (DINO) ja OpenAI CLIP-i multimodaalseid võimalusi, et automaatselt määrata pildisegmentidele semantiline tähendus, kaotades seeläbi vajaduse käsitsi märgistamise järele ning lisades eraldatud tunnustele semantilise kihi. Oluliselt rakendab FedCAPE föderatiivset K-means klasterdamist, et ühiselt joondada ja täiustada avastatud kontseptsioone klientide vahel, tagades, et jagatud kontseptuaalne teadmine tekib ilma toorandmeid vahetamata.

Selle föderaalse lähenemise kaudu saavutab FedCAPE otsast lõpuni interpreteeritavuse, parema kontseptsioonide joondamise ja suurema läbipaistvuse mudeli ennustustes — ületades nii traditsioonilisi kui ka tipptasemel automaatseid lähenemisi privaatsuse säilitamise, mastaapsuse ja seletatavuse osas. Eksperimentaalne hindamine mitme jaotatud kliendi puhul näitas tugevat semantilist järjepidevust klientide vahel, kusjuures inimhindajad eelistasid FedCAPE klastri tulemusi juhuslikele võrdlusnäidetele 80–100% juhtudest. Kvantitatiivselt saavutas FedCAPE kõrgeid TCAV-skoore silmapaistvate kontseptsioonide puhul, ületades mõnel juhul tsentraliseeritud baastaseme, vältides samas üleklaasterdamist ja parandades klastrite puhtust.

Need tulemused toovad esile FedCAPE potentsiaali olla sillaks interpreteeritava tehisintellekti ja privaatsust säilitava masinõppe vahel jaotatud keskkondades

Võtmesõnad: Hajusõpe, seletatav AI, TCAV, mõistejoondus, andmekaitse

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Acknowledgments

I dedicate this work to the memory of my beloved father, whose courageous battle with cancer ended too soon. His strength, integrity, and unwavering belief in me continue to inspire and guide every step I take.

My deepest gratitude goes to my wife, Nesma, whose steadfast faith in me and constant encouragement have been my anchor throughout this journey. To my daughter, Haya, your light and laughter have been a source of joy and purpose, especially during the most challenging moments. I am profoundly thankful to my mother for her unconditional love, and to my siblings for their unfailing support and belief in my path.

I owe a special and heartfelt debt of gratitude to my supervisor, Dr. Radwa, whose unwavering support, insightful guidance, and patience have shaped the direction and quality of this work. Her ability to challenge my thinking while encouraging my independence has been instrumental in my academic growth. This thesis stands as much a reflection of her mentorship as of my own efforts.

I also wish to express my sincere appreciation to my managers, Randa and Mina, whose trust and support made it possible for me to embark on this academic chapter. Their encouragement created the space and opportunity I needed to pursue this degree.

Mahmoud

August, 2025

Contents

1. Introduction

The increasing deployment of artificial intelligence (AI) systems in domains such as healthcare, finance, autonomous driving, and law enforcement has intensified the demand for models whose decision-making processes can be understood and trusted by humans. In high-stakes scenarios, stakeholders—including regulators, domain experts, and end-users—require explanations not only to ensure fairness, safety, and compliance but also to facilitate model debugging and foster acceptance. This has led to the emergence of *Explainable Artificial Intelligence* (XAI) as a key research area, aiming to make the reasoning processes of complex models more transparent, interpretable, and accountable.

Deep neural networks have achieved remarkable success across a range of domains, often surpassing human-level performance in complex tasks such as image classification, speech recognition, and natural language processing. However, the opacity of these models—their characterization as "black boxes"—poses a fundamental challenge to trust, adoption, and regulatory acceptance, particularly in high-stakes applications.

Traditional interpretability techniques have primarily relied on *feature attribution*, either through model-agnostic methods such as LIME and SHAP, or model-specific approaches including Integrated Gradients, Layer-wise Relevance Propagation (LRP), and Grad-CAM [2–7]. In the vision domain, *pixel-wise saliency methods* highlight the influence of individual pixels using gradients [7] or relevance propagation [5], while region-based enhancements such as XRAI provide semantically coherent attributions [8]. Figure 1 illustrates an example output of a pixel-saliency method. Despite their utility, these methods often fall short of providing

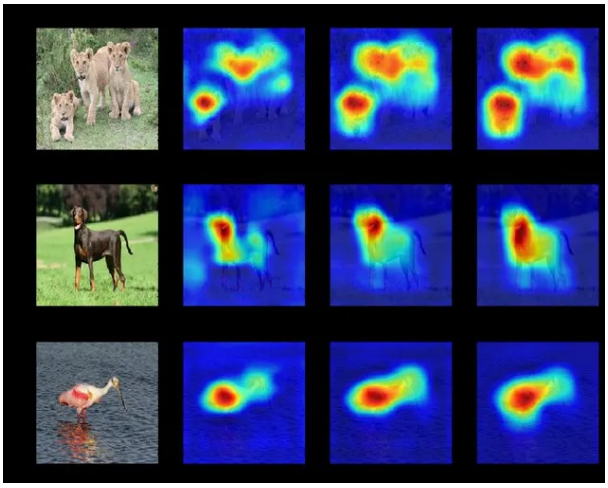


Figure 1. Pixel-saliency (source: medium.com/@bijil.subhash)

actionable or human-understandable insights. Their outputs are sensitive to perturbations, may lack semantic coherence, and often fail to align with the way humans naturally reason about decisions—sometimes even contradicting them.

Recent research has shifted towards *concept-based explanations*, which aim to interpret model behavior in terms of higher-level, human-meaningful concepts rather than low-level features. In this context, [9] introduced *Quantitative Testing with Concept Activation Vectors* (TCAV), a pioneering methodology that maps simple human knowledge into concepts. This approach, considered a paradigm shift, enables model explanations to be both global and inherently interpretable, offering users a summary of "what the model has learned" in terms of familiar, reusable knowledge. Figure 2 from [9] illustrates an example in which concepts such as "striped," "dotted," and "zigzagged" are used to interpret model predictions, showing alignment with human reasoning.

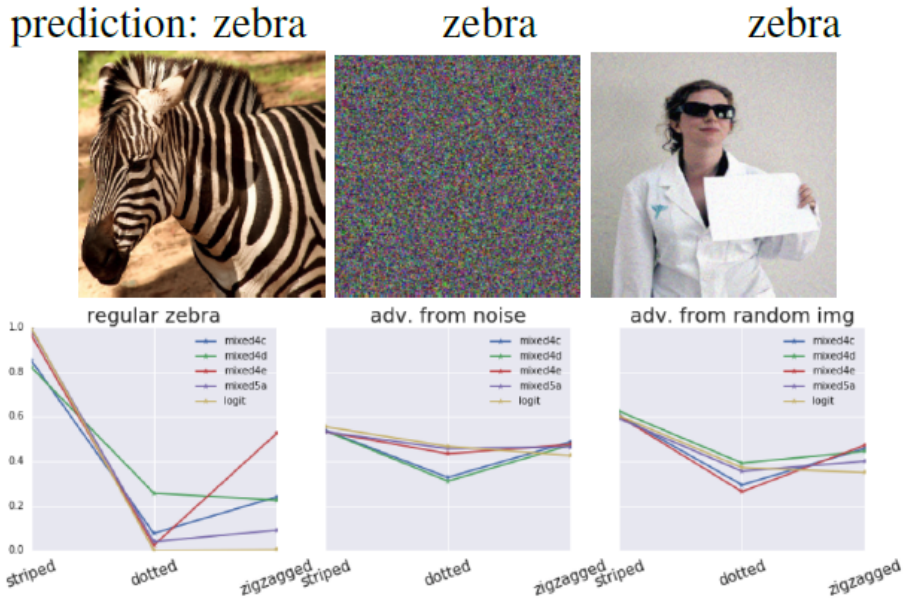


Figure 2. TCAV on Zebra classification model (source: Been Kim et al. [9])

Ghorbani et al., in *Towards Automatic Concept-based Explanations* (ACE) [10], extended TCAV to develop a framework that automates the explanation process by extracting the most meaningful, coherent, and essential concepts from an image, which can then be used to interpret model predictions. Figure 3 from [10] illustrates the concept extraction algorithm.

The process begins with a super-pixel segmentation at multiple resolutions from the same image, followed by clustering the segments and filtering out less relevant clusters, and concludes with computing the salient clusters using TCAV.

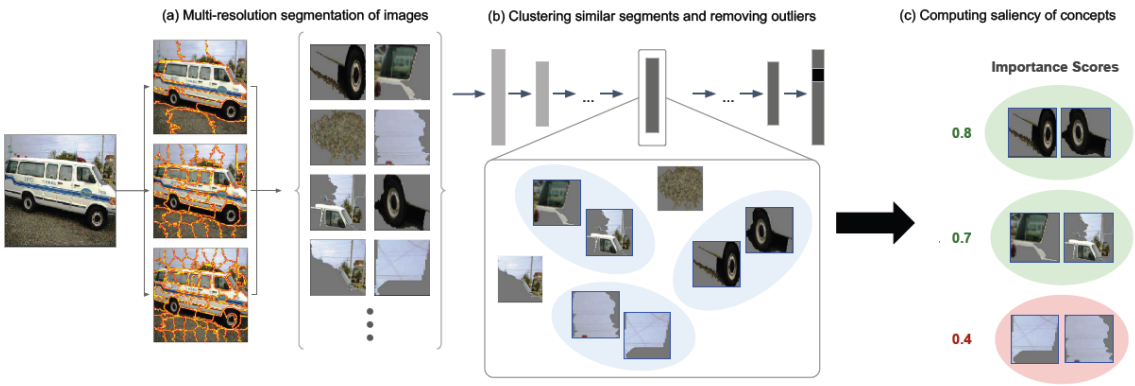


Figure 3. Concept extraction algorithm in ACE (source: [10])

Despite these advances, significant challenges remain. Most notably, current concept-based explanation frameworks assume centralized access to all data and models. In distributed environments, such as those in which image datasets are located across multiple sites, extracting and sharing concepts without compromising privacy becomes non-trivial. This limitation restricts the applicability of such frameworks in privacy-sensitive or federated learning scenarios.

In this thesis, to the best of our knowledge, no prior work has addressed the automatic extraction of concepts within a federated learning setting. Building upon the foundation laid by ACE and its successors, this work advances the state of concept-based explainability by enhancing robustness and scalability through the integration of recent advances in computer vision and machine learning.

At a high level, the proposed *Federated Concept Alignment for Privacy-Preserving Explanations (FedCAPE)* framework operates in three core stages. First, each client locally extracts concepts from images using SAM-based segmentation [11], CLIP-based semantic filtering [12], and DINO feature embeddings [13]. Second, a federated K-Means process aligns these locally discovered concepts into a shared global concept dictionary without exchanging raw data. Finally, TCAV is applied to quantify the importance of each concept for model predictions, enabling coherent, interpretable explanations across distributed clients while preserving privacy.

The main contributions of this thesis are as follows:

- **Integration of automatic concept-based extraction with federated learning:** Introducing the first framework that combines these two paradigms, enabling automatic concept extraction in privacy-sensitive, distributed environments.

- **Advanced segmentation pipeline:** Leveraging the Segment Anything Model (SAM) [11] for high-quality object segmentation and CLIP [12] for semantically meaningful filtering, producing cleaner and more relevant segments.
- **Universal feature extraction:** Utilizing DINO [13], a self-supervised Vision Transformer, to generate rich, universal feature representations for more contextualized concept labeling.
- **Federated K-Means clustering:** Implementing a privacy-preserving clustering method in the DINO activation space, ensuring that only feature embeddings—not raw data—are shared between clients.
- **Cross-client concept alignment:** Promoting consistent interpretation of concepts across clients despite heterogeneous local datasets, thus advancing the practical deployment of interpretable AI in real-world distributed systems.

By systematically addressing the limitations of existing methods, this thesis bridges the gap between the promise of concept-based interpretability and its reliable application in operational machine learning systems.

2. Background

This section reviews the foundational concepts, methods, and tools relevant to this thesis. It first introduces the notion of interpretability in machine learning, focusing on concept-based methods. Then it reviews key contributions in the literature, including human-in-the-loop concept learning, quantitative concept testing, and automated concept discovery.moreover, it concludes an overview of production-ready tools—such as SAM, CLIP, and DINOv2—and federated learning frameworks that support the proposed methodology.

2.1 Taxonomy of Interpretability Methods

Interpretability in machine learning refers to the extent to which a human can understand the cause of a decision or the internal logic of a model [14]. Practically, an explanation is an interface between an AI system and a human, conveying the model's reasoning in a form that a target audience can comprehend [15]. Building on Guidotti et al. [15] and recent surveys [16, 17], We organize the landscape of interpretability methods along four key dimensions [15, 16]:

(i) **Scope:** *Global* vs. *Local* — whether the explanation aims to describe the model's overall behavior or only specific predictions.

(ii) **Model dependency:** *Model-specific* vs. *Model-agnostic* — whether the method requires access to internal model structure or can operate on any black-box predictor.

(iii) **Timing:** *Transparent / Intrinsic* vs. *Post-hoc* — whether interpretability is built into the model itself or generated after training.

(iv) **Explanation format:** The representation of the explanation, such as rules, surrogate models, feature attribution, example-based reasoning, or concept-based representations.

Global vs Local Interpretability

Global interpetability aims to understand the model's overall behavior (e.g., a small decision tree whose logic can be audited end-to-end) [15].

Local interpretability explains individual predictions for specific instances (e.g., LIME and SHAP identify features most influential for one input) [18, 19].

Model-Specific vs Model-Agnostic Methods

Model-specific methods exploit internal structure (e.g., gradients and activations in neural networks, node splits in trees) [20, 21]. *Model-agnostic* methods treat the predictor as a black

box and use only input–output behavior [18, 19]; they are broadly applicable but may provide less structural insight.

Transparent vs Post-hoc Explanations

Transparent (intrinsic) models are interpretable by design—e.g., shallow decision trees, sparse linear models, or rule lists; here, the model itself is the explanation [22]. *Post-hoc* approaches explain an already-trained black box with surrogates, feature importance, saliency maps, textual or example-based rationales [15, 16].

Technique Families

Following on the post-hoc approaches main families which are:

- **Transparent models:** inherently interpretable predictors (decision trees, linear models, rule lists) offering global auditability.
- **Surrogate models:** simplified models (trees, rules) trained to mimic a black box globally or locally [18].
- **Feature attribution:** importance scores for inputs via perturbations (model-agnostic) [18, 19] or gradients (model-specific) [20, 21].
- **Rule-based:** extracting if–then rules from a trained model or learning rules as surrogates [15].
- **Example-based:** prototypes, criticisms, and counterfactuals to explain decisions with cases [23].
- **Concept-based:** explanations with human-recognisable concepts (e.g., TCAV [24] and ACE [10]) that quantify or discover concepts in latent spaces.

Evaluation Criteria

Consistent In line with prior surveys [15, 16, 22], interpretability methods are commonly assessed using the following key criteria:

- *Interpretability*: simplicity and human comprehensibility (e.g., few rules, shallow depth).
- *Fidelity*: agreement of the explanation (e.g., surrogate) with the black box's outputs.
- *Accuracy*: predictive performance of the interpretable model on the task.
- *Generalisability*: whether explanations are stable across inputs and robust to perturbations.
- *Ethical alignment*: fairness, privacy preservation, robustness, and accountability [16, 17].

A concise summary of these interpretability categories is shown in Table 1.

Table 1. Summary of interpretability method categories and properties.

Category	**Scope**	**Intrinsic / Post-hoc**	**Model-Agnostic?**	**Examples**
Transparent models	Global	Intrinsic	N/A	Decision trees, rule lists, linear/logistic regression [22]
Global surrogates	Global	Post-hoc	Yes	Decision-tree or rule surrogates of neural networks; rule extraction [15]
Local explanations	Local	Post-hoc	Yes	LIME, SHAP local attributions [18, 19]
Feature attribution	Local / Global	Post-hoc	Varies	Integrated Gradients, Grad-CAM, permutation importance [20, 21]
Example-based	Local / Global	Post-hoc	Yes	Prototypes, criticisms, counterfactuals [23]
Concept-based	Local / Global	Post-hoc or intrinsic	Typically specific	TCAV, ACE, concept bottleneck models [10, 24]

Positioning of FedCAPE

Within this taxonomy, **FedCAPE** belongs to the *concept-based* branch (post-hoc, model-specific with respect to internal representations), producing *global* concept alignment across decentralized datasets. It extends TCAV/ACE [10, 24] to federated settings while preserving privacy (ethical alignment) and supporting coherent, human-meaningful explanations at scale.

2.2 Related work

2.2.1 Learning Interpretable Concept-Based Models with Human Feedback

In many domains, particularly healthcare, there is a gap between the low-level features used by machine learning models and the high-level concepts understood by domain experts. Yeh et al. (2022) [25] address the challenges of model understanding and interpretability in domains characterized by high-dimensional, tabular, and count data, particularly emphasizing healthcare settings. In such contexts, raw data often consist of granular diagnostic or procedural codes (e.g.

prescriptions for specific medications or coded diagnoses). In contrast, clinicians and domain experts reason about patient status in terms of higher-level, abstracted conditions (e.g., ”anxiety” or ”hypertension”). The misalignment between machine-learned features and human mental models presents a significant barrier to adopting machine learning in clinical decision support.

A central limitation of many existing concept-based models is that they define concepts as functions of black-box classifiers, which may not align with user expectations or mental models. These black-box concepts risk encoding spurious correlations or perpetuating bias, and without tools for transparent concept interpretation, users may struggle to trust or meaningfully interact with the resulting explanations. Furthermore, prior work typically assumes the availability of ground-truth concept annotations for at least a subset of the data. This assumption is impractical in healthcare due to the high cost and time required for manual chart review. This contrasts with the image domain, where labeling is relatively fast and inexpensive.

To address these challenges, Yeh et al. [25] propose an interactive framework in which concept definitions are iteratively refined based on human feedback. Their approach allows users to label small numbers of exemplars and counter-examples, which the system then uses to improve the alignment of model-derived concepts with user intent. The method employs a probabilistic concept encoder $g(\cdot)$ that projects high-dimensional raw features x into a lower-dimensional concept space $c = g(x)$. Human feedback is used to guide the learning of g such that the resulting concept activations better reflect the semantics the user desires. The framework also introduces metrics to measure concept interpretability and predictive utility, supporting both quantitative and qualitative evaluation of concept quality.

Mathematically, given raw features x, the model learns:

$$c = g(x)$$

Where c is a vector of interpretable concept activations, the user provides feedback y_c (labels or corrections) for selected examples, and the model adjusts g to maximize the agreement between c and y_c under the supervision loss:

$$\mathcal{L}_{\text{concept}} = \text{CE}(y_c, g(x))$$

CE denotes the cross-entropy or another appropriate loss function, depending on the feedback type.

The added value of this work is twofold. First, it demonstrates that human-guided, interactive concept learning can substantially improve both the interpretability and accuracy of concept-

based explanations, even in complex, high-dimensional domains where ground-truth concepts are challenging to annotate. Second, it establishes a general, domain-agnostic template for integrating human feedback into concept-based model development, thereby setting the stage for user-centered, trustworthy AI in sensitive applications such as healthcare.

2.2.2 Quantitative Testing with Concept Activation Vectors (TCAV)

Traditional feature attribution methods, such as saliency maps and gradient-based explanations, can be complex for users to interpret, especially when reasoning about complex, high-level concepts rather than low-level feature activations, as they mainly highlight individual features or pixels that contribute most to a prediction. Which might not be aligned with human knowledge, Kim et al. [24] present a novel approach to model interpretability that addresses these limitations by Testing with Concept Activation Vectors (TCAV), which introduced to bridge this gap by allowing practitioners to quantify the influence of user-defined, semantically meaningful concepts on model predictions. The TCAV framework is particularly valuable in domains such as computer vision, where researchers and practitioners are often interested in understanding whether abstract properties (e.g., ”stripes,” ”wheels,” or ”texture”) play a significant role in the model's reasoning process. The method works by learning a *concept activation vector* (CAV) for each user-supplied concept, typically by training a linear classifier to distinguish examples of the concept from random counterparts within the internal activation space of a neural network.

A key strength of TCAV is its generality and user-centered flexibility: it can be applied to any differentiable model and supports arbitrary, human-defined concepts, requiring only sets of positive and negative examples for each concept. The directional influence of a concept is then tested by measuring the alignment between the CAV and the gradient of the output logit with respect to the internal activations. The importance of a concept C for predicting class k is quantified by the TCAV score, defined as the fraction of examples for which increasing the presence of the concept increases the logit for class k.

Formally, the TCAV score is computed as:

$$\text{TCAV}_{k,C} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left(\nabla h_k(f(x_i)) \cdot v_C > 0 \right)$$

where h_k is the logit function for class k, $f(x_i)$ is the activation of the internal layer for input x_i, v_C is the learned concept activation vector, and $\mathbb{I}(\cdot)$ is the indicator function.

TCAV enables practitioners to ask targeted, global questions such as ”To what extent does the concept of ‘stripes’ influence the model's prediction of a 'zebra'?” rather than being limited to

local feature attributions for individual samples. This capability has made TCAV foundational in the field of concept-based interpretability, fostering a shift toward model explanations that are more aligned with human cognition and semantics, and providing a robust, quantitative tool for testing model reliance on user-relevant concepts.

2.2.3 Towards Automatic Concept-based Explanations (ACE)

While prior approaches such as TCAV rely on human-supplied concept examples, Ghorbani et al. (2019) [10] address the challenge of automating concept discovery in complex data domains—particularly in computer vision. The need for scalable, label-free methods is driven by the recognition that many real-world applications, such as large-scale image classification, lack explicit annotations or easily obtainable examples for high-level, semantically meaningful concepts. Manual labeling is not only costly and labor-intensive but can also introduce subjective bias.

To overcome these limitations, ACE (Automated Concept-based Explanations) introduces an end-to-end pipeline that discovers and evaluates visual concepts directly from data, without human intervention. The method first segments input images at multiple resolutions, using algorithms such as SLIC superpixels, to produce a diverse set of candidate segments representing different visual scales and granularities. These segments are then embedded into a deep feature space (typically using internal activations from a pre-trained convolutional neural network) and clustered—often via k-means—to group visually similar segments together. Each resulting cluster is hypothesized to correspond to a meaningful visual concept.

The novelty of ACE lies in its use of TCAV to quantitatively evaluate the importance of each discovered concept for a given class prediction. By treating each cluster as a candidate concept and applying the TCAV framework, ACE determines which automatically discovered concepts are most influential in the model's decision-making process. This enables the construction of global, human-interpretable explanations that describe, for instance, "what concepts are most relevant for classifying a zebra."

Mathematically, for a given class k and discovered concept C, ACE computes the TCAV score as:

$$\text{TCAV}_{k,C} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left(\nabla h_k(f(x_i)) \cdot v_C > 0 \right)$$

where v_C is the concept activation vector for cluster C and $f(x_i)$ is the activation of the relevant network layer for segment x_i.

By automating concept extraction and evaluation, ACE significantly expands the applicability of concept-based interpretability, allowing explanations to be constructed in a data-driven manner without reliance on user-supplied annotations. The approach advances the field by providing a scalable solution for extracting coherent, human-meaningful concepts and systematically evaluating their influence on deep learning model predictions. In this thesis, we build upon the ACE methodology to develop the FedCAPE pipeline, which extends automatic concept-based explanations to federated and privacy-preserving settings.

2.2.4 Logical Reasoning–Based Explainable Federated Learning (LR-XFL)

Zhang and Yu (2024) [26] propose **LR-XFL**, a framework that integrates *logic rule–based* explanations into federated learning (FL) while maintaining strict data privacy. The method transforms symbolic reasoning from a passive explanatory tool into an active component of the FL optimisation process by:

(i) Extracting *local logic rules* from each client's data using an entropy-based neural network [27].

(ii) Combining these rules at the server via a logical connector (-AND- $\wedge$ or -OR- $\vee$) selected automatically from client statistics.

(iii) Using the *explanatory utility* of a client's rules to determine its weight in model aggregation.

Motivation and challenges

In standard FL, the server receives only model updates, not raw data, making it challenging to build global explanations. Earlier concept-based methods such as TCAV [24] or ACE [10] quantify or discover concepts but assume a centralised setting. LR-XFL adapts reasoning over concepts to FL, with two explicit goals:

- Generate *global*, human-auditable logic rules without sharing data.
- Use those rules to directly influence the federated aggregation process.

Local rule induction

On each client, an entropy-based linear layer [27] produces sparse, binary concept activations. These are translated into literals (f_i or $\neg f_i$) and combined into *sample-level* rules of the form:

$$r \equiv f_{i_1}^{\sigma_1} \wedge f_{i_2}^{\sigma_2} \wedge \cdots \wedge f_{i_\ell}^{\sigma_\ell},$$

where $\sigma_j \in \{+, -\}$ indicates whether the literal is positive or negated. For each class, sample rules can be merged into a class-level rule, but naïvely joining all with OR can overgeneralise, and always using AND can be too restrictive.

Co-occurrence matrix

To analyse the relationships between features in the rules, each client constructs a *Positive co-occurrence matrix* $\mathbf{P} \in \mathbb{N}^{n \times n}$, where n is the number of features. Each entry p_{ij} counts the number of rules in which features f_i and f_j appear together as positive literals. Similarly, a *negative co-occurrence matrix* $\mathbf{Q}$ is built, where q_{ij} counts the number of rules containing f_i with the negation $\neg f_j$. These matrices capture feature dependencies and potential conflicts.

Automatic connector selection

LR-XFL chooses between $\wedge$ and $\vee$ based on two statistics derived from the co-occurrence matrices:

(i) **Diagonality** — measures the degree to which features appear *in isolation*, i.e., how dominant self-co-occurrence p_{ii} is compared to all co-occurrences:

$$D = \frac{\sum_{i=1}^{n} p_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij}}.$$

A high D indicates mutual exclusivity between features, suggesting that an $\vee$ connector may better capture the decision logic.

(ii) **Exclusivity** — measures the extent of *negative co-occurrence* between features, indicating logical conflict:

$$E = \frac{\max\limits_{i \in \{1,\ldots,n\}} \left(\sum_{j=1}^{n} q_{ij} \right)}{\sum_{m=1}^{M} l_{r_m}},$$

where M is the number of rules and l_{r_m} is the length (number of literals) of rule r_m. A high E suggests that many features co-occur with the negation of others, again favouring $\vee$.

If $D > 0.9$ or $E > 0.8$ (thresholds tuned on validation data in [26]), the client recommends $\vee$; otherwise it recommends $\wedge$. The server determines the *global* connector via majority vote across clients for each class.

Global rule synthesis

The server receives candidate rules $r_{c,k}$ for each class c from clients k whose local models exceed an accuracy threshold. Using the fixed global connector, it applies *beam search* [28] to select a compact subset of rules that maximises rule accuracy on a server validation set.

Rule-weighted aggregation

Clients are assigned aggregation weights proportional to the number of their rules selected into the global rule set:

$$w_k = \frac{p_k}{\sum_{i=1}^{K} p_i},$$

where p_k is the number of rules from client k included in the global set. Clients with no selected rules are assigned $w_k = 0$, reducing the influence of noisy or unhelpful contributors.

Performance and robustness

Across MNIST (even/odd), CUB, V-Dem, and MIMIC-II, LR-XFL achieved on average:

- **+1.19%** model accuracy
- **+5.81%** rule accuracy
- **+5.41%** rule fidelity

compared to a FedAvg+logic baseline. Under noise injection (random label shuffling in up to 60% of clients), LR-XFL's rule accuracy remained relatively stable and performance degraded more gracefully than baselines such as DDT [29] and FedAvg-Logic.

Relation to FedCAPE

LR-XFL assumes that input features are already semantically meaningful (conceptually labelled), which is not always true in real-world applications. In contrast, the **FedCAPE** framework developed in this thesis automatically performs *concept discovery and extraction*, thus extending LR-XFL–style logic-based interpretability to raw, unlabeled data.

2.2.5 Concept-Guided Interpretable Federated Learning (FedCBM)

Yang and Long (2023) [30] propose **FedCBM**, a federated *concept bottleneck model* that embeds interpretability directly into the training pipeline rather than applying it post-hoc.

Core Idea

FedCBM follows the *concept bottleneck paradigm* [31]: instead of learning arbitrary latent features, each model is trained to first predict a set of human-understandable concepts, and then derive the final class prediction from these concept activations. This ensures that every prediction can be explained in terms of concept presence and influence.

In the federated setting, a major challenge arises: due to data heterogeneity, clients may represent the same concept inconsistently. FedCBM addresses this through **shared Concept Activation Vectors (CAVs)**, which act as a global semantic basis. Each client maps its internal

representations to this common concept space, ensuring that, for example, “concept 1” has the same meaning across all clients.

Training Objective

Each client k minimizes a joint loss:

$$\mathcal{L}_k = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{concept}}, \tag{1}$$

where $\mathcal{L}_{\text{cls}}$ is the standard classification loss (e.g., cross-entropy) between predicted labels and ground truth, $\mathcal{L}_{\text{concept}}$ enforces accuracy of predicted concept activations, and λ balances the two terms.

A concept alignment step ensures the client’s concept representations align with the *shared* CAVs. These CAVs are updated on the server using a FedAvg-style aggregation:

$$v_g = \frac{1}{K} \sum_{k=1}^{K} v_k, \tag{2}$$

where v_k are the client-specific CAVs and v_g is the global shared vector.

Interpretability and Robustness

The shared CAV mechanism ensures:

- **Consistent semantics** across clients, enabling coherent explanations in non-IID settings.
- **Faithful explanations** at the concept level, since predictions are explicitly routed through concept activations.

Empirically, FedCBM improves both accuracy and interpretability on federated benchmarks, and demonstrates resilience to heterogeneous data partitions.

Position Among Concept-Based XAI Methods

FedCBM extends centralized concept bottleneck models [31] to federated learning, solving the semantic alignment problem that arises in distributed environments. Unlike post-hoc concept attribution methods such as ACE [10] or TCAV [24], FedCBM enforces interpretability *by design* through architectural constraints.

Relation to FedCAPE

Similar to LR-XFL, FedCBM assumes that a set of predefined, human-understandable concepts is available. In contrast, **FedCAPE** (proposed in this thesis) operates without concept annotations, automatically performing *concept discovery,extraction* from raw data. In this sense, FedCAPE can serve as a precursor to FedCBM, supplying the concept banks that FedCBM could then align across clients for interpretable federated reasoning.

2.3 Utilized Technologies and Frameworks

This section introduces the main models and tools used in the implementation of the proposed framework. Each tool was selected based on its performance, suitability for the task, and integration potential within a federated, concept-based explainability pipeline.

2.3.1 Segment Anything Model (SAM2) for Image and Video Segmentation

The Segment Anything Model (SAM2) [11] represents a transformative step in the field of universal image and video segmentation. SAM2 is designed to generalize across an extensive range of object segmentation tasks, utilizing large-scale pre-training to robustly identify and delineate objects, parts, and regions within diverse visual contexts. A key advancement of SAM2 is its zero-shot capability: the model can generate accurate segmentations for new, unseen objects or scenes without task-specific fine-tuning. This universality addresses a significant bottleneck in segmentation—the need for model retraining or domain adaptation for each new use case.

For the development of concept-based explanations, SAM2 offers critical utility as a generic and high-quality segmentation backbone. Its ability to produce flexible, high-granularity segment proposals provides a rich set of candidates for downstream concept discovery, clustering, and semantic labeling, regardless of the specific dataset or domain. Integrating SAM2 into explainability pipelines not only enhances the granularity of identified visual concepts but also ensures that discovered concepts are grounded in robust, adaptable segmentations. The official implementation is openly available for research use [32].

2.3.2 CLIP: Learning Transferable Visual Models From Natural Language Supervision

CLIP (Contrastive Language-Image Pre-Training) [12] bridges computer vision and natural language processing by training models to jointly embed images and text into a shared semantic space. This approach overcomes the bottleneck of fixed-category, label-dependent visual models by allowing open-vocabulary, language-guided understanding of visual data. In practice, CLIP is pre-trained on a large corpus of image-caption pairs, enabling the model to learn generalizable associations between textual descriptions and visual features.

Within explainable AI and concept discovery pipelines, CLIP provides a powerful mechanism to semantically score or filter image segments or clusters, assessing their correspondence to natural language concepts without the need for manual labeling. This enables flexible,

language-driven exploration of discovered visual patterns and supports more intuitive, human-aligned explanations. The official codebase for CLIP is available for integration and further experimentation [33].

2.3.3 DINOv2: Learning Robust Visual Features without Supervision

DINOv2 [13] advances self-supervised learning for vision transformers, enabling the extraction of universal, robust visual features without reliance on annotated datasets. DINOv2 employs self-distillation and a contrastive learning objective to train models that capture semantic and structural regularities in images, facilitating transferability across downstream visual tasks.

For concept-based explainability, DINOv2's high-quality, context-rich feature embeddings form an ideal basis for both fine-grained segmentation and meaningful clustering of image regions. The use of DINOv2 features enhances the coherence and discriminative power of discovered concepts, supporting automated pipelines such as with stronger, more generalizable representations. The official DINOv2 implementation is available for research use and reproducibility [34].

2.3.4 Federated Machine Learning for Explainable AI

Federated Machine Learning (FML) [35, 36] has emerged as a transformative paradigm for collaborative training of machine learning models across distributed, often heterogeneous, data silos. The primary motivation for FML is to enable institutions or devices to jointly contribute to model improvement without the need to centralize or share sensitive raw data, thus upholding privacy, regulatory compliance, and data sovereignty. This is particularly salient in sectors such as healthcare, finance, and telecommunications, where data cannot be freely exchanged due to legal and ethical constraints.

FML Paradigms and Architectures

FML encompasses a variety of architectures including: Horizontal federated learning where clients share the same feature space but have different samples. Vertical federated learning different feature spaces but overlapping samples. Federated transfer learning different feature and sample spaces. The standard workflow typically involves each client—which can range from large organizations (”cross-silo” federated learning, e.g., hospitals, banks, or enterprises) to individual edge devices (”cross-device” federated learning, e.g., smartphones or IoT sensors)—training a local model on its own data. Periodically, each client shares model updates (such as gradients or parameters) with a central server, which securely aggregates these updates and broadcasts the improved global model back to all participating clients. Crucially, at no

point is raw data transmitted between parties, thereby preserving data privacy and compliance. Cross-silo FL is often used for collaborations among a small number of institutions with rich datasets, while cross-device FL enables large-scale learning across millions of devices with limited local data.

Challenges in Federated Explainability.

Integrating explainable AI (XAI) within the federated learning framework introduces a new set of challenges and opportunities:

- **Distributed Concept Discovery:** In concept-based explanations, such as those generated by ACE or TCAV, concepts may be distributed or fragmented across clients, each holding only a partial view of the underlying data distribution. Ensuring that discovered concepts are coherent, representative, and semantically aligned across clients requires robust methods for federated clustering, concept alignment, and aggregation.
- **Privacy-Preserving Feature Aggregation:** Achieving meaningful interpretability while maintaining privacy constraints necessitates the development of protocols for aggregating features or concepts without exposing sensitive information. Approaches such as differential privacy, secure multi-party computation, and homomorphic encryption have been proposed to further mitigate risks of data leakage during model or concept aggregation.
- **Heterogeneity and Non-IID Data:** Real-world federated scenarios often involve clients with highly heterogeneous and non-IID (not independently and identically distributed) data, leading to challenges in aligning explanations and ensuring that global concept dictionaries or saliency maps remain consistent and reliable.
- **Scalability and Communication Efficiency:** Federated deployments at scale require efficient communication protocols and the ability to operate under limited network bandwidth, especially in edge or mobile settings.

Federated Learning Frameworks and Vendor Solutions.

Several software frameworks have been developed to support FML at scale, each with varying levels of support for explainable AI workflows:

- **Flower** [37]: An open-source, extensible federated learning framework that is model-agnostic and supports flexible client-server orchestration. Flower is particularly suited

for research on federated concept discovery and explainability, as it enables customized aggregation logic, client selection, and evaluation routines. Its integration with popular machine learning libraries (e.g., PyTorch, TensorFlow) allows rapid prototyping of federated explainable AI pipelines.

- **TensorFlow Federated (TFF)**: Developed by Google, TFF is designed for experimentation and deployment of federated learning algorithms, especially in the context of mobile and edge devices. TFF provides primitives for secure aggregation and supports integration with TensorFlow's explainability tooling.
- **PySyft**: An open-source framework that emphasizes privacy-preserving machine learning using federated learning, secure multi-party computation, and differential privacy. PySyft is particularly useful for research into privacy guarantees in explainable AI.
- **FedML**: A research-oriented federated learning library that supports edge, mobile, and cloud environments. FedML offers plug-and-play algorithms for federated averaging, personalized learning, and supports real-world benchmarking.
- **Meta's ExecuTorch and NVIDIA FLARE**: Recently announced collaboration delivering a three-tier hierarchical architecture for large-scale cross-device federated learning. ExecuTorch enables efficient on-device inference and training (used across billions of devices in Meta's Instagram, WhatsApp, Messenger) using a light-weight variation of SAM, called **SqueezeSAM**, for segmentation. NVIDIA FLARE orchestrates massive-scale device fleets in the federated process, abstracting on-device complexity and enabling hierarchical communication across millions of endpoints [38]
- **Scaleout FEDn** [39]: An open-source federated learning framework designed by Scaleout Systems, FEDn focuses on robust, scalable, and production-ready cross-silo federated learning. Its architecture features a relay-based aggregation protocol that efficiently handles communication between distributed silos (clients) and supports deployment in diverse and regulated environments such as healthcare, finance, and manufacturing. FEDn is especially noted for its horizontal scalability, reliability, and integration with container orchestration platforms (e.g., Kubernetes), making it suitable for large-scale, real-world deployments where strict data privacy and compliance requirements are in place. Official tools and documentation are available for researchers and industry practitioners alike.

Federated machine learning holds significant promise for deploying interpretable AI at scale in multi-institutional and privacy-sensitive settings. However, realizing this promise requires addressing complex challenges around privacy, heterogeneity, and semantic alignment of explanations. The integration of FML frameworks like Flower with concept-based explanation pipelines is a key enabler for scalable, trustworthy, and human-aligned AI in real-world applications.

3. FedCAPE Design and Implementation

This section presents the design and step-by-step implementation of the FedCAPE algorithm. It outlines the core components, underlying methodologies, and integration details that enable privacy-preserving, concept-based extraction in a federated learning environment.

3.1 FedCAPE Design

Ghorani et al. [1] presented a framework, which is considered a bedrock for FedCAPE core design

Federated Concept-Based Extraction Algorithm automates the discovery and quantification of concepts within distributed client image datasets, as illustrated in Figure 4, enabling evaluation of how these concepts influence a model's predictions for specific classes. It processes a pre-trained classifier (e.g., Inception V3) alongside labeled images from multiple known classes to generate class-specific concept explanations in a federated learning environment.

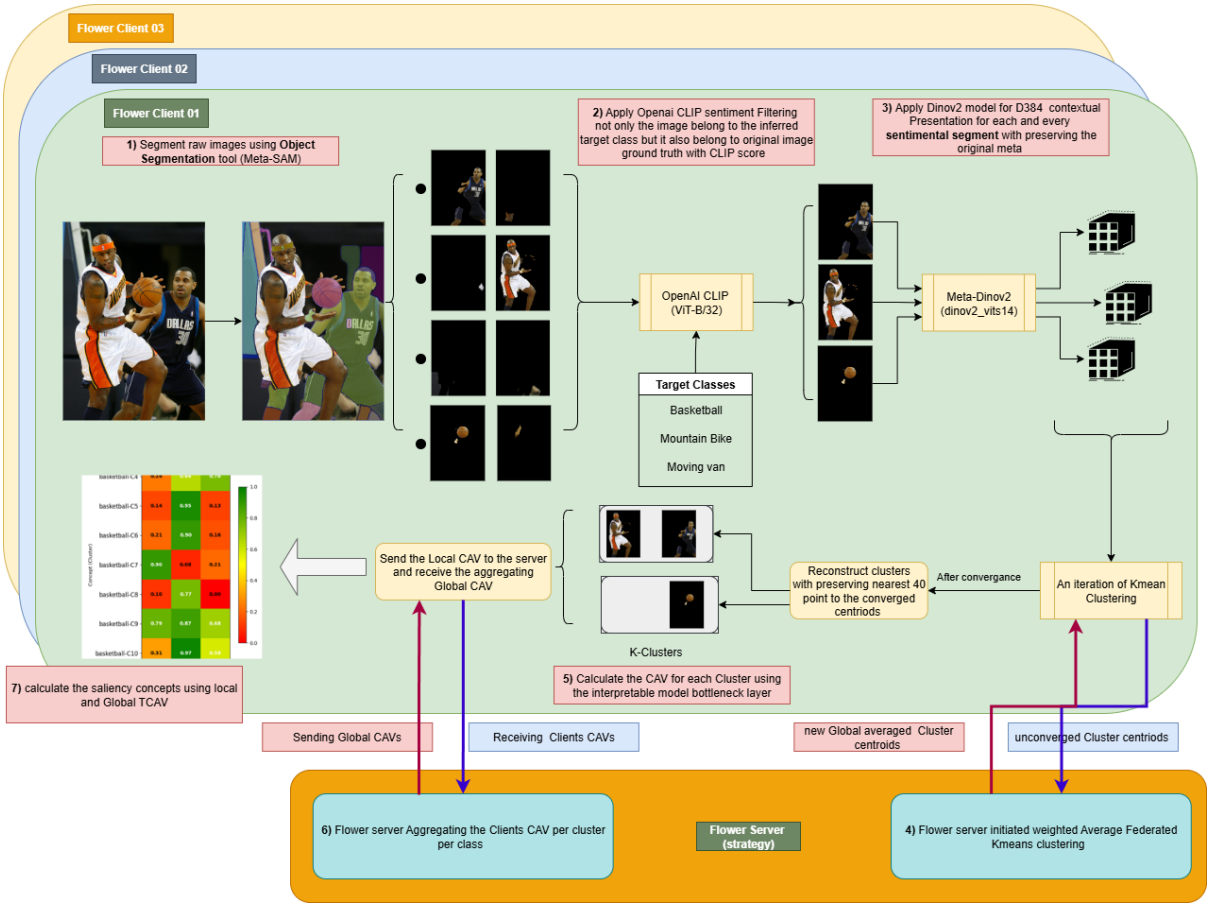


Figure 4. FedCAPE Algorithm

Step 1 Each client starts by performing object segmentation on every image, dividing it into multiple regions using a high-quality segmenter, such as SAM2, aiming to capture diverse concepts, from fine texture to full objects.

Step 2 CLIP-based Segment Filtering, each segment is classified using a pre-trained CLIP model comparing its features to a set of target classes (’basketball’, ’moving van’,’ mountain bike’), only segments that cross the CLIP similarity threshold to any of the target classes will be retained. To reduce the false positives, like the existence of the segments that belong to ”mountain bike” in the ”basketball”, we make sure that only the predicted CLIP class is the same as the ground truth. This last step is important to ensure that the class clusters are not polluted with other classes and that the class clusters are semantically aligned.

Step 3 Retained semantically aligned Segments are transformed and processed to extract features using a vision backbone SSL model Dinov2. Features are batched, normalized, and stored alongside labels and metadata for clustering preparation.

Step 4 Federated K-Means clustering begins with the Flower server instantiating a random centroid for each predefined cluster in every class, which is then distributed across the flower cluster registered clients. Iteratively, the clients update cluster assignments and local centroids for their own segment features, sending local statistics to the central server. The server aggregates these local centroids and updates the global centroids until convergence, measured by the shifts of centroids between rounds. The process continues until cluster assignments are stable across clients.

Step 5 Cluster Assignment and Curation, Upon convergence, each client assigns every segment to its nearest cluster centroid. For each cluster, the closest N=40 segments are retained and sorted with their CLIP scoring to ensure high cluster purity and coherence.

Step 6 Concept Activation vector Calculation [40]. Using the pretrained classifier Inception v3 that needs to be interpreted and understood, for each class cluster/Concept, CAV is extracted as [1] from the mixed_8 layer bottleneck layer, such as the logical regression model, where positive samples correspond to cluster members and negative samples are randomly selected from segments outside the same class cluster or from clusters of other classes.

Step 7 Each client sends the computed CAVs for every cluster across classes to the server, which aggregates them to construct a global CAV as a weighted mean for the received CAV. The

weight assigned to each cluster CAV is proportional to its size, giving larger clusters greater influence on the resulting global CAV.

Step 8 TCAV Scores are computed to know what the salient class concepts are for the client dataset. For each test image of each class, the gradients are taken from the same bottleneck layer of the Inception v3 model. The proportion of test samples where the concept direction positively contributes to class prediction is the concept TCAV score.

Algorithm 1 FedCAPE: Client-side Pipeline

Require: Local dataset D_c with class labels Y, target classes T, CLIP threshold τ (Sec. 3.3)

Ensure: Local clusters $\{C\}$, local CAVs $\{v_C\}$, global CAVs $\{v_C^{\text{glob}}\}$, TCAV scores $S^{\text{local}}, S^{\text{glob}}$

1: **for** each image $x \in D_c$ **do**
2: $S_x \leftarrow$ SAM2_SEGMENT$(x, \texttt{min_mask_region_area})$ (Sec. 3.3)
3: $S'_x \leftarrow \{s \in S_x \mid \text{CLIP\_SIM}(s, T) \geq \tau \ \wedge \ \text{PRED\_CLASS}(s) \in Y\}$
4: $F_x \leftarrow \{\text{DINO\_EMBED}(s) \mid s \in S'_x\}$ (Sec. 3.3)
5: **end for**
6: **Group extracted features/metadata by class label** ▷ prep for class-wise clustering
7: **for** each class $k \in T$ **do**
8: $X_k \leftarrow \{f \mid f \text{ labeled as } k\}$
9: Assign X_k to global centroids from server (K clusters per class)
10: Update local centroids μ_k^{local}
11: **end for**
12: Send $\{\mu_k^{\text{local}}, \text{counts}\}$ to server; receive updated $\{\mu_k\}$
13: Repeat lines 7–11 until server signals convergence (δ tolerance) ▷ *— After K-means convergence —*
14: **for** each cluster C **do**
15: Retain top-N closest segments to μ_C (Sec. 3.3)
16: Train logistic regression to get local CAV v_C (Sec. 3.3)
17: **end for**
18: **Send** $\{(v_C, \ |C|)\}$ **to server for global aggregation**
19: **Receive broadcast** $\{v_C^{\text{glob}}\}$ **from server**
20: **for** each class k **do**
21: Compute local TCAV S_k^{local} using $\{v_C\}$ and test protocol (Sec. 3.3)
22: Compute global TCAV S_k^{glob} using $\{v_C^{\text{glob}}\}$ and same protocol
23: **end for**

Algorithm 2 Server-side Federated K-Means & Global CAV Aggregation (per class)

Require: K clusters per class, initial centroids $\mu_k^{(0)}$, tolerance δ (Sec. 3.3)

Ensure: Converged global centroids $\mu_k^{(*)}$ and aggregated global CAVs $\{v_C^{\text{glob}}\}$

$t \leftarrow 0$

repeat

  Broadcast $\{\mu_k^{(t)}\}$ to all clients

  Collect $\{(\mu_k^c, n_k^c)\}$ from all clients

  for each class k and cluster j **do**

    $\mu_{kj}^{(t+1)} \leftarrow \dfrac{\sum_c n_{kj}^c \mu_{kj}^c}{\sum_c n_{kj}^c}$ ▷ weighted average

  end for

  $\Delta \leftarrow \max_{k,j} \|\mu_{kj}^{(t+1)} - \mu_{kj}^{(t)}\|_2$

  $t \leftarrow t + 1$

until $\Delta < \delta$ ▷ converged K-means centroids

Switch to CAV phase: instruct clients to send local CAVs

Collect per-concept $\{(v_{kj}^c,\ n_{kj}^c)\}$ from all clients

for each class k and cluster j **do**

  $\tilde{v}_{kj} \leftarrow \dfrac{\sum_c n_{kj}^c\, v_{kj}^c}{\sum_c n_{kj}^c}$ ▷ weighted average of CAVs

  $v_{kj}^{\text{glob}} \leftarrow \tilde{v}_{kj} / \|\tilde{v}_{kj}\|_2$ ▷ normalize direction

end for

Broadcast $\{v_{kj}^{\text{glob}}\}$ to all clients for *global* TCAV scoring

(Optionally) terminate when all clients confirm global TCAV completion

Algorithm 3 CAV and TCAV Computation

Require: Cluster C, negatives N_C, classifier $h(\cdot)$, layer ℓ (Sec. 3.3)

Ensure: CAV v_C, TCAV scores per class k

Train logistic regression g (config in Sec. 3.3) on positives from C vs negatives from N_C
$v_C \leftarrow \text{normalize}(g.\text{coef})$
for each class k **do**
    for each test image x of class k **do**
        $a \leftarrow \text{ACTIVATIONS}(x, \ell)$
        $s(x, k, C) \leftarrow \text{sign}(\nabla_a h_k(x) \cdot v_C)$
    end for
    $\text{TCAV}_{k,C} \leftarrow \frac{1}{|X_k|} \sum_x [s(x, k, C) > 0]$
end for

3.2 FedCAPE System and Implementation details

3.2.1 High-level Architecture

FedCAPE is implemented -[41]- as a modular, privacy-preserving, distributed system using the Flower federated learning framework. Its execution on the Tartu HPC cluster leverages multiple GPU/CPU nodes and distributed actors for efficient large-scale processing. The system architecture is illustrated in Figure 5.

- **Data Preprocessing Pipeline**: As illustrated experimentally, the Data Preprocessing pipeline is handled by a main function outside the Flower cluster. To reduce the Memory footprint on the simulation, the Ray Flower cluster clients it is responsible for preparing the data for every client in a good shape, as follows.
  - Fetch the available classes and their associated data images and apply the needed transformation on them (resize, mean, and standard deviation).
  - Images are segmented with SAM2, filtered using a ViT-based CLIP model, then converted to DINOv2 features, of course, with tracing the meta for each segment.
  - Retained semantic segments, with their metadata and feature representations, are saved efficiently for reuse by the flower clients.
- **Federated Learning Platform (Flower)**: The flower simulation is the platform used for simulating the needed number of clients, and it is composed of the following:

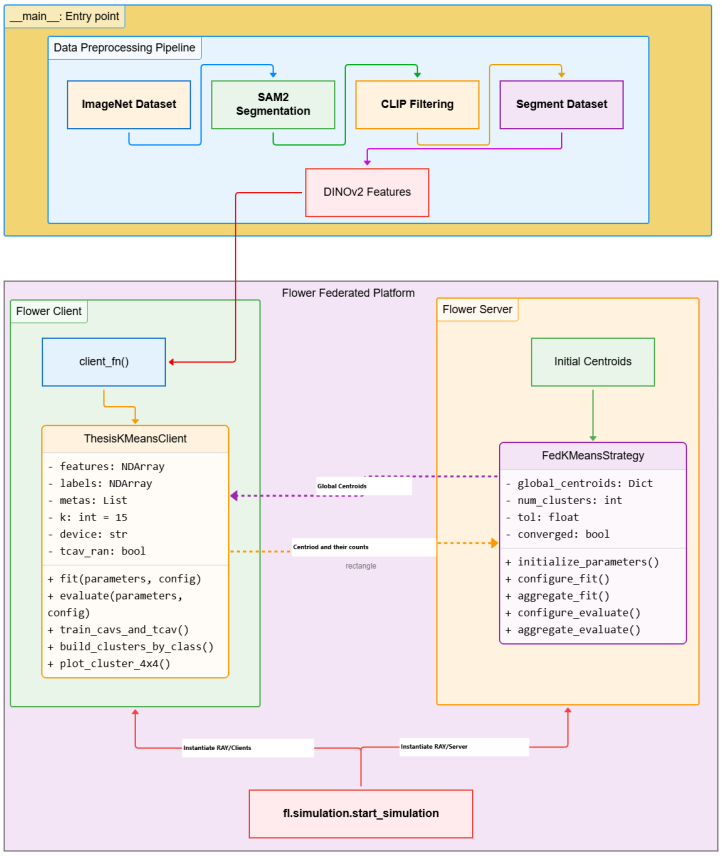


Figure 5. FedCAPE HL-Architecture

– The central *Flower server* instantiates the `FedKMeansStrategy`, which maintains global and local centroids for each class on each client and coordinates federated K-Means clustering and their convergence.

– Each *Flower client* loads its precomputed features, labels, and segment metadata (from disk) using `client_fn()` and instantiates the `ThesisKMeansClient`, which manages local clustering, executes local K-Means steps, and shares only aggregated centroid updates/statistics with the server. CAV and TCAV logic locally.

– The Flower orchestrator (`fl.simulation.start_simulation`) manages process launch, resource allocation (CPUs, GPUs), and inter-process communication between the cluster components.

3.3 Hyperparameter Selection and Justification

The following hyperparameters were chosen based on both our exploratory experiments and [10] author's exploratory experiments as well, and empirical evaluation on validation data to balance accuracy, interpretability, and communication efficiency.

Number of clusters per class (K)

We evaluated $K = \{15\}$ for each class, selecting the value that maximized the mean silhouette score in the DINOv2 embedding space while ensuring that at least $N \geq 40$ high-purity members were retained in each cluster. This threshold provided sufficient positive samples for CAV training without diluting semantic purity.

CLIP similarity threshold (τ) We selected the smallest τ that might ensure semantically meaningful segments. This choice minimizes off-class pollution while preserving diversity within-class visual parts.

SAM2 segmentation parameters.

In addition to `min_mask_region_area`, which was set between 0.5%–5% of the image area depending on the target object size (see above), several other SAM2 hyperparameters were tuned for optimal concept discovery:

- **`pred_iou_threshold`**: Set to 0.88. This controls the minimum predicted IoU score for retaining a mask. A relatively high value was chosen to ensure that retained masks are geometrically well-aligned with the predicted object boundaries, reducing noise in fine-grained concepts such as "basketball net threads" or "bike spokes." Practical tests showed that lowering this threshold increased recall but introduced many overlapping or partial masks, which degraded CLIP filtering precision.
- **`stability_score_threshold`**: Set to 0.8. This threshold filters out masks whose boundaries are unstable under small perturbations of the mask generation process. A high stability score ensures robustness of the concept regions, which is critical because unstable masks often produce inconsistent DINOv2 embeddings and hurt clustering purity.
- **`min_mask_region_area`**: Set adaptively between $256 - 512$ for 224 by 224 image. Smaller thresholds were used for datasets where fine parts (e.g., textures, handles, logos) were semantically relevant, while larger thresholds filtered out uninformative background

details. This adaptive setting provided the best trade-off between recall of fine details and noise exclusion.

These parameter values were not arbitrarily selected; they were chosen after iterative spot-checking on validation images and measuring their downstream effect on cluster purity (silhouette score) and CAV consistency. The goal was to ensure that retained segments are (i) semantically meaningful, (ii) geometrically stable, and (iii) non-redundant, so that the subsequent CLIP filtering and DINOv2 feature extraction stages operate on high-quality candidate concepts.

Federated K-Means convergence tolerance (δ)

The server–client centroid update loop was terminated when the maximum centroid shift across all classes and clusters fell below $\delta = 0.5$ in normalized DINO space.

Cluster curation size (N)

The $N = 40$ nearest segments (ranked by combined cosine similarity to the centroid and CLIP score) were retained for each converged cluster. Basically, the numbers were taken from the referenced paper [10], but empirically, it shows a good one in terms of concept similarity and the CAV efficiency.

CAV training configuration

We used logistic regression with class-balanced weights. Positives were all curated members of a cluster, while negatives were drawn in a $1{:}1$ ratio from (i) other clusters of the same class and (ii) clusters from other classes.

CAV Embeddings Such a configuration, we meant to be aligned with [10], we used the Inception v3 pretrained model with layer mixed_8 to explain.

4. Experiment Results and Evaluation

This chapter presents the experimental results of the FedCAPE framework, highlighting both qualitative and quantitative evaluations. On the qualitative side, it incorporates a pre-designed user study questionnaire (as described in Section 4.2.3) to assess whether the discovered concepts are meaningful and coherent from a human perspective. This is complemented by visual inspections of concept clusters, attention overlays, and projection plots, illustrating how the extracted concepts align with semantically relevant parts within each class.

On the quantitative side, the analysis applies Testing with Concept Activation Vectors (TCAV) to evaluate the importance of each concept, alongside metrics such as clustering compactness and class separability. TCAV was chosen because FedCAPE, like the Automated Concept-based Explanation (ACE) method by Ghorbani et al., aims to discover high-level human-interpretable concepts and assess their causal influence on model predictions. In ACE, TCAV served as the standard quantitative tool to validate whether discovered clusters correspond to meaningful and predictive concepts; we adopt the same principle here to ensure comparability with established concept-based interpretability research.

Together, these evaluations—both the planned human study and the quantitative metrics—demonstrate FedCAPE's ability to learn robust, interpretable concepts from decentralized data, while also identifying its limitations and guiding directions for further improvement.

4.1 Clients Data

4.1.1 Dataset Distribution and Heterogeneity

In federated learning, client datasets are often distributed in a **non-IID** (non-independent and identically distributed) manner, which can introduce significant class imbalance and diversity across clients. Such imbalances can negatively affect model convergence, fairness, and generalization. To mitigate these confounding effects and to establish a controlled baseline, we ensure that the data is nearly **IID**, which is evenly partitioned across all participating clients in our experimental setup.

Figure 6 shows the class-wise label distribution for each client. Each stacked bar represents the percentage of samples from the "basketball," "mountain bike," and "moving van" classes within each client's local dataset. As illustrated, the class distribution is virtually identical across all three clients. This controlled partitioning ensures that any observed differences in federated

training outcomes are not attributable to class imbalance, but instead to the learning algorithms or system configurations under investigation.

It is important to note, however, that an equal partitioning of class labels does not guarantee an identical distribution of **concepts** across clients. Variability in underlying data concepts may still introduce subtle heterogeneity that could affect model learning and interpretation even within the same class.

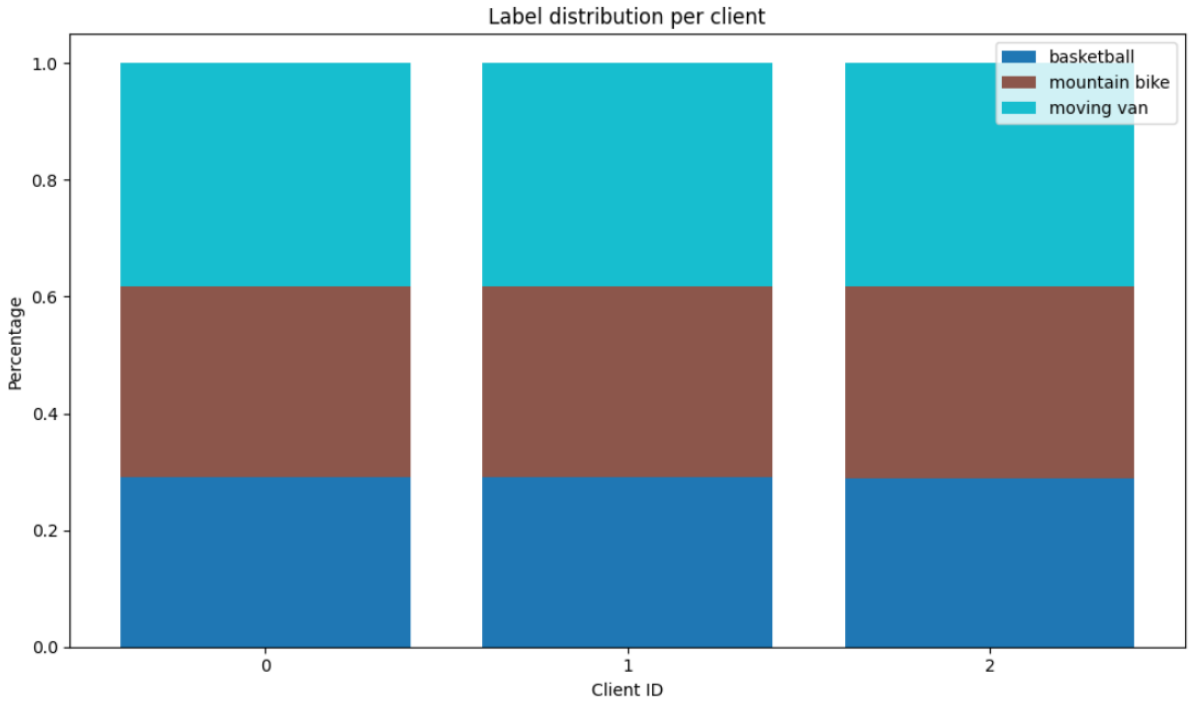


Figure 6. Clients Dataset Distribtuion

4.1.2 Local Segmentation and Filtering

In contrast to the SLIC super-pixel segmentation approach to discover the objects used by Ghorbani et al. [1]. this study employs the Segment Anything Model (SAM2) 2.3.1 for precise object segmentation. The tool also provide configurations options that enable the detection of fine-grained textures in addition to object boundaries. An example of the resulting segmentation output is shown in Figure 7.

Regarding the CLIP similarity-based filtering of the Target classes, Figure 8 shows the code log.

(a) Basketball

(b) Basketball Player

(c) Mountain Bike

(d) Helmet

(e) Moving Van

(f) Van

(g) ACE Super pixel segmentation

Figure 7. Visualization of 3 classes images and segmentation pairs in FedCAPE (left), ACE Segmentation (right)

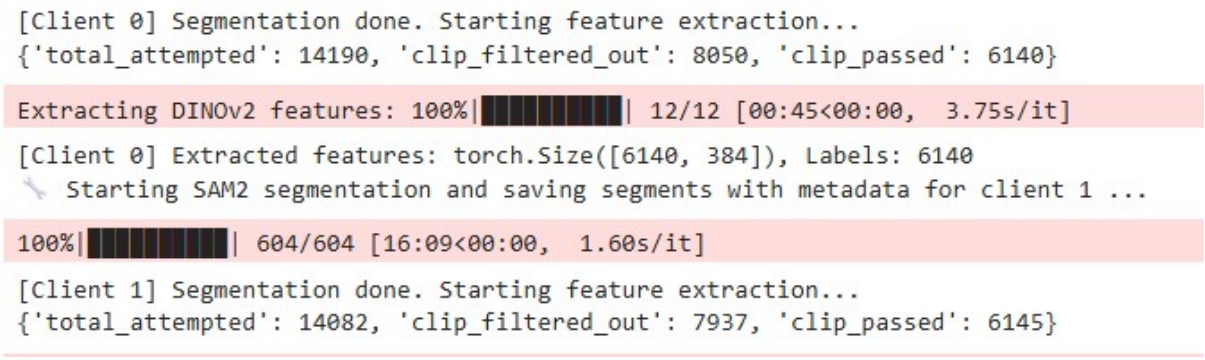


Figure 8. OpenAI CLIP model for **object segmentation** filtering

4.2 Concept Discovery and Clustering

4.2.1 Concept Clustering Procedure

The FedCAPE pipeline adopts a federated K-means clustering mechanism for concept centroids to facilitate global interpretability while preserving data privacy. Each client independently performs clustering on their locally discovered concept embeddings, resulting in a set of local centroids that capture prominent concepts within their data partition.

These local centroids are periodically transmitted to a central server, which aggregates them—typically via averaging in the embedding space—to construct a unified global concept dictionary. This global aggregation ensures that concepts discovered on different clients are harmonized, enabling cross-client explainability and downstream analysis. It is worth mentioning that the severe aggregation confirms the convergence if the maximum difference between the client class centroids is an empirical value of .5 units in Dinov2 embedded space.

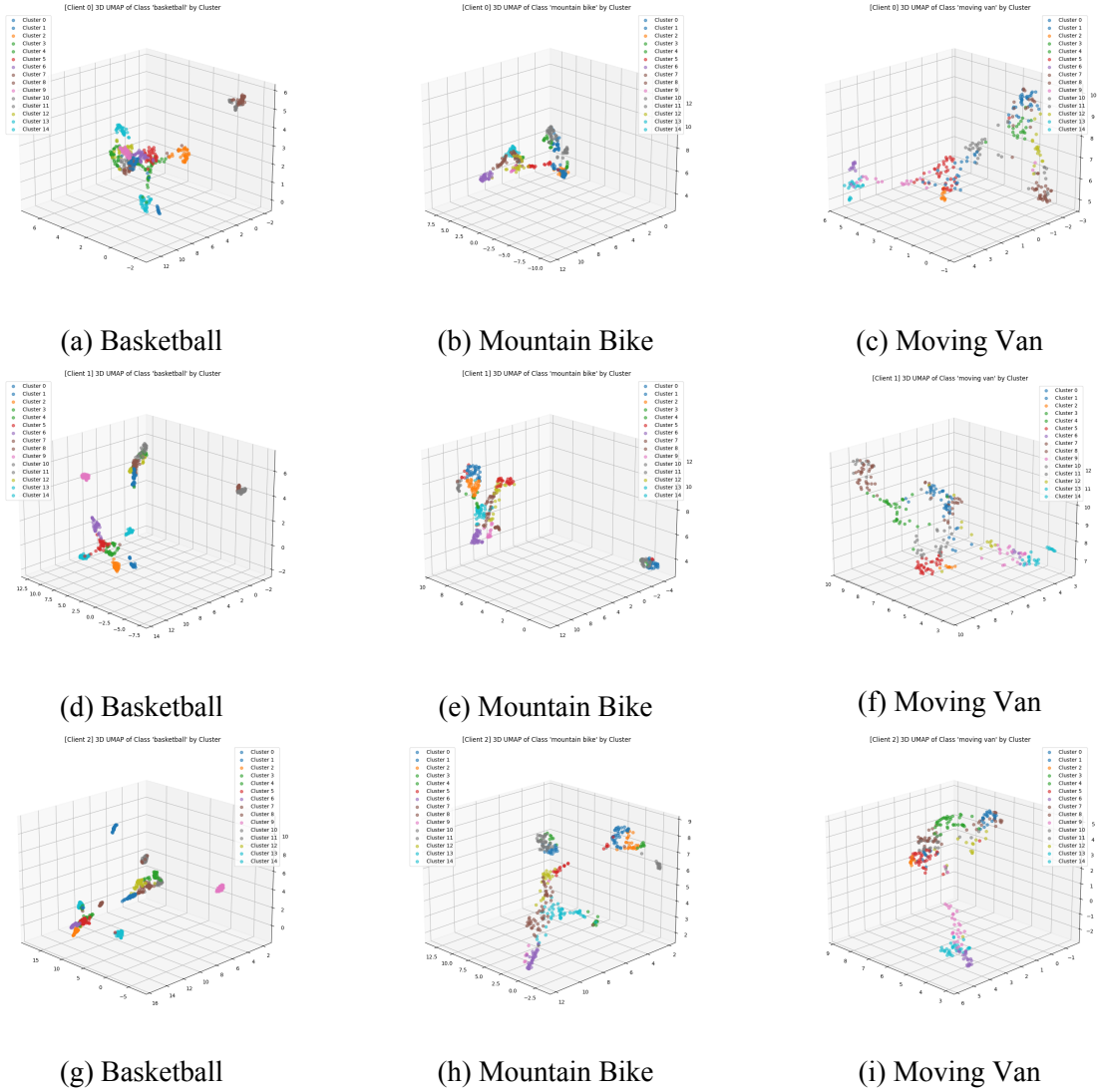
Figure 9 presents UMAP projections of the clustered concept embeddings for all clients and classes. The 3×3 grid structure shows, for each client (rows), the distribution of clusters for each semantic class (columns: *Basketball*, *Mountain Bike*, *Moving Van*). Each point represents a segment or concept embedding, color-coded by cluster assignment.

This visualization enables a qualitative assessment of three key aspects:

- **Cluster separation:** Distinct, well-separated color-coded clusters suggest coherent concepts, whereas overlapping clusters may indicate ambiguity or redundancy.
- **Inter-client variation:** Comparing rows reveals the degree of consistency or diversity in concepts discovered across different clients.

- **Class-wise structure:** Comparing columns highlights variations in concept structure between classes.

It is worth noting that: - The number of clusters is manually set, which may affect cluster purity and separation. Nonetheless, the emergence of distinct, color-coded clusters across clients and classes demonstrates the effectiveness of the pipeline in discovering meaningful and interpretable concepts in a distributed setting. Additionally, The Data Distribution might be in a way that is hard to cluster using the K-Means. In this case, there should be a fallback mechanism to test one more suitable for non-spherical data distributions.



(a) Basketball (b) Mountain Bike (c) Moving Van

(d) Basketball (e) Mountain Bike (f) Moving Van

(g) Basketball (h) Mountain Bike (i) Moving Van

Figure 9. UMAP clustering of concepts per class and client. Each row corresponds to a different client (top: Client 0, middle: Client 1, bottom: Client 2); each column shows the clustering for a specific class ("Basketball", "Mountain Bike", "Moving Van").

4.2.2 Concept Coherency and Interpretability

The concept/cluster coherency depends totally on the data preprocessing phase; the CLIP class purifies the object segments by filtering out what is irrelevant to the class semantically and ensuring the segment label's alignment with the target class. The following are samples from different clients with different clusters across classes to illustrate the cluster content.

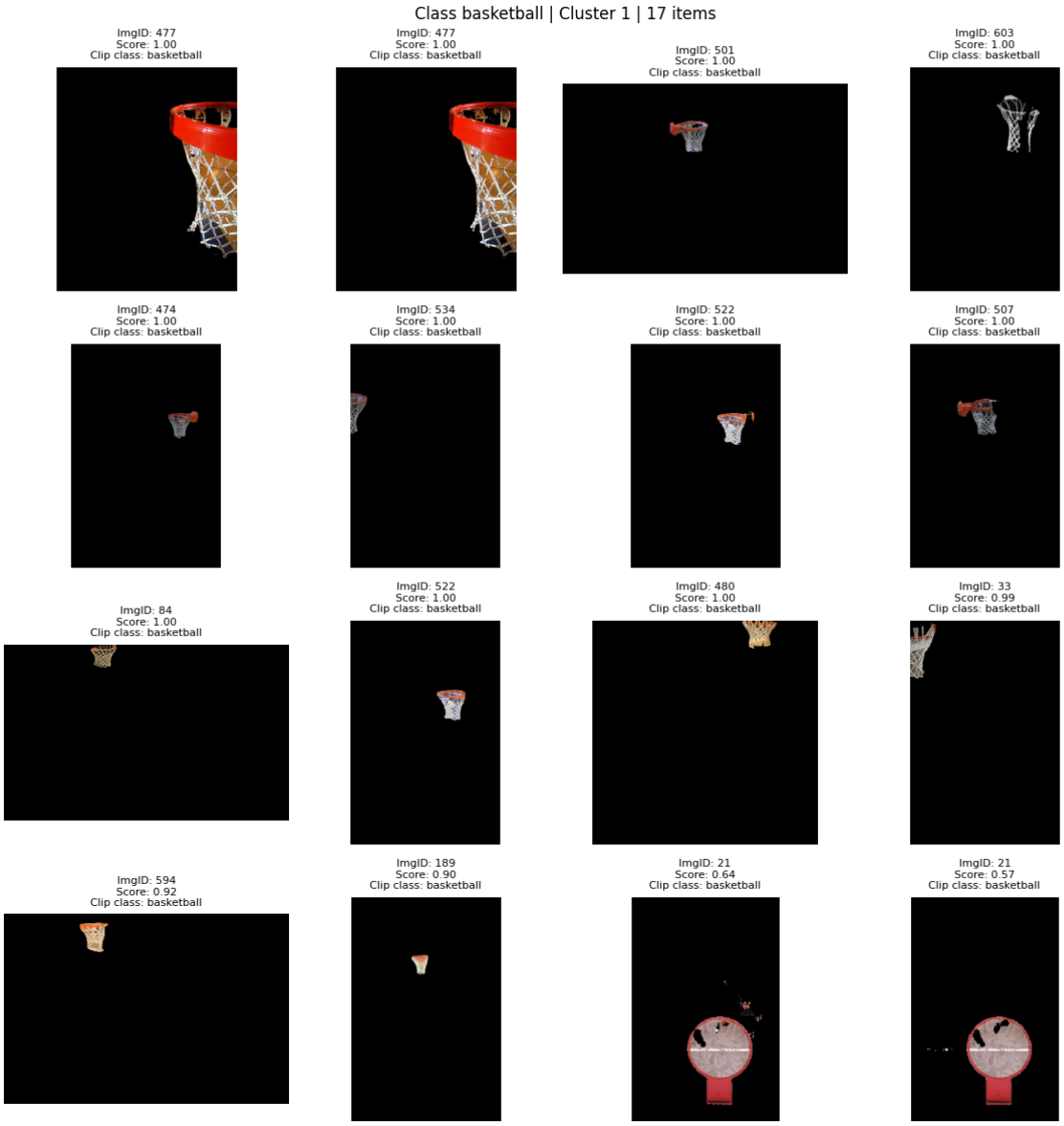


Figure 10. Representative samples from Client 0 for the *Basketball* class, with salient segments highlighted from Cluster 1

-

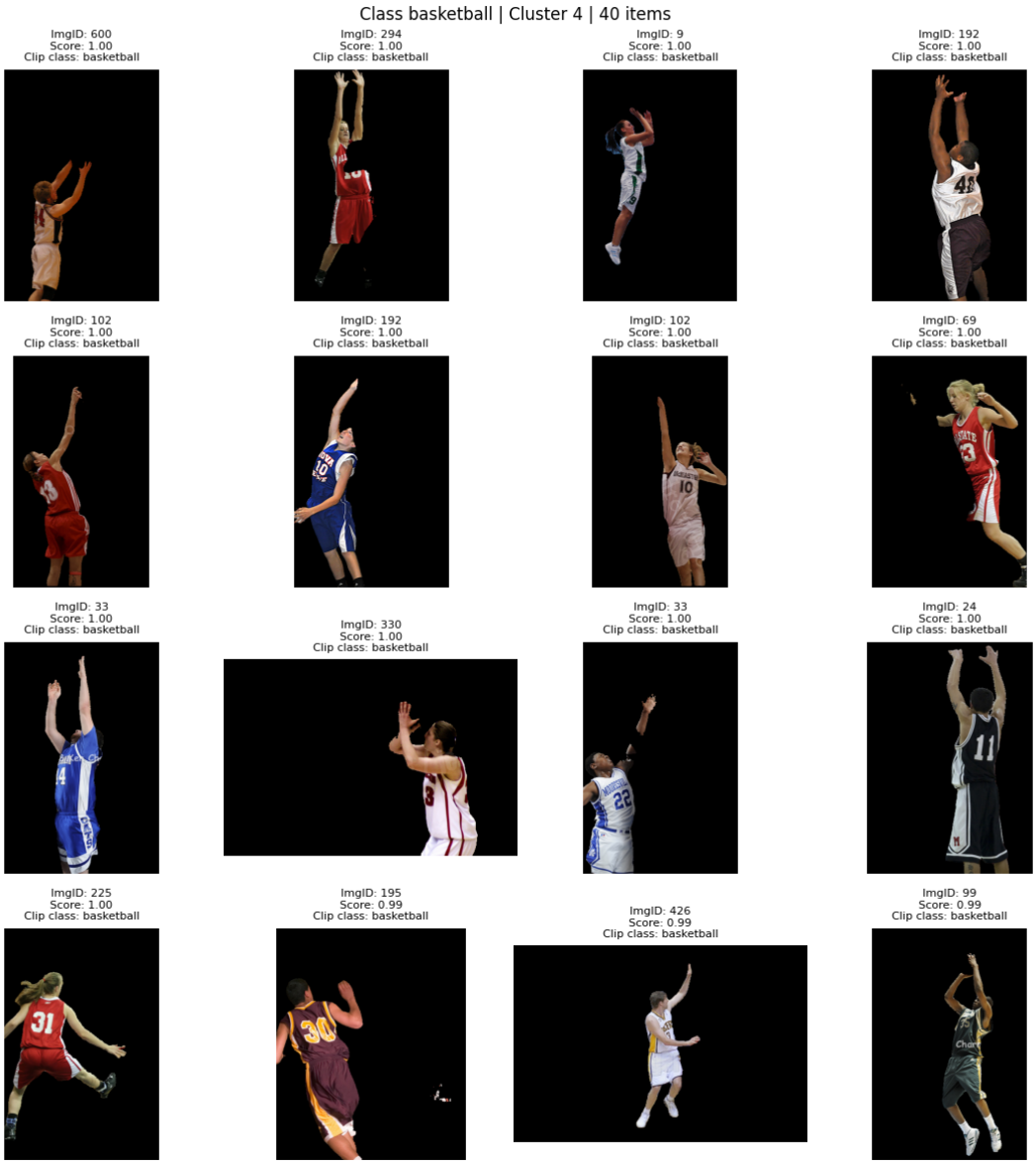


Figure 11. Representative samples from Client 0 for the *Basketball* class, with salient segments highlighted from Cluster 4

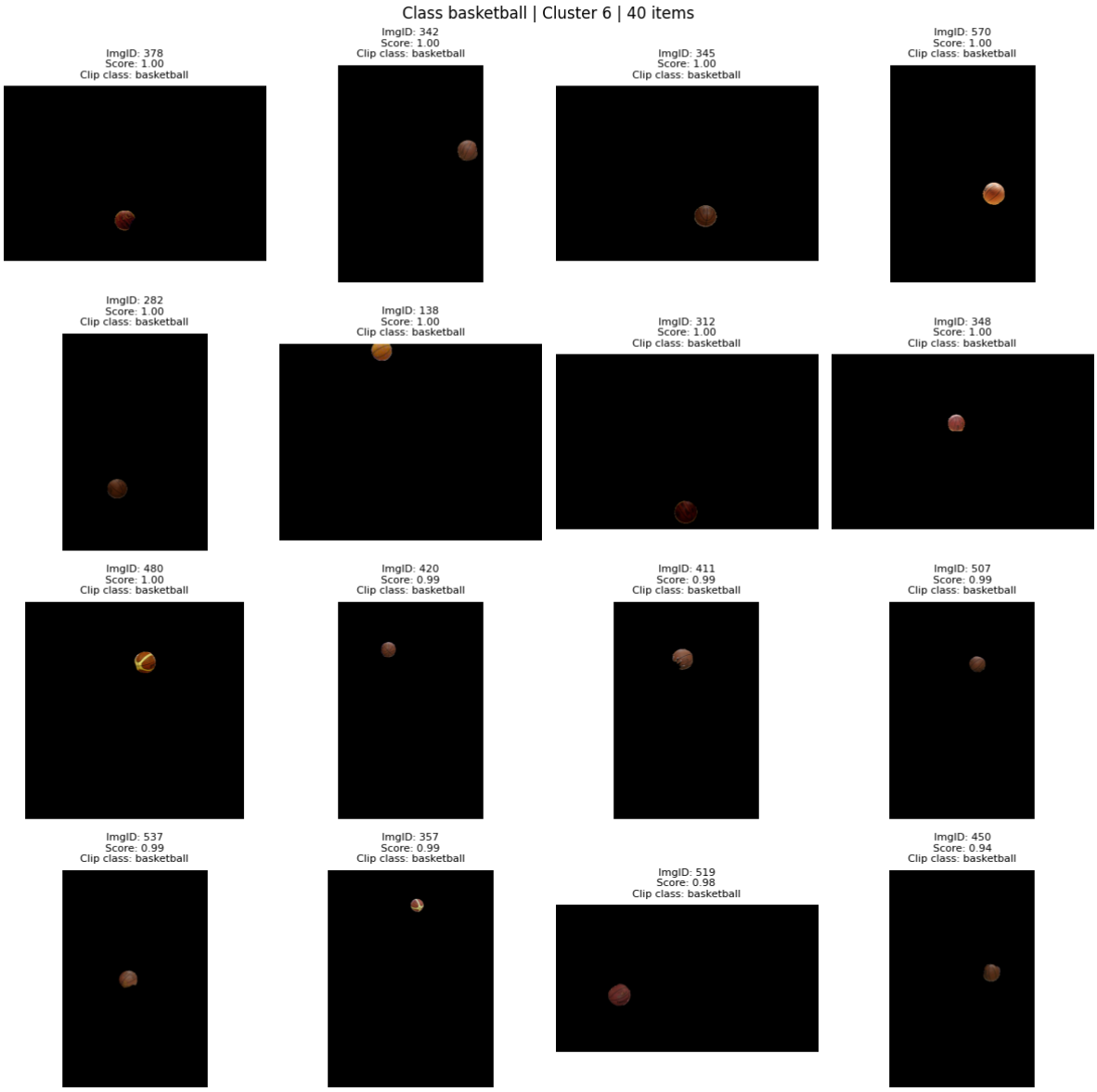


Figure 12. Representative samples from Client 0 for the *Basketball* class, with salient segments highlighted from Cluster 6

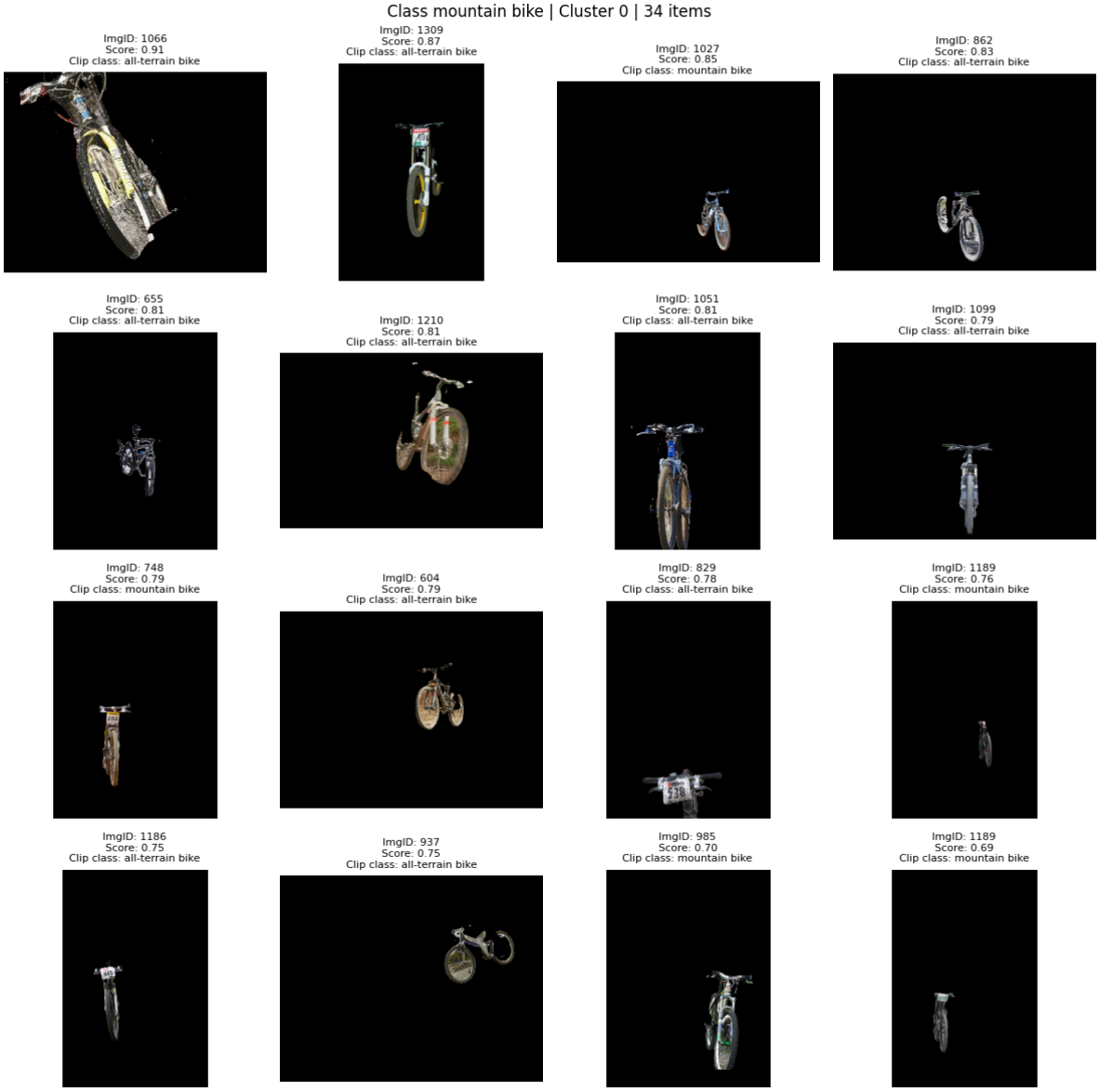


Figure 13. Representative samples from Client 1 for the *mountain bike* class, with salient segments highlighted from Cluster 0

Figure 14. Representative samples from Client 1 for the *mountain bike* class, with salient segments highlighted from Cluster 1

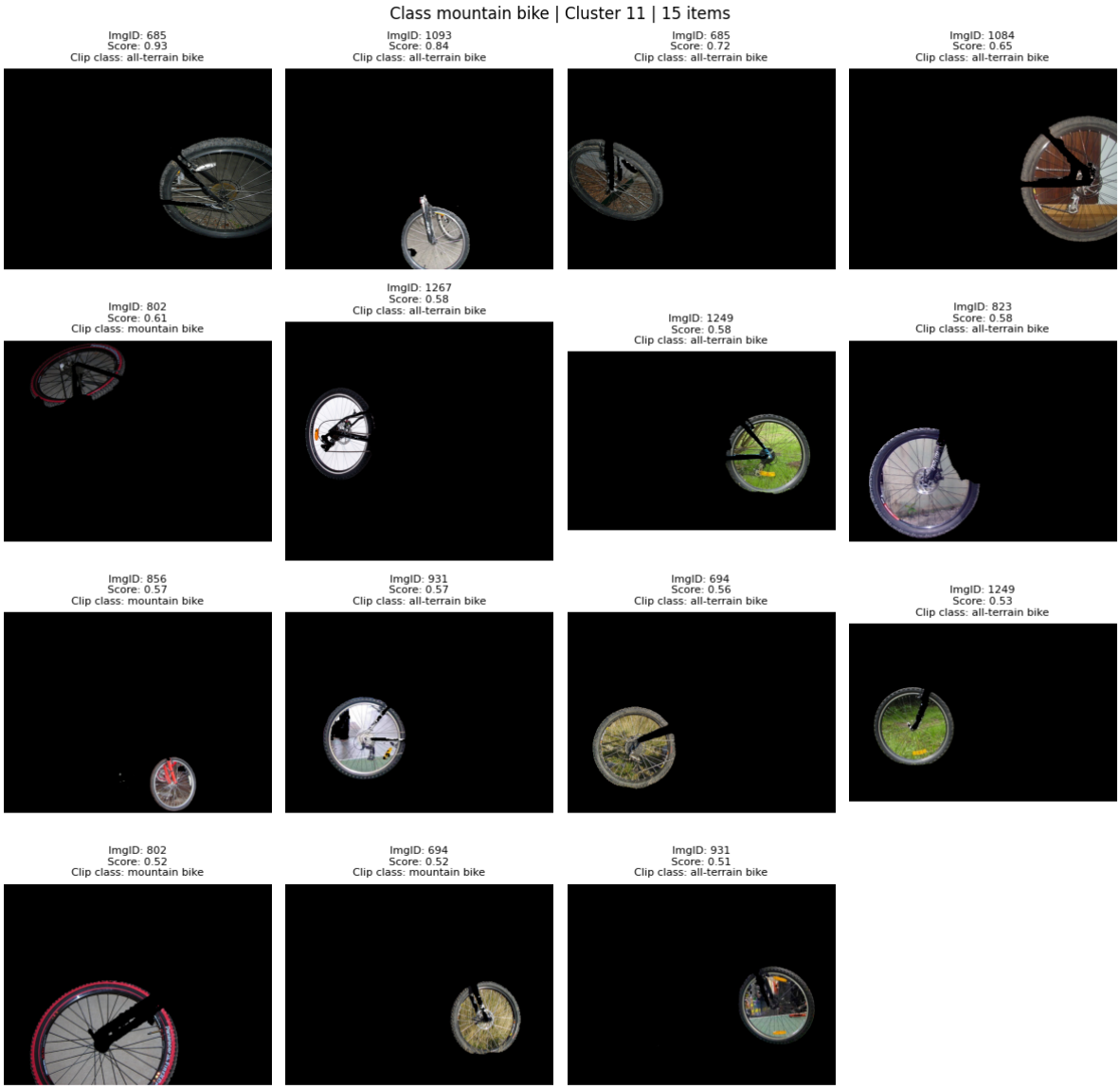


Figure 15. Representative samples from Client 1 for the *mountain bike* class, with salient segments highlighted from Cluster 11

Figure 16. Representative samples from Client 2 for the *moving van* class, with salient segments highlighted from Cluster 0

Figure 17. Representative samples from Client 2 for the *moving van* class, with salient segments highlighted from Cluster 2

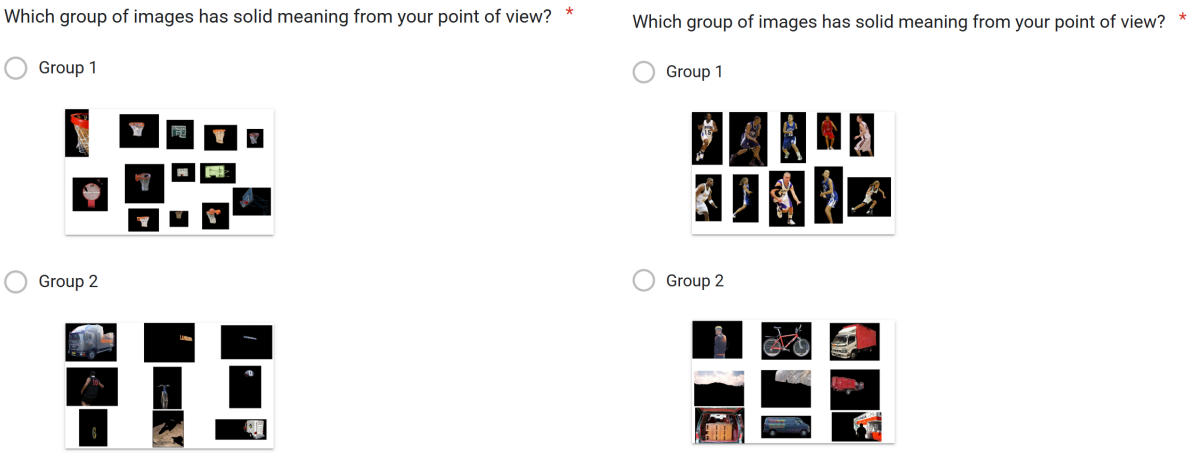
Figure 18. Representative samples from Client 2 for the *moving van* class, with salient segments highlighted from Cluster5

4.2.3 Qualitative Evaluation:Human Evaluation of Concept Cluster Coherence

To assess whether the concepts discovered by FedCAPE are meaningful to humans, we conducted a small user study [42] using a forced–choice interface to consider the ”basketball” class clusters. For each task (Figure 19), a participant saw two images:

(i) **Group 1** — Segments drawn from a single FedCAPE cluster (candidate concept).

(ii) **Group 2** — Segments randomly mixed from different clusters and classes.



Group 1: Salient Basketball concepts *Group 2: Random concepts*

Figure 19. Example **basketball** salient clusters (concepts) used in the human evaluation

The task was:

> “Which group of images has a solid, consistent meaning?”

We collected relevant individual background information (highest degree, years of professional experience).

Vote-Based Agreement

Across all trials, participants consistently preferred Group 1 clusters, with agreement rates ranging from **80% to 100%** depending on the cluster. This strong preference indicates that the majority found FedCAPE-derived clusters to be more visually coherent than randomly mixed examples.

Participant-Generated Labels

To further investigate semantic alignment, participants were asked to describe what they perceived in Their selection. The most common responses for selected *Basketball* class clusters are summarized below:

Table 2. Most common participant-generated labels for *Basketball* clusters and their semantic alignment with intended cluster meanings.

Cluster (Group 1)	**Common Participant Labels**	**Semantic Alignment with Intended Cluster**
Basketball Hoop	"Basketball hoops", "Basketball"	Matches the intended hoop/rim focus — strong consensus on object identity.
Players	"Players", "Basketball player"	Matches intended player-focused cluster — participants noticed human figures in uniforms.
Basketball Floor	"Basketball floor", "Court", "Matchground"	Aligns perfectly with floor/court cluster — users recognised spatial context.
Player in Shooting Action	"Basketball player in shooting action"	Matches intended action-based cluster — motion and ball trajectory noticed.

Interpretation

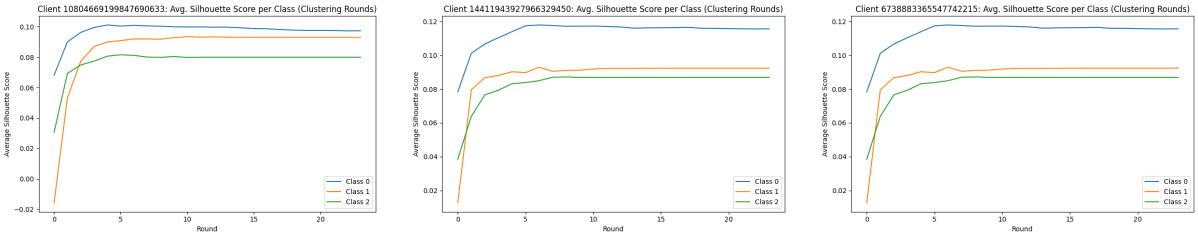
The close match between **participant-generated labels** and **the semantic intention of each cluster** demonstrates that FedCAPE produces visually and semantically coherent concepts. In contrast, Group 2 responses were inconsistent and generic (e.g., "sports scenes"), underscoring the lower semantic purity of mixed clusters.

This finding confirms that **qualitative interpretability** in FedCAPE is not merely a by-product of the segmentation pipeline, but results from **meaningful concept discovery** that aligns with human perception. Combined with the **quantitative TCAV analysis** (Section 4.4.1), this provides a multi-faceted validation of the framework's interpretability.

4.3 Federated Aggregation of Concepts

4.3.1 Federated Kmeans convergance

As discussed in the 3.1 section, the Federated K-means converged and stopped when reaching an empirical point (0.5 Dino space unit). The current implementation doesn't cover the silhouette score tracking for every cluster, but the following graph shows the average score tracking per class across the clients.



(a) FedCAPE Client 0 (b) FedCAPE Client 1 (c) FedCAPE Client 2

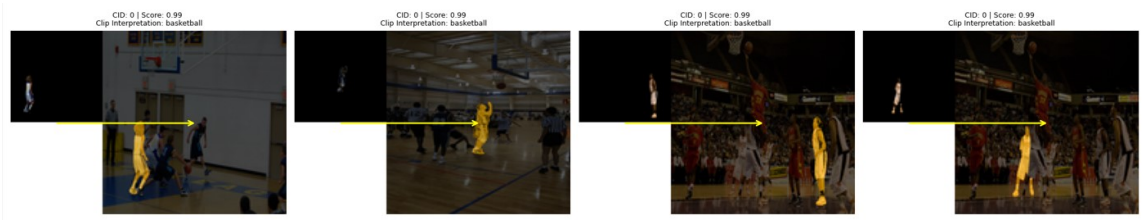
Figure 20. Salient Clusters in FedCAPE and Centralized

4.3.2 Cross-Client Consistency

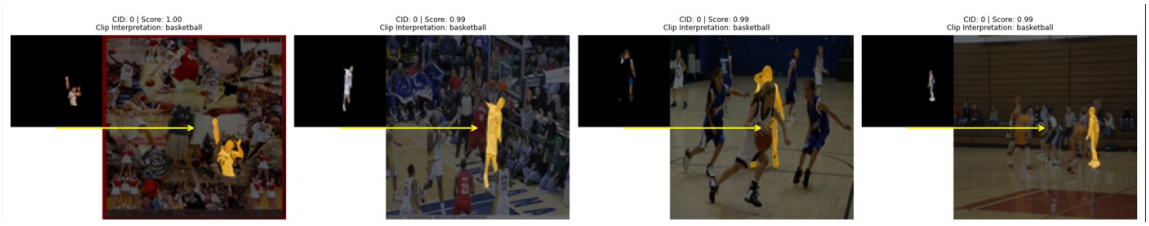
The discovered salient concept clusters—specifically clusters (0–1, 4–7) for the basketball class—show strong consistency across clients. As illustrated in Figure 21, Each row shows representative samples from one client, with the identified cluster segments highlighted. Across all clients, the segments are visually and semantically coherent, typically corresponding to the same concept (e.g., basketball player in action). This demonstrates that the federated pipeline can achieve robust cross-client concept alignment, despite the variations in the local data distributions.

4.4 Concept Importance Assessment

Following the methodology of Ghorbani et al. [10], the relevance of the discovered concepts is quantitatively evaluated using **TCAV (Testing with Concept Activation Vectors)**. For each class, TCAV scores are calculated for every cluster, measuring how much each concept influences the model's classification decisions. This provides a direct and interpretable metric for assessing the contribution of individual concepts within the proposed pipeline.

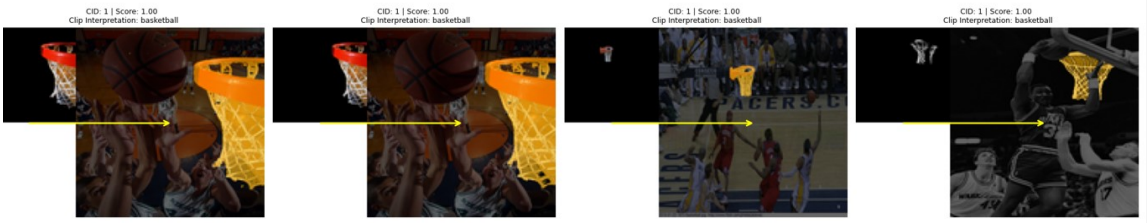


(a) Client 0: Cluster 0



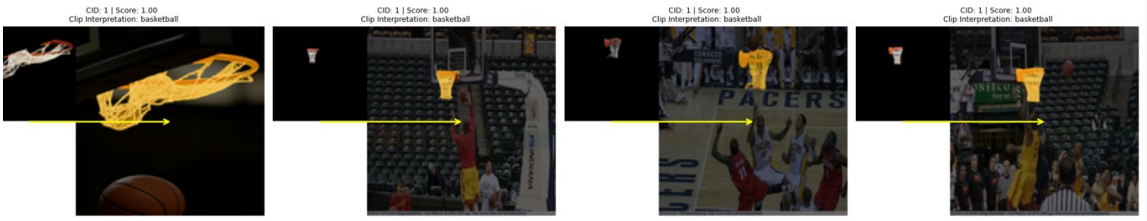
(b) Client 1: Cluster 0

(c) Client 2: Cluster 0



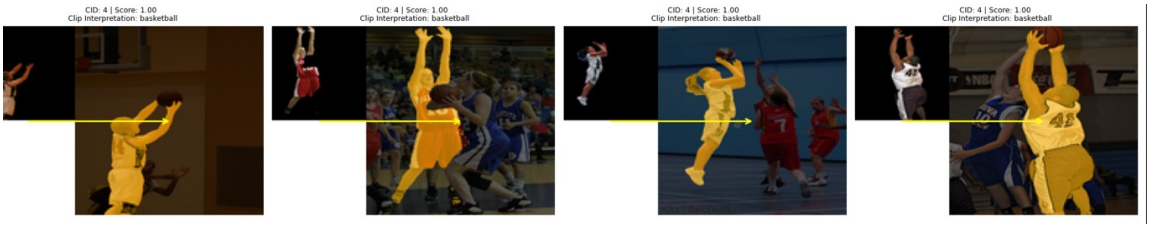
(d) Client 0: Cluster 1

(e) Client 1: Cluster 1

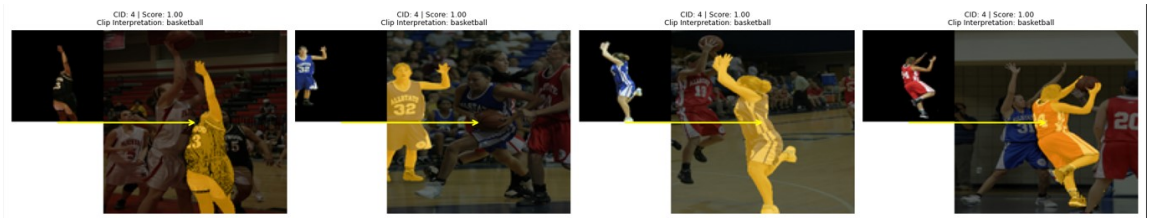


(f) Client 2: Cluster 1

Figure 21. Basketball clusters across clients (part 1)

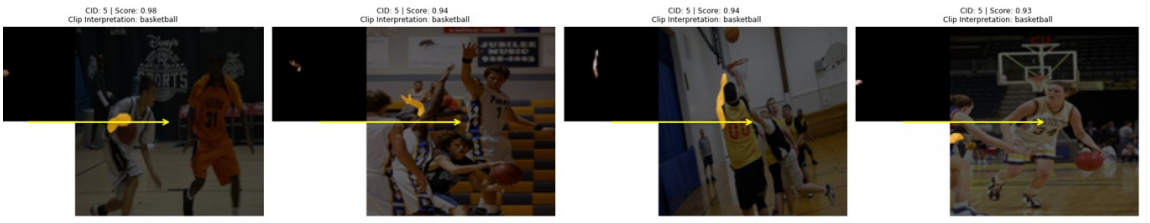


(g) Client 0: Cluster 4



(h) Client 1: Cluster 4

(i) Client 2: Cluster 4



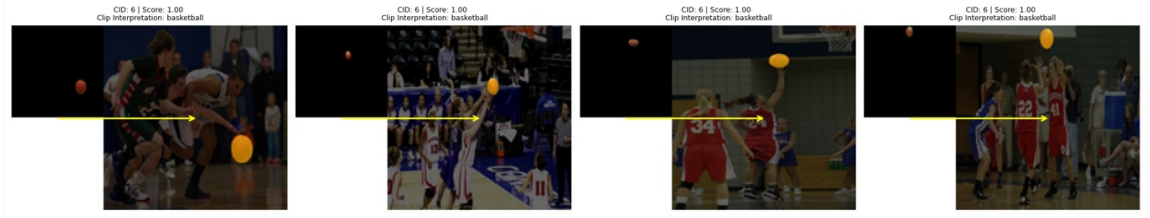
(j) Client 0: Cluster 5

(k) Client 1: Cluster 5

(l) Client 2: Cluster 5

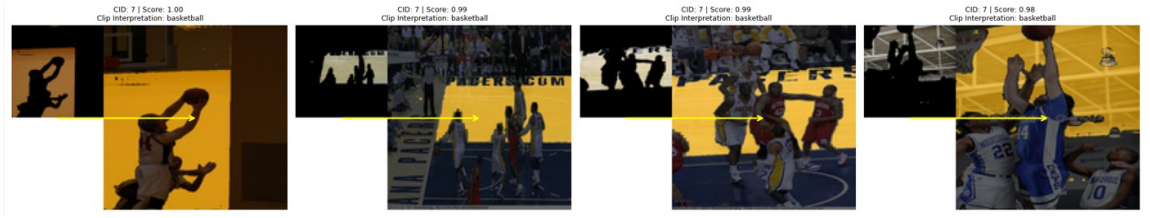
Figure 21. Basketball clusters across clients (part 3)

(m) Client 0: Cluster 6



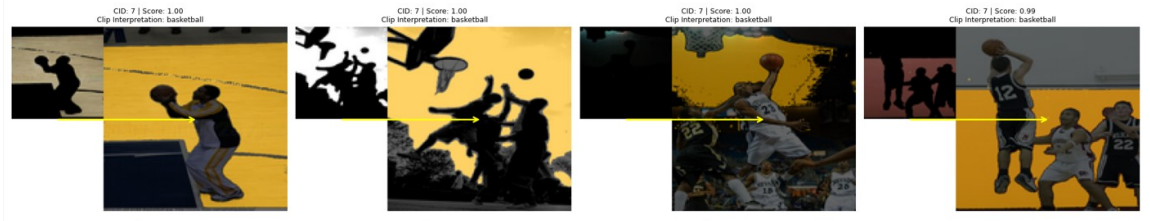
(n) Client 1: Cluster 6

(o) Client 2: Cluster 6



(p) Client 0: Cluster 7

(q) Client 1: Cluster 7



(r) Client 2: Cluster 7

Figure 21. Basketball clusters across clients (part 4)

4.4.1 FedCAPE vs. Centralized TCAV Comparison

As outlined in Algorithm 3.1, the basketball class is used as a case study to compare FedCAPE and centralized setups based on TCAV scores. Table 3 reports the basketball class's top three salient clusters/concepts, evaluated on the basketball test dataset under identical experimental conditions. Results show that FedCAPE can surpass the centralized setup in certain configurations, likely due to differences in dataset distribution. Figure 22 presents representative samples from each salient cluster in both setups.

FedCAPE (Client 0, basketball)			**Centralized**	
Client-0 Cluster (short semantic description)	**Global**	**Local**	**Cluster**	**TCAV**
basketball-C0 (Basketball Player I Standing)	0.28	0.24	central-C14	0.21
basketball-C1 (Basketball Net)	**0.93**	**1.00**	central-C12	**0.97**
basketball-C2 (Outlair samples)	0.00	0.03	central-C11	0.0
basketball-C3 (Basketball clothes I)	0.52	0.59	-	-
basketball-C4 (Basketball Player II shooting)	0.07	0.03	central-C8	0.24
basketball-C5 (Basketball Player far Arm)	0.38	0.38	central-C1	0.10
basketball-C6 (Basketball far Ball II)	0.34	0.69	central-C10	0.72
basketball-C7 (Basket ball Surrounding)	0.83	0.62	central-C6	0.45
basketball-C8 (Basketball Player III)	0.10	0.10	central-C7	0.21
basketball-C9 (Basketball near Ball I)	**0.90**	**0.90**	central-C4	**0.83**
basketball-C10 (Basketball Player IV)	0.24	0.34	central-C13	0.31
basketball-C11 (Basketball Playground)	**0.97**	0.90	central-C5,C3	**0.97,0.83**
basketball-C12 (Basketball Player V)	0.24	0.24	central-C2	0.14
basketball-C13 (Basketball Player near Arm)	0.07	0.07	central-C0	0.21
basketball-C14 (Basketball clothes II)	0.45	0.62	central-C9	0.48

Table 3. TCAV scores for the *basketball* class, Left: FedCAPE (Global & Local), Right: Centralized

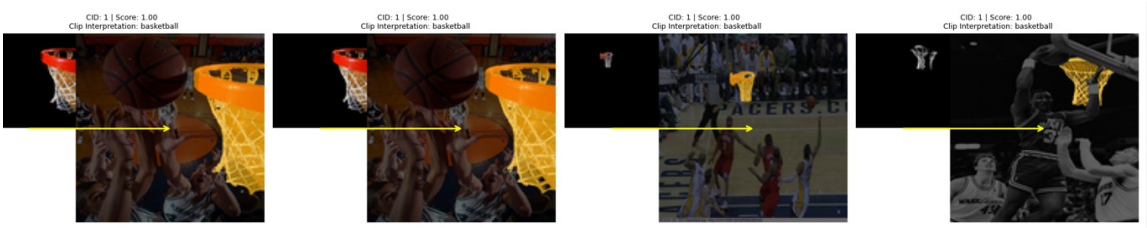
having the clustering is done using the same applicable hyper parameter like the number of cluster From the table we can notice that a symptom of over clustering happened in the centralized usecase as both the 2 clusters (central-C3 and central-C5) presenting the same concept which is the "Basketball Playground", moreover the federated Clustering recognize the basketball-C3 which Presents fine elements like "player shoes" and other elements is having less of a presentations

by its ouwn and merged with other cluster.it is also worth mentioning that the mapping between the FedCAPE clusters and the centralized was done manually by visual inspection.

4.4.2 Interpretation of Salient Concepts

Salient clusters are those with high TCAV scores for their own class indicating strong class-specific interpretability. For example, in Figure 23, clusters C1, C9, and C11 show high importance for the ”basketball” class. Similar patterns are observed for the other classes.

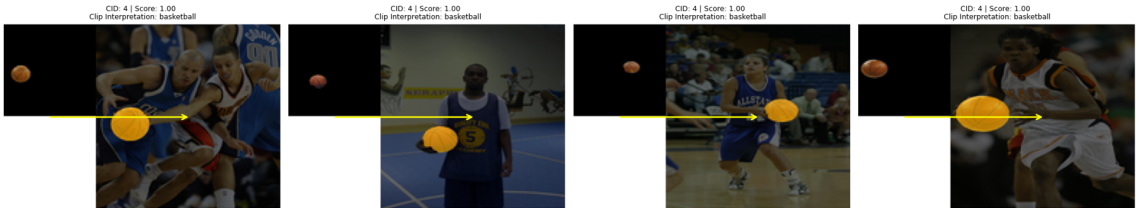
Additionally, qualitative inspection of the most salient clusters (see visualizations in Section 21) shows these clusters correspond to semantically meaningful image regions, supporting the quantitative analysis.



(a) FedCAPE Client 0: Cluster 1

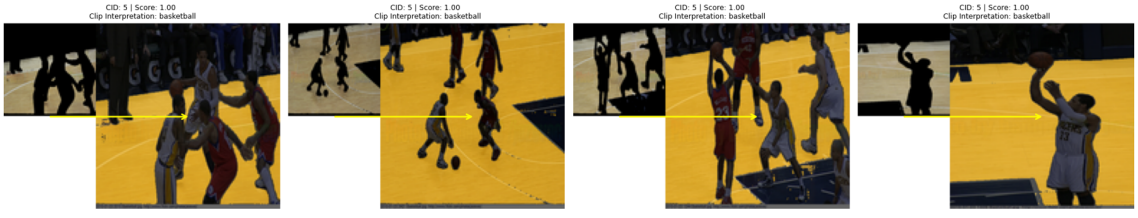
(b) centralized : Cluster 12

(c) FedCAPE Client 0: Cluster 9



(d) centralized : Cluster 4

(e) FedCAPE Client 0: Cluster 11



(f) centralized : Cluster 5

Figure 22. Salient Clusters in FedCAPE and Centralized

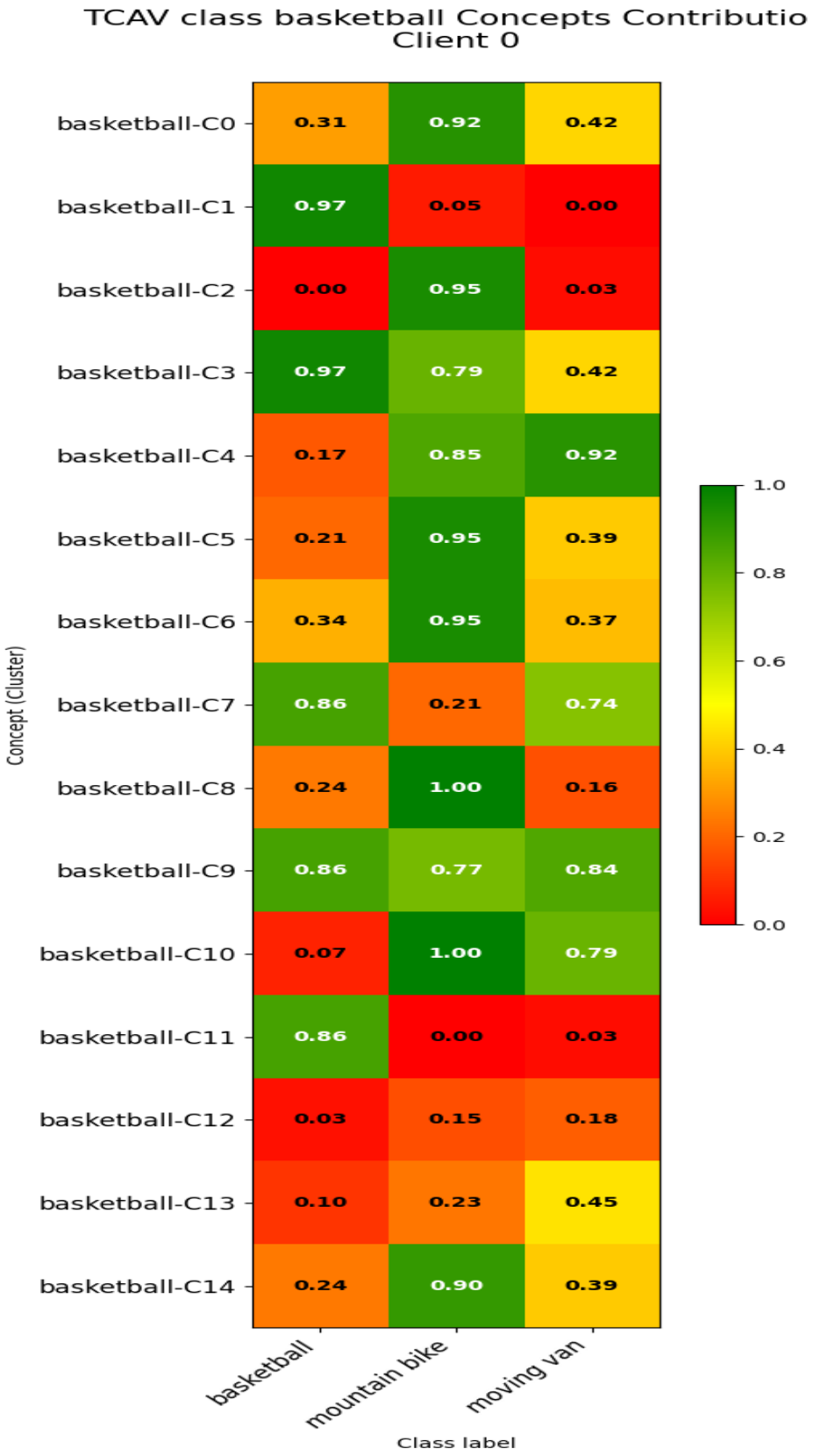


Figure 23. TCAV scores heatmap for **basketball** concepts (Client 0)

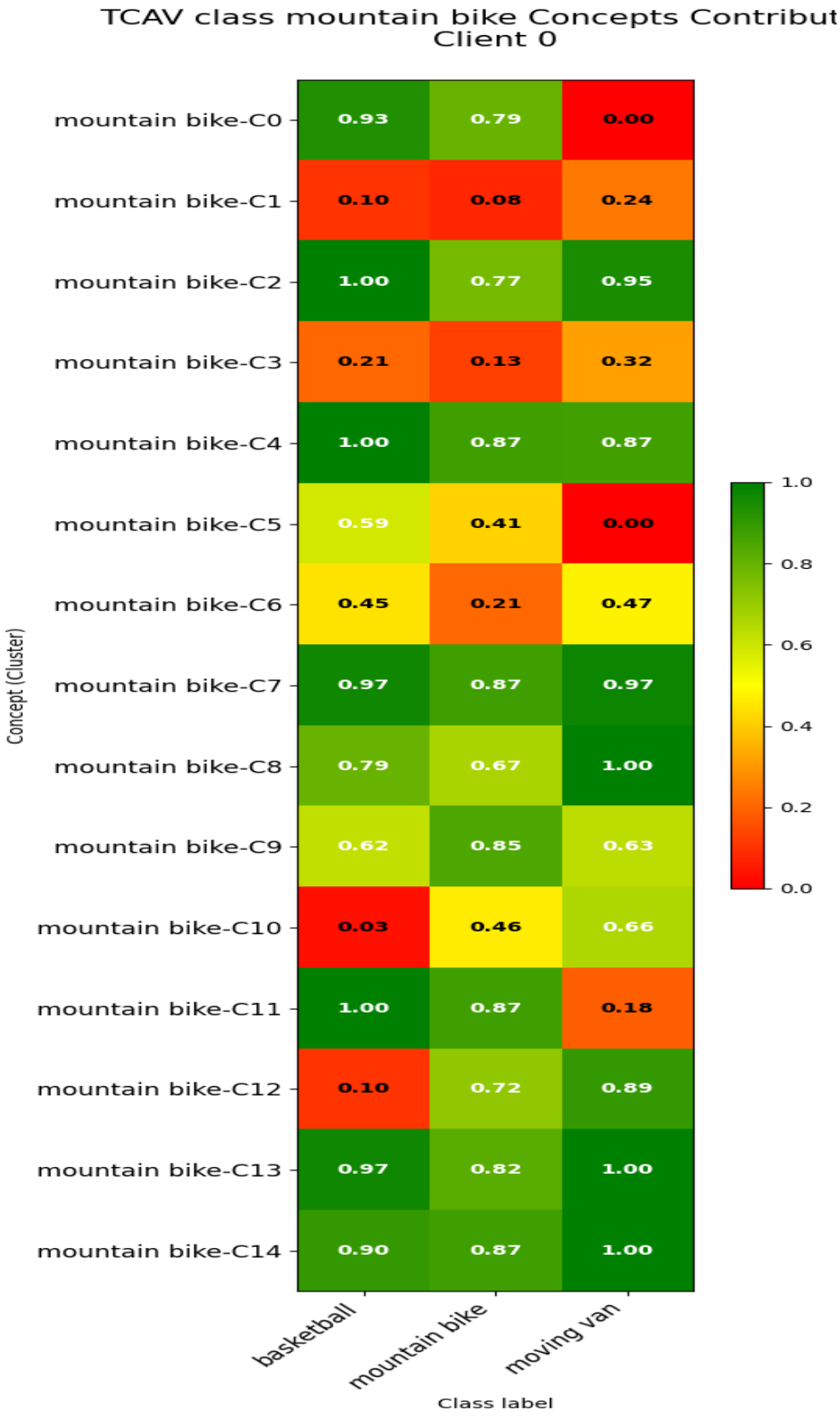


Figure 24. TCAV scores heatmap for **mountain bike** concepts (Client 0)

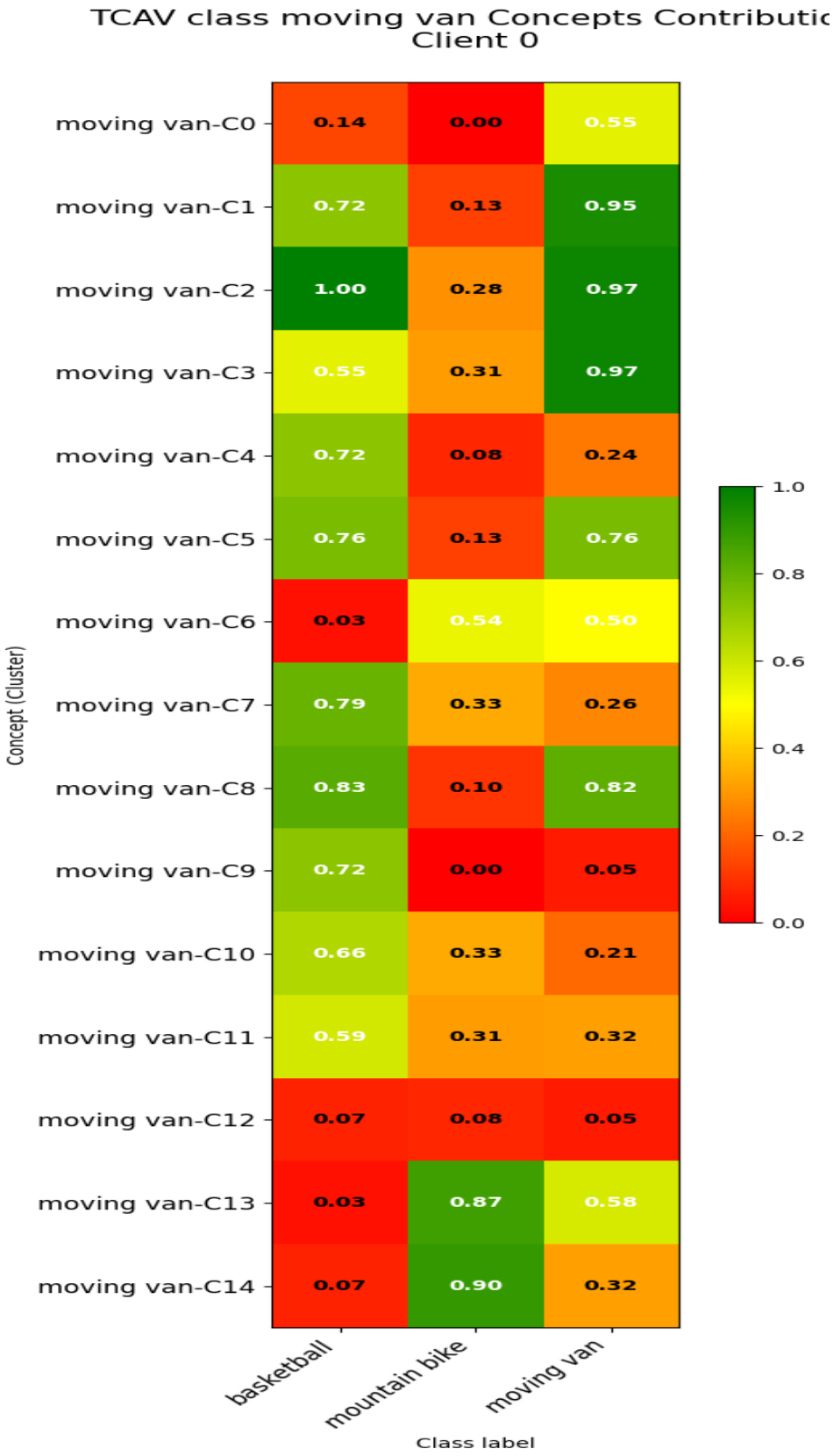


Figure 25. TCAV scores heatmap for **moving van** concepts (Client 0)

5. Limitations and Future Work

While FedCAPE provides a novel framework for interpretable federated learning through concept alignment, it is not without limitations. This chapter outlines known constraints, observed challenges during implementation, and proposes directions for future research.

5.1 Limitations

5.1.1 Concept Drift and Stability

A significant challenge in federated concept learning arises from **concept drift**—the phenomenon where local client data distributions evolve over time, either due to the arrival of new data or the appearance of novel, fine-grained details within existing classes. As a consequence, the clusters representing concepts may undergo pollution or fragmentation; previously salient and coherent groups can become diluted, with outlier or unrelated features joining the cluster. This often necessitates discovering a larger number of clusters to maintain concept precision, consequently reducing interpretability.

These effects manifest as:

- **Inconsistent cluster formation or disappearance of prior concepts:** As distributions shift, previously discovered concepts may dissolve, while new ones materialize unpredictably, leading to instability across federated rounds.
- **Polluted or fragmented clusters:** The need to cover emerging patterns can generate clusters with mixed or unclear semantics, challenging both local and global alignment.
- **Reduced TCAV reliability:** If concepts evolve or hybridize, the meaningfulness and stability of TCAV-based attributions decline, undermining the trustworthiness of explanations.

5.1.2 Foundation Model Accuracy: DINO, CLIP, and SAM

The robustness of FedCAPE's pipeline heavily relies on the accuracy and discriminative power of foundational models leveraged for feature extraction (DINO), semantic filtering (CLIP), and image segmentation (SAM/SAM2). Limitations in these models directly propagate through the pipeline:

- **DINO Feature Limitations:** If DINO features do not sufficiently separate classes or fail to capture meaningful texture/shape cues, resulting concepts may be blurred, noisy, or lack intra-class cohesion.

- **CLIP Semantic Filtering:** Imperfect CLIP filtering can result in retaining irrelevant segmentations or discarding true positives, especially when class tokens are ambiguous or CLIP's training vocabulary does not cover certain image semantics.
- **SAM Segmentation Error:** Over-segmentation, under-segmentation, or incorrect object boundaries due to SAM's limitations can lead to clusters representing partial, incomplete, or background regions.
- **Propagation of Error:** Errors or inadequacies at any foundational stage act as a bottleneck, reducing the stability, interpretability, and utility of the aligned concepts.

These challenges highlight the need for continual evaluation and possibly modular upgrading of the feature, segmentation, and filtering models as improved versions or domain-adapted variants become available.

5.1.3 Evaluation Benchmarks

There is currently no established benchmark for evaluating federated interpretability frameworks in automatic concept extraction. This limits:

- Direct comparisons to other methods.
- Quantitative validation of "explanation quality" beyond TCAV.

It is worth noting that the framework proposed in LR-XFL [26] provides an interpretation methodology grounded in rule-based reasoning, evaluated through well-defined metrics such as model accuracy, rule accuracy, and rule fidelity. This framework could be leveraged to assess the extent to which the concepts discovered by FedCAPE contribute to performance in an established benchmark setting.

5.1.4 Privacy Leakage via Gradients

While FedCAPE avoids sharing raw features, Dino-based signatures can carry indirect information about local data.

5.2 Future Work

5.2.1 LLM-Assisted Concept Evaluation

Large language models (LLMs) can be used to describe and semantically validate discovered concepts:

- Generate natural language descriptions of each concept.

- Score semantic coherence across aligned global concepts.

This could improve transparency for end-users and enable natural language querying of concept dictionaries.

5.2.2 Dynamic Concept Dictionaries

Enable time-evolving global dictionaries that support:

- Versioning of concepts
- Online learning from new clients or data streams
- Deprecation of stale or redundant concepts

This would enhance FedCAPE's ability to function in continual learning scenarios.

5.2.3 End-to-End Classifier on Concept Space

Train a final federated classifier purely on aligned concept activations:

- Replaces raw features entirely with interpretable inputs
- Allows full pipeline explanation of predictions
- Potentially improves robustness by abstracting away from raw feature noise

5.2.4 Integration with Privacy-Preserving Techniques

FedCAPE can benefit from formal privacy guarantees:

- Differential privacy noise applied to concept signatures
- Secure multi-party computation for distributed alignment

These techniques can reduce the risk of leakage through signatures while maintaining interpretability alignment.

5.2.5 Advanced Concept Representations and Hyperplanes

Developing more complex, robust concept representations is a promising avenue for future work. Instead of limiting concepts to linear boundaries in latent space (as used with Concept Activation Vectors, CAVs), future research can explore:

- Combining activations from multiple bottleneck or latent layers to form **hyperplane**-based or non-linear concept definitions.

- Leveraging multi-layer or non-linear models to better capture nuanced, stable, and composite semantic factors.

Such advances could help stabilize concept discovery under drift and improve the semantic fidelity of federated explanations in complex, evolving environments.

5.2.6 Handling Corner Cases in Federated Concept Discovery

While FedCAPE demonstrates robust concept discovery and alignment under typical federated scenarios, several corner cases remain open challenges for future work. These cases can significantly influence both the quality and trustworthiness of the global concept dictionary:

- **Severe Non-IID Distributions:** When certain clients contain data distributions drastically different from others (e.g., due to local domain drift), concept discovery may produce client-specific concepts that have little or no overlap with the majority. This raises the question: *should such concepts be published to the global dictionary, and if so, under what trust criteria?*
- **Biased or Overrepresented Clients:** If a concept is discovered predominantly or exclusively by a single client, it may still be valuable globally (e.g., rare but important phenomena), or it may be an artifact of local bias. Deciding whether to propagate such Concept Activation Vectors (CAVs) without corroboration from other clients is a non-trivial governance issue.
- **Unverified Concept Propagation:** Publishing CAVs from a single client without contributions from others could lead to:
  (i) *Inflated Concept Importance:* Misleading TCAV scores in the global space if the concept is irrelevant elsewhere.
  (ii) *Semantic Drift:* Gradual divergence between the concept's intended meaning and its global usage.
  (iii) *Security Risks:* Potential leakage of unique or identifiable local features.
- **Potential Treatments:** Future research should investigate:
  (i) Establishing **confidence thresholds** or **support requirements** before a concept is accepted into the global dictionary.

(ii) Applying **cross-client validation** (e.g., verifying that a CAV produces consistent activations on other clients' data before global publication).

(iii) Introducing **trust weighting** in alignment, where concepts from single-client origins are down-weighted until corroborated.

(iv) Using LLM-assisted semantic checks to detect potential local overfitting in concepts prior to alignment.

Addressing these corner cases is critical for ensuring that FedCAPE's global dictionary remains both semantically coherent and trustworthy, especially when deployed in real-world non-IID and heterogeneous federated environments.

6. Conclusion

This thesis introduced **FedCAPE**, a framework for *automatic concept extraction and alignment in federated learning environments*. The central contribution lies in enabling each client to autonomously discover semantically meaningful concepts from its local data and collaboratively align them into a global concept dictionary—without sharing raw data or sensitive features. This is achieved through an implemented **federated K-Means clustering** mechanism, which aggregates concept representations across clients in a privacy-preserving manner.

Unlike conventional approaches that require centralized data or model aggregation, FedCAPE's pipeline integrates self-supervised segmentation, feature extraction, and clustering with federated concept alignment. The result is a distributed yet coherent concept representation across heterogeneous clients, paving the way for interpretable federated models.

The main contributions are:

- **Federated self-supervised concept discovery** using SAME, CLIP, and DINOv2 for segmentation, feature extraction, and per-class clustering.
- **Privacy-preserving concept alignment** via the implemented **federated K-Means clustering** Over Dino embedding space.
- **Concept importance evaluation** via TCAV and complementary clustering metrics.

Experimental results, supported by both qualitative (user studies, concept visualizations, projection plots) and quantitative (TCAV scores, clustering compactness, separability) evaluations, confirm that FedCAPE can reliably extract and federate meaningful concepts even under strict privacy constraints.

Broader Impact

By providing an automated, privacy-preserving approach to concept extraction in federated learning, FedCAPE extends the applicability of interpretable AI to sensitive domains such as healthcare, edge computing, and collaborative analytics—where data centralization is infeasible. This strengthens trust, compliance, and transparency in distributed intelligent systems.

Final Remarks

FedCAPE establishes a baseline for federated concept-based interpretability, with the primary benefit being its ability to perform automatic concept discovery across distributed data silos

through **federated K-Means clustering**. Future work should explore scaling to larger federations, extending to multimodal data, and integrating with human-in-the-loop refinement to further improve explanation quality.

References

[1] Ghorbani A., Wexler J., Zou J. Y., and Kim B. Towards Automatic Concept-based Explanations. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Wallach H. M., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E. B., and Garnett R. 2019, pp. 9273–9282. https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html.

[2] Ribeiro M. T., Singh S., and Guestrin C. Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[3] Lundberg S. M. and Lee S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[4] Sundararajan M., Taly A., and Yan Q. Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning*. 2017, pp. 3319–3328.

[5] Bach S., Binder A., Montavon G., Klauschen F., Müller K.-R., and Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* 10.7 (2015), e0130140.

[6] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., and Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 618–626.

[7] Simonyan K., Vedaldi A., and Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ICLR Workshop*. 2014.

[8] Kapishnikov A., Alon U., Hassidim A., Lublin O., Reif E., Wang Y., Yarom M., Yogev O., Ofek E., Sorodoc I., et al. XRAI: Better Attributions Through Regions. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4948–4957.

[9] Kim B., Wattenberg M., Gilmer J., Cai C., Wexler J., Viegas F., and Sayres R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *International Conference on Machine Learning*. 2018, pp. 2668–2677.

[10] Ghorbani A., Wexler J., Zou J. Y., and Kim B. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[11] Kirillov A., Mintun E., Ravi N., Mao H., Rolland K., Gustafson L., Xiao T., Whitehead S., Clegg A., Boyko A., et al. Segment Anything. *arXiv preprint arXiv:2304.02643* (2023).

[12] Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., and Sutskever I. Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.

[13] Oquab M., Darcet T., Moutakanni T., Vo H., Szafraniec M., Khalidov V., Fernandez P., Haziza D., Massa F., Assran M., et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[14] Lipton Z. C. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16.3 (2018), pp. 31–57. DOI: 10.1145/3236386.3241340.

[15] Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., and Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 51.5 (2018), 93:1–93:42. DOI: 10.1145/3236009.

[16] Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R., et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[17] Sokol K. and Flach P. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*. 2020, pp. 56–67. DOI: 10.1145/3351095.3372870.

[18] Ribeiro M. T., Singh S., and Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[19] Lundberg S. M. and Lee S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 4768–4777.

[20] Sundararajan M., Taly A., and Yan Q. Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 3319–3328.

[21] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., and Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[22] Molnar C. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/. Leanpub, 2019.

[23] Wachter S., Mittelstadt B., and Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31.2 (2017), pp. 841–887.

[24] Kim B., Wattenberg M., Gilmer J., Cai C. J., Wexler J., Viegas F., and Sayres R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *International Conference on Machine Learning*. PMLR. 2018, pp. 2668–2677.

[25] Yeh C.-K., Kim B., Arik S. O., Li C.-L., Pfister T., and Ravikumar P. Learning Interpretable Concept-Based Models with Human Feedback. *International Conference on Machine Learning*. PMLR. 2022, pp. 25462–25485.

[26] Zhang Y. and Yu H. LR-XFL: Logical Reasoning-based Explainable Federated Learning. 2023. arXiv: 2308.12681 [cs.AI]. https://arxiv.org/abs/2308.12681.

[27] Barbiero P., Ciravegna G., Giunchiglia E., Marra G., Tiddi I., Lio P., and Diligenti M. Entropy-based logic explanations of neural networks. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization. 2022, pp. 2145–2151.

[28] Lowerre B. T. The HARPY speech recognition system. PhD thesis. Carnegie Mellon University, 1976.

[29] Quinlan J. R. Induction of decision trees. *Machine Learning* 1.1 (1986), pp. 81–106.

[30] Yang J. and Long G. Concept□Guided Interpretable Federated Learning. *AI 2023: Advances in Artificial Intelligence*. Lecture Notes in Computer Science. Springer, 2023.

[31] Koh P. W., Nguyen T., Tang Y. S., Mussmann S., Pierson E., Kim B., and Liang P. Concept bottleneck models. *International Conference on Machine Learning (ICML)*. PMLR. 2020, pp. 5338–5348.

[32] Research M. A. Segment Anything: Official Code. https://github.com/facebookresearch/segment-anything. Accessed: 2025-07-30. 2023.

[33] OpenAI. CLIP: Official Code. https://github.com/openai/CLIP. Accessed: 2025-07-30. 2021.

[34] Research M. A. DINOv2: Official Code. https://github.com/facebookresearch/dinov2. Accessed: 2025-07-30. 2023.

[35] Yang Q., Liu Y., Chen T., and Tong Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019), p. 12.

[36] Kairouz P., McMahan H. B., Avent B., et al. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning* 14.1–2 (2021), pp. 1–210.

[37] Beutel D. J., Topal T., Mathur A., Qiu X., Parcollet T., Gusmão P. P. de, and Lane N. D. Flower: A friendly federated learning framework. *arXiv preprint arXiv:2204.03042* (2022).

[38] AI M. ExecuTorch powers on-device machine learning in Meta's family of apps. https://engineering.fb.com/2025/07/28/android/executorch-on-device-ml-meta-family-of-apps/. Accessed: 2025-07-30. 2025.

[39] AB S. S. FEDn: Open Federated Learning Framework. https://scaleoutsystems.com/fedn. Accessed: 2025-07-30. 2021.

[40] Kim B., Wattenberg M., Gilmer J., Cai C. J., Wexler J., Viégas F. B., and Sayres R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Dy J. G. and Krause A. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2673–2682. http://proceedings.mlr.press/v80/kim18d.html.

[41] FEDCAPE implementation. https://github.com/M-Shash/Thesis. Accessed: 2025-07-30. 2025.

[42] Shash M. FedCAPE Human Evaluation Survey. https://docs.google.com/forms/d/e/1FAIpQLSdj91VY2N__Ov4Z7XBkIxOAtgMCRmXpPb7N3PNyuxl8CxmSPA/viewform?usp=header. Accessed: 2025-08-12. 2025.

License

I, Mahmoud Said Hosny Elsayed Karamalla ,

(i) grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis **Federated Concept Alignment for Privacy-Preserving Explanations (FedCAPE)**, supervised by **Dr.Radwa ElShawi**;

(ii) grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

(iii) am aware of the fact that the author retains the rights specified in points 1 and 2;

(iv) confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Mahmoud Said Hosny Elsayed Karamalla,
12/08/2025