# F$^3$OCUS - Federated Finetuning of Vision-Language Foundation Models with Optimal Client Layer Updating Strategy via Multi-objective Meta-Heuristics

Supplementary Material

The **Supplementary Material** is organized as follows:

- **Section A** details all the algorithmic components of $F^3OCUS$ on server and clients.
- **Section B** shows the convergence analysis and theoretical motivation of our proposed method.
- **Section C** provides more details on the dataset and experimental setup as well as more implementation details including model architecture and training description.
- **Section D** provides analysis and discussion of experimental results reported in the main paper, and additional experimental results.
- **Section E** provides further discussion and clarification regarding different aspects of the main paper.

# A. Algorithms

## A.1. Client-level Layer Selection via Layerwise Neural Tangent Kernel

**Algorithm 1** Layer Selection via Principal Eigenvalue of LNTK

**Require:**
- Model $M$: Neural network model
- Logits `logits`: Output logits from the model
- Layers `layers`: Layers to evaluate

**Ensure:**
- Principal eigenvalues of NTK for the layers
- Top $k$ layers with the highest principal eigenvalues

**Main Algorithm**
Initialize `gradients_dict_ce` ← COMPUTEGRADIENTS($M$, `logits`, `layers`)
`principal_eigenvalues_dict` ← COMPUTEPRINCIPALEIGENVALUES(`gradients_dict_ce`, `layers`)
Reset gradients using `model.zero_grad()`
`TopLayers` ← SELECTTOPLAYERS(`principal_eigenvalues_dict`, $k$)
**return** `TopLayers`
**function** COMPUTEGRADIENTS($M$, `logits`, `layers`)
    Initialize `gradients_dict_ce` for all layers in `layers`
    **for** $b = 1$ to `logits.shape`[0] **do** ▷ Iterate over batch size
        **for** $i = 1$ to `logits.shape`[1] **do** ▷ Iterate over number of classes
            **for** each `name`, `param` in `model.named_parameters()` **do**
                **if** `name` ∈ `layers` **and** `param.grad` ≠ `None` **then**
                    `gradients_dict_ce[name]` ← `gradients_dict_ce[name]` +(`param.grad`)$^2$
                **end if**
            **end for**
        **end for**
    **end for**
    **return** `gradients_dict_ce`
**end function**
**function** COMPUTEPRINCIPALEIGENVALUES(`gradients_dict`, `layers`)
    Initialize `principal_eigenvalues_dict`
    **for** each `name`, `gradients` in `gradients_dict.items()` **do**
        **if** `name` ∈ `layers` **and** `gradients.numel()` $> 0$ **then**
            `J` ← `gradients.view(gradients.shape[0], -1)` ▷ Flatten gradients
            `NTK` ← `J @ J.T` ▷ Compute NTK matrix
            `eigenvalues` ← `torch.linalg.eigvalsh(NTK.cpu())`
            `principal_eigenvalues_dict[name]` ← `torch.max(eigenvalues)`
        **end if**
    **end for**
    **return** `principal_eigenvalues_dict`
**end function**
**function** SELECTTOPLAYERS(`principal_eigenvalues_dict`, $k$)
    Sort layers by principal eigenvalues in descending order
    Select top $k$ layers
    **return** Selected layers
**end function**

## A.2. Layer Selection on server using Genetic Algorithm

**Algorithm 2** Genetic Algorithm for Layer Selection

**Require:** Client-specific layer importance scores, layers per client, population size, mutation rate, number of generations.
**Ensure:** Best layer assignment across clients.

$population \leftarrow$ INITIALIZE_POPULATION
**for** $generation \leftarrow 1$ to $num\_generations$ **do**
  $fronts \leftarrow$ NON_DOMINATED_SORT($population$)
  $new\_population \leftarrow \emptyset$
  **for all** $front$ **in** $fronts$ **do**
    **if** $|new\_population| + |front| \leq population\_size$ **then**
      Add $front$ to $new\_population$
    **else**
      Add the first $(population\_size - |new\_population|)$ elements of $front$ to $new\_population$
      **break**
    **end if**
  **end for**
  **while** $|new\_population| < population\_size$ **do**
    $parent1 \leftarrow$ SELECT($population$)
    $parent2 \leftarrow$ SELECT($population$)
    $child1, child2 \leftarrow$ CROSSOVER($parent1, parent2$)
    Add MUTATE($child1$) to $new\_population$
    **if** $|new\_population| < population\_size$ **then**
      Add MUTATE($child2$) to $new\_population$
    **end if**
  **end while**
  $population \leftarrow new\_population$
  $best_individual \leftarrow \text{argmin}_{ind \in population}$CALCULATE_DIVERSITY$(ind)$
  $best\_importance \leftarrow$ CALCULATE_IMPORTANCE($best\_individual$)
  $best\_diversity \leftarrow$ CALCULATE_DIVERSITY($best\_individual$)
  **Print:** "Generation", $generation$, "Best Importance:", $best\_importance$, "Best Diversity:", $best\_diversity$
**end for**
  **return** $best\_individual$

**Algorithm 3** INITIALIZE_POPULATION

**Require:** Client-specific layer importance scores, layers per client.
**Ensure:** Initialized population.

$population \leftarrow \emptyset$
**for** $i \leftarrow 1$ to $population\_size$ **do**
  $individual \leftarrow \emptyset$
  **for** $client\_idx \leftarrow 1$ to $num\_clients$ **do**
    $num\_layers \leftarrow layers\_per\_client[client\_idx]$
    $scores \leftarrow importance\_scores[client\_idx]$
    $probabilities \leftarrow$ NORMALIZE($scores$)
    $selected \leftarrow$ RANDOM_CHOICES($probabilities, num\_layers$)
    Add $selected$ to $individual$
  **end for**
  Add $individual$ to $population$
**end for**
  **return** $population$

**Algorithm 4** CALCULATE_DIVERSITY

**Require:** Individual layer assignments.
**Ensure:** Diversity score.

$layer\_counts \leftarrow \{0 \text{ for all layers}\}$
**for all** $client\_layers$ **in** individual **do**
  **for all** $layer$ **in** $client\_layers$ **do**
    $layer\_counts[layer] \leftarrow layer\_counts[layer] + 1$
  **end for**
**end for**
$mean \leftarrow \text{MEAN}(layer\_counts)$
$variance \leftarrow \text{VARIANCE}(layer\_counts)$
$diversity \leftarrow \sqrt{variance}$
**return** $diversity$

**Algorithm 5** CALCULATE_IMPORTANCE

**Require:** Individual layer assignments.
**Ensure:** Importance score.

$importance \leftarrow 0$
**for all** $(client\_idx, client\_layers)$ **in** individual **do**
  **for all** $layer$ **in** $client\_layers$ **do**
    $importance \leftarrow importance + scores[client\_idx][layer]$
  **end for**
**end forreturn** $importance$

**Algorithm 6** NON_DOMINATED_SORT

**Require:** Population.
**Ensure:** Fronts of non-dominated solutions.

$fronts \leftarrow \{\}$
**for all** $individual1$ **in** population **do**
  $dominance\_count \leftarrow 0$
  **for all** $individual2$ **in** population **do**
    **if** DOMINATES($individual1, individual2$) **then**
      Add $individual2$ to $dominated\_solutions[individual1]$
    **else if** DOMINATES($individual2, individual1$) **then**
      $dominance\_count \leftarrow dominance\_count + 1$
    **end if**
  **end for**
  **if** $dominance\_count == 0$ **then**
    Add $individual1$ to $fronts[0]$
  **end if**
**end for**
**return** $fronts$

**Algorithm 7** DOMINATES

**Require:** Two solutions, `solution1` and `solution2`.
**Ensure:** `True` if `solution1` Pareto-dominates `solution2`, otherwise `False`.

$imp1, div1 \leftarrow$ CALCULATE_IMPORTANCE$(solution1)$, CALCULATE_DIVERSITY$(solution1)$
$imp2, div2 \leftarrow$ CALCULATE_IMPORTANCE$(solution2)$, CALCULATE_DIVERSITY$(solution2)$
**return** $(imp1 \geq imp2$ **and** $div1 \leq div2)$ **and** $(imp1 > imp2$ **or** $div1 < div2)$

**Algorithm 8** SELECT

**Require:** Population of solutions.
**Ensure:** Selected individual from the first Pareto front.

$fronts \leftarrow$ NON_DOMINATED_SORT(population) ▷ Sort population into Pareto fronts
$selected\_front \leftarrow fronts[0]$ ▷ Focus on the first Pareto front
$distances \leftarrow$ CALCULATE_CROWDING_DISTANCE([`population[i]` for $i \in selected\_front$])
▷ Check for non-finite values in distances
**for** $i \leftarrow 1$ to $|distances|$ **do**
    **if not** ISFINITE(distances[i]) **then**
        $distances[i] \leftarrow 1e-6$
    **end if**
**end for**
$epsilon \leftarrow 1e-6$ ▷ Add a small value to avoid zero probabilities
$selection\_probs \leftarrow [(dist + epsilon)/\sum \text{distances}\ \forall dist \in distances]$
$selected\_index \leftarrow$ RANDOM_CHOICES($selected\_front$, weights=$selection\_probs$, k=1) **return** `population[selected_index]`

**Algorithm 9** CALCULATE_CROWDING_DISTANCE

**Require:** A front of solutions.
**Ensure:** Crowding distances for each solution in the front.

$num\_individuals \leftarrow |front|$
$distances \leftarrow$ [0.0 for each solution in front]
**for** $m \leftarrow 1$ to 2 **do** ▷ Loop over objectives: 1 for importance, 2 for diversity
    **if** $m = 1$ **then**
        Sort $front$ **by** CALCULATE_IMPORTANCE
    **else**
        Sort $front$ **by** CALCULATE_DIVERSITY
    **end if**
    $distances[0] \leftarrow distances[-1] \leftarrow \infty$ ▷ Boundary solutions are always selected
    **for** $i \leftarrow 2$ to $num\_individuals - 1$ **do**
        **if** $m = 1$ **then**
            $distances[i] \leftarrow distances[i] + (\text{CALCULATE\_IMPORTANCE}(front[i+1]) - \text{CALCULATE\_IMPORTANCE}(front[i-1]))$
        **else**
            $distances[i] \leftarrow distances[i] + (\text{CALCULATE\_DIVERSITY}(front[i+1]) - \text{CALCULATE\_DIVERSITY}(front[i-1]))$
        **end if**
    **end for**
**end for**
**return** $distances$

**Algorithm 10** CROSSOVER

**Require:** Two parent solutions, `parent1` and `parent2`.
**Ensure:** Two child solutions, `child1` and `child2`.

$child1, child2 \leftarrow \emptyset, \emptyset$
**for** $client\_idx \leftarrow 1$ to $num\_clients$ **do**
    $combined \leftarrow$ UNION($parent1[client\_idx]$, $parent2[client\_idx]$)
    $child1[client\_idx] \leftarrow$ RANDOM_SAMPLE($combined$, $layers\_per\_client[client\_idx]$)
    $child2[client\_idx] \leftarrow$ RANDOM_SAMPLE($combined$, $layers\_per\_client[client\_idx]$)
**end forreturn** $child1, child2$

**Algorithm 11** MUTATE

**Require:** Solution `individual`.
**Ensure:** Mutated solution.

**if** RANDOM(0, 1) $< mutation\_rate$ **then**
    $client\_idx \leftarrow$ RANDOM_INTEGER($0, num\_clients - 1$)
    $num\_layers \leftarrow layers\_per\_client[client\_idx]$
    $individual[client\_idx] \leftarrow$ RANDOM_SAMPLE($range(num\_layers)$, $num\_layers$)
**end ifreturn** $individual$

## A.3. Layer Selection on server using MOPSO Algorithm

**Algorithm 12** MOPSO with Pareto Optimization

**Require:** Client-specific layer importance scores, layers per client, population size, number of iterations, inertia weight, cognitive and social constants.
**Ensure:** Pareto-optimal set of layer assignments across clients.

$population, velocities \leftarrow$ INITIALIZE_PARTICLES
$personal\_best \leftarrow population$
$personal_best_values \leftarrow \{($CALCULATE_IMPORTANCE$(p),$ CALCULATE_DIVERSITY$(p)) \, \forall p \in population\}$
$pareto\_archive \leftarrow$ NON_DOMINATED_SORT($population$)
$global\_best \leftarrow$ **random_choice**($pareto\_archive$)
**for** $iteration \leftarrow 1$ to $num\_iterations$ **do**
    **for** $i \leftarrow 1$ to $population\_size$ **do**
        $population[i], velocities[i] \leftarrow$ UPDATE_VELOCITY_POSITION($population[i], velocities[i], personal\_best[i], global\_best$)
        $importance \leftarrow$ CALCULATE_IMPORTANCE($population[i]$)
        $diversity \leftarrow$ CALCULATE_DIVERSITY($population[i]$)
        **if** $importance \geq personal\_best\_values[i][0]$ **and** $diversity \leq personal\_best\_values[i][1]$ **then**
            $personal\_best[i] \leftarrow population[i]$
            $personal\_best\_values[i] \leftarrow (importance, diversity)$
        **end if**
    **end for**
    $pareto\_archive \leftarrow$ NON_DOMINATED_SORT($population$)
    $global\_best \leftarrow$ **random_choice**($pareto\_archive$)
    $best\_importance \leftarrow$ CALCULATE_IMPORTANCE($global\_best$)
    $best\_diversity \leftarrow$ CALCULATE_DIVERSITY($global\_best$)
    **Print:** "Iteration", $iteration$, "Pareto Set Size:", $|pareto\_archive|$, "Best Importance:", $best\_importance$, "Best Diversity:", $best\_diversity$
**end for**
    **return** $pareto\_archive$

**Algorithm 13** INITIALIZE_PARTICLES

**Require:** Client-specific layer importance scores, layers per client.
**Ensure:** Initialized population of particles and velocities.

$population, velocities \leftarrow \emptyset$
**for** $i \leftarrow 1$ to $population\_size$ **do**
  $particle \leftarrow \emptyset, velocity \leftarrow \emptyset$
  **for** $client\_idx \leftarrow 1$ to $num\_clients$ **do**
    $num\_layers \leftarrow layers\_per\_client[client\_idx]$
    $selected\_layers \leftarrow$ RANDOM_SAMPLE($range(num\_layers), num\_layers$)
    Add $selected\_layers$ to $particle$
    Add **random velocities** to $velocity$
  **end for**
  Add $particle$ to $population$, $velocity$ to $velocities$
**end for**
  **return** $population, velocities$

**Algorithm 14** UPDATE_VELOCITY_POSITION

**Require:** Particle, velocity, personal best, global best.
**Ensure:** Updated particle and velocity.

$new\_velocity, new\_particle \leftarrow \emptyset$
**for** $client\_idx \leftarrow 1$ to $num\_clients$ **do**
  $current\_position \leftarrow particle[client\_idx]$
  $p\_best\_position \leftarrow personal\_best[client\_idx]$
  $g\_best\_position \leftarrow global\_best[client\_idx]$
  $new\_velocity\_client, new\_position\_client \leftarrow \emptyset$
  **for** $i \leftarrow 1$ to $|current\_position|$ **do**
    $r1, r2 \leftarrow$ **random numbers in** $[0, 1]$
    $cognitive \leftarrow cognitive\_constant \cdot r1 \cdot (p\_best\_position[i] - current\_position[i])$
    $social \leftarrow social\_constant \cdot r2 \cdot (g\_best\_position[i] - current\_position[i])$
    $v\_new \leftarrow inertia\_weight \cdot velocity[client\_idx][i] + cognitive + social$
    $position\_new \leftarrow$ ROUND($current\_position[i] + v\_new$)
    $position\_new \leftarrow$ CLAMP($position\_new, 0, num\_layers - 1$)
    Add $v\_new$ to $new\_velocity\_client$
    Add $position\_new$ to $new\_position\_client$
  **end for**
  $new\_position\_client \leftarrow$ REMOVE_DUPLICATES($new\_position\_client$)
  **while** $|new\_position\_client| < num\_layers\_for\_client$ **do**
    Add **random unique layers** to $new\_position\_client$
  **end while**
  Add $new\_velocity\_client$ to $new\_velocity$
  Add $new\_position\_client$ to $new\_particle$
**end for**
  **return** $new\_particle, new\_velocity$

**Algorithm 15** CALCULATE_DIVERSITY

**Require:** Particle layer assignments.
**Ensure:** Diversity score.

$layer\_counts \leftarrow \{0 \text{ for all layers}\}$
**for all** $client\_layers$ **in** particle **do**
  **for all** $layer$ **in** $client\_layers$ **do**
    $layer\_counts[layer] \leftarrow layer\_counts[layer] + 1$
  **end for**
**end for**
$mean \leftarrow \text{MEAN}(layer\_counts)$
$variance \leftarrow \text{VARIANCE}(layer\_counts)$
$diversity \leftarrow \sqrt{variance}$
**return** $diversity$

**Algorithm 16** CALCULATE_IMPORTANCE

**Require:** Particle layer assignments.
**Ensure:** Importance score.

$importance \leftarrow 0$
**for all** $(client\_idx, client\_layers)$ **in** particle **do**
  **for all** $layer$ **in** $client\_layers$ **do**
    $importance \leftarrow importance + importance\_scores[client\_idx][layer]$
  **end for**
**end for**
**return** $importance$

**Algorithm 17** NON_DOMINATED_SORT

**Require:** Population of solutions.
**Ensure:** Pareto archive of non-dominated solutions.

$fronts \leftarrow [[]]$ ▷ Initialize empty Pareto fronts
$pareto\_archive \leftarrow []$ ▷ Initialize Pareto archive
**for** $i \leftarrow 1$ to $|population|$ **do**
  $is\_dominated \leftarrow$ **False**
  **for all** $individual2 \in population$ **do**
    **if** DOMINATES($individual2$, $population[i]$) **then**
      $is\_dominated \leftarrow$ **True**
      **break**
    **end if**
  **end for**
  **if not** $is\_dominated$ **then**
    Add $population[i]$ to $pareto\_archive$
  **end if**
**end for**
**return** $pareto\_archive$

**Algorithm 18** DOMINATES

**Require:** Two individuals $individual1$ and $individual2$.
**Ensure:** **True** if $individual1$ dominates $individual2$, otherwise **False**.

$imp1, div1 \leftarrow$ CALCULATE_IMPORTANCE($individual1$), CALCULATE_DIVERSITY($individual1$)
$imp2, div2 \leftarrow$ CALCULATE_IMPORTANCE($individual2$), CALCULATE_DIVERSITY($individual2$)
**if** $(imp1 \geq imp2$ **and** $div1 \leq div2)$ **and** $(imp1 > imp2$ **or** $div1 < div2)$ **then**
    **return True**
**else**
    **return False**
**end if**

## A.4. Layer Selection on server using Simulated Annealing Algorithm

**Algorithm 19** Simulated Annealing for Layer Assignment

**Require:** Client-specific layer importance scores, layers per client, initial and final temperatures, cooling rate, number of iterations.
**Ensure:** Pareto-optimal set of layer assignments across clients.

$current\_solution \leftarrow$ INITIALIZE_SOLUTION
$current\_importance \leftarrow$ CALCULATE_IMPORTANCE($current\_solution$)
$current\_diversity \leftarrow$ CALCULATE_DIVERSITY($current\_solution$)
$temperature \leftarrow initial\_temperature$
$pareto\_archive \leftarrow \{current\_solution\}$
**for** $iteration \leftarrow 1$ to $num\_iterations$ **do**
    $new\_solution \leftarrow$ PERTURB_SOLUTION($current\_solution$)
    $new\_importance \leftarrow$ CALCULATE_IMPORTANCE($new\_solution$)
    $new\_diversity \leftarrow$ CALCULATE_DIVERSITY($new\_solution$)
    **if** DOMINATES($new\_solution, current\_solution$) **or** ACCEPT_WORSE_SOLUTION($current\_importance, current\_diversity, neu$ **then**
        $current\_solution \leftarrow new\_solution$
        $current\_importance \leftarrow new\_importance$
        $current\_diversity \leftarrow new\_diversity$
    **end if**
    $pareto\_archive \leftarrow$ UPDATE_PARETO_ARCHIVE($pareto\_archive, current\_solution$)
    $temperature \leftarrow temperature \cdot cooling\_rate$
    **Print:** "Iteration", $iteration$, "Temp:", $temperature$, "Best Importance:", $current\_importance$, "Best Diversity:", $current\_diversity$, "Pareto Archive Size:", $|pareto\_archive|$
**end for**
    **return** $pareto\_archive$

**Algorithm 20** INITIALIZE_SOLUTION

**Require:** Client-specific layer importance scores, layers per client.
**Ensure:** Initial solution for layer assignments.

$solution \leftarrow \emptyset$
**for** $client\_idx \leftarrow 1$ to $num\_clients$ **do**
    $num\_layers \leftarrow layers\_per\_client[client\_idx]$
    $selected\_layers \leftarrow$ RANDOM_SAMPLE($range(num\_layers), num\_layers$)
    Add $selected\_layers$ to $solution$
**end for**
    **return** $solution$

**Algorithm 21** PERTURB_SOLUTION

**Require:** Current solution.
**Ensure:** Perturbed solution (neighbor).
1: $new\_solution \leftarrow current\_solution$
2: $client\_idx \leftarrow$ RANDOM_INTEGER$(0, num_clients)$
3: $num\_layers \leftarrow layers\_per\_client[client\_idx]$
4: $new\_layers \leftarrow$ RANDOM_SAMPLE$(range(num_layers), num_layers)$
5: $new\_solution[client\_idx] \leftarrow new\_layers$
**return** $new\_solution$

**Algorithm 22** DOMINATES

**Require:** Two solutions, `solution1` and `solution2`.
**Ensure:** `True` if `solution1` Pareto-dominates `solution2`, otherwise `False`.
1: $imp1, div1 \leftarrow$ CALCULATE_IMPORTANCE$(solution1)$, CALCULATE_DIVERSITY$(solution1)$
2: $imp2, div2 \leftarrow$ CALCULATE_IMPORTANCE$(solution2)$, CALCULATE_DIVERSITY$(solution2)$
3: **return** $(imp1 \geq imp2$ **and** $div1 \leq div2)$ **and** $(imp1 > imp2$ **or** $div1 < div2)$

**Algorithm 23** CALCULATE_DIVERSITY

**Require:** Solution layer assignments.
**Ensure:** Diversity score.
1: $layer\_counts \leftarrow \{0$ for all layers$\}$
2: **for all** $client\_layers$ **in** solution **do**
3: **for all** $layer$ **in** $client\_layers$ **do**
4: $layer\_counts[layer] \leftarrow layer\_counts[layer] + 1$
5: **end for**
6: **end for**
7: $mean \leftarrow$ MEAN$(layer_counts)$
8: $variance \leftarrow$ VARIANCE$(layer_counts)$
9: $diversity \leftarrow \sqrt{variance}$
**return** $diversity$

**Algorithm 24** CALCULATE_IMPORTANCE

**Require:** Solution layer assignments.
**Ensure:** Importance score.
1: $importance \leftarrow 0$
2: **for all** $(client\_idx, client\_layers)$ **in** solution **do**
3: **for all** $layer$ **in** $client\_layers$ **do**
4: $importance \leftarrow importance + importance\_scores[client\_idx][layer]$
5: **end for**
6: **end for**
**return** $importance$

**Algorithm 25** ACCEPT_WORSE_SOLUTION

**Require:** Current and new importance/diversity scores, temperature.
**Ensure:** Whether to accept the worse solution.
**if** $temperature \leq final\_temperature$ **then return False**
**end if**
$delta \leftarrow (new\_importance - current\_importance) + (current\_diversity - new\_diversity)$
$acceptance\_probability \leftarrow \exp(-delta/temperature)$
**return** RANDOM_VALUE $< acceptance\_probability$

**Algorithm 26** UPDATE_PARETO_ARCHIVE

**Require:** Current Pareto archive, new solution.
**Ensure:** Updated Pareto archive.
$non\_dominated \leftarrow \{s \in archive : \neg\text{DOMINATES}(new\_solution, s)\}$
**if** $\neg\exists s \in archive$ :DOMINATES$(s, new_solution)$ **then**
    Add $new\_solution$ to $non\_dominated$
**end if**
**return** $non\_dominated$

## A.5. Layer Selection on server using Ant Colony Optimization Algorithm

**Algorithm 27** Ant Colony Optimization for Layer Assignment

**Require:** Client-specific layer importance scores, layers per client, pheromone parameters, number of ants, number of iterations.
**Ensure:** Pareto-optimal set of layer assignments across clients.
$pareto\_archive \leftarrow \emptyset$
**for** $iteration \leftarrow 1$ to $num\_iterations$ **do**
    $ants_solutions \leftarrow \{$INITIALIZE_ANT_SOLUTION $\forall$ant $\in \{1, \ldots, num\_ants\}\}$
    **for all** $solution \in ants\_solutions$ **do**
        $importance \leftarrow$ CALCULATE_IMPORTANCE$(solution)$
        $diversity \leftarrow$ CALCULATE_DIVERSITY$(solution)$
        $pareto\_archive \leftarrow$ UPDATE_PARETO_ARCHIVE$(pareto_archive, solution)$
    **end for**
    UPDATE_PHEROMONES$(pareto_archive)$
    $best\_solution \leftarrow$ PICK_BEST_SOLUTION$(pareto_archive)$
    **Print:** "Iteration", $iteration$, "Pareto Archive Size:", $|pareto\_archive|$, "Best Importance:", CALCULATE_IMPORTANCE$(best_solution)$, "Best Diversity:", CALCULATE_DIVERSITY$(best_solution)$
**end for**
**return** $pareto\_archive$

**Algorithm 28** INITIALIZE_ANT_SOLUTION

**Require:** Client-specific layer importance scores, pheromone matrix, pheromone parameters.
**Ensure:** Single ant's solution for layer assignments.

$solution \leftarrow \emptyset$
**for** $client\_idx \leftarrow 1$ to $num\_clients$ **do**
    $num\_layers \leftarrow layers\_per\_client[client\_idx]$
    $importance\_scores \leftarrow importance\_scores[client\_idx]$
    $probabilities \leftarrow \{(\text{pheromone}[layer]^{\alpha} \cdot \text{importance}[layer]^{\beta}) \, \forall layer \in \text{num\_layers}\}$
    Normalize $probabilities$
    $selected\_layers \leftarrow$ RANDOM_CHOICES$(range(num_layers), weights = probabilities, k = num_layers)$
    Remove duplicates and fill missing layers until $|selected\_layers| = num\_layers$
    Add $selected\_layers$ to $solution$
**end for**
**return** $solution$

**Algorithm 29** CALCULATE_DIVERSITY

**Require:** Solution layer assignments.
**Ensure:** Diversity score.

$layer\_counts \leftarrow \{0 \text{ for all layers}\}$
**for all** $client\_layers \in solution$ **do**
    **for all** $layer \in client\_layers$ **do**
        $layer\_counts[layer] \leftarrow layer\_counts[layer] + 1$
    **end for**
**end for**
$mean \leftarrow$ MEAN$(layer_counts)$
$variance \leftarrow$ VARIANCE$(layer_counts)$
$diversity \leftarrow \sqrt{variance}$
**return** $diversity$

**Algorithm 30** CALCULATE_IMPORTANCE

**Require:** Solution layer assignments.
**Ensure:** Importance score.

$importance \leftarrow 0$
**for all** $(client\_idx, client\_layers) \in solution$ **do**
    **for all** $layer \in client\_layers$ **do**
        $importance \leftarrow importance + importance\_scores[client\_idx][layer]$
    **end for**
**end for**
**return** $importance$

**Algorithm 31** UPDATE_PHEROMONES

**Require:** Current Pareto archive.
**Ensure:** Updated pheromone matrix.

**for** $client\_idx \leftarrow 1$ to $num\_clients$ **do**
  **for** $layer\_idx \leftarrow 1$ to $num\_layers$ **do**
    pheromone$[client_idx][layer_idx] \leftarrow$ pheromone$[client_idx][layer_idx] \cdot (1 -$ pheromone_evaporation$)$
  **end for**
**end for**
**for all** $solution \in pareto\_archive$ **do**
  **for all** $(client\_idx, client\_layers) \in solution$ **do**
    **for all** $layer \in client\_layers$ **do**
      pheromone$[client_idx][layer] \leftarrow$ pheromone$[client_idx][layer] +$ pheromone_deposit
    **end for**
  **end for**
**end for**

**Algorithm 32** DOMINATES

**Require:** Two solutions, `solution1` and `solution2`.
**Ensure:** `True` if `solution1` Pareto-dominates `solution2`, otherwise `False`.

$imp1, div1 \leftarrow$ CALCULATE_IMPORTANCE$(solution1)$, CALCULATE_DIVERSITY$(solution1)$
$imp2, div2 \leftarrow$ CALCULATE_IMPORTANCE$(solution2)$, CALCULATE_DIVERSITY$(solution2)$
**return** $(imp1 \geq imp2$ **and** $div1 \leq div2)$ **and** $(imp1 > imp2$ **or** $div1 < div2)$

**Algorithm 33** UPDATE_PARETO_ARCHIVE

**Require:** Current Pareto archive, new solution.
**Ensure:** Updated Pareto archive.

$non_dominated \leftarrow \{s \in archive : \neg$DOMINATES$(new_solution, s)\}$
**if** $\neg\exists s \in archive$ :DOMINATES$(s, new_solution)$ **then**
  Add $new\_solution$ to $non\_dominated$
**end if**
  **return** $non\_dominated$

**Algorithm 34** PICK_BEST_SOLUTION

**Require:** Pareto-optimal solutions, weights for importance and diversity.
**Ensure:** Best solution based on weighted score.

$best\_solution \leftarrow \emptyset, best\_score \leftarrow -\infty$
**for all** $solution \in pareto\_set$ **do**
  $importance \leftarrow$ CALCULATE_IMPORTANCE$(solution)$
  $diversity \leftarrow$ CALCULATE_DIVERSITY$(solution)$
  $score \leftarrow weight\_importance \cdot importance - weight\_diversity \cdot diversity$
  **if** $score > best\_score$ **then**
    $best\_solution \leftarrow solution$
    $best\_score \leftarrow score$
  **end if**
**end for**
  **return** $best\_solution$

### A.6. Layer Selection on server using Artificial Bee Colony Algorithm

**Algorithm 35** Artificial Bee Colony Optimization with Pareto Optimization

**Require:** Client-specific layer importance scores, layers per client, number of bees, number of iterations, limit of trials for scout bees.
**Ensure:** Pareto-optimal set of layer assignments across clients.

$bee_solutions \leftarrow \{$INITIALIZE_SOLUTION $\forall$bee $\in \{1, \ldots, num\_bees\}\}$
$trial\_counter \leftarrow [0]$ for each bee
$pareto\_archive \leftarrow \emptyset$
**for** $iteration \leftarrow 1$ to $num\_iterations$ **do**
▷ Employed Bees Phase
    EMPLOYED_BEES($bee\_solutions$, $trial\_counter$)
▷ Onlooker Bees Phase
    ONLOOKER_BEES($bee\_solutions$)
▷ Scout Bees Phase
    SCOUT_BEES($bee\_solutions$, $trial\_counter$)
▷ Update Pareto Archive
    **for all** $solution \in bee\_solutions$ **do**
        $pareto\_archive \leftarrow$ UPDATE_PARETO_ARCHIVE($pareto\_archive$, $solution$)
    **end for**
▷ Log Progress
    $best\_solution \leftarrow$ PICK_BEST_SOLUTION($pareto\_archive$)
    **Print:** "Iteration", $iteration$, "Pareto Archive Size:", $|pareto\_archive|$, "Best Importance:", CALCULATE_IMPORTANCE($best\_solution$), "Best Diversity:", CALCULATE_DIVERSITY($best\_solution$)
**end for**
    **return** $pareto\_archive$

**Algorithm 36** INITIALIZE_SOLUTION

**Require:** Client-specific layer importance scores, layers per client.
**Ensure:** Initial solution for layer assignments.

$solution \leftarrow \emptyset$
**for** $client\_idx \leftarrow 1$ to $num\_clients$ **do**
    $num\_layers \leftarrow layers\_per\_client[client\_idx]$
    $importance\_scores \leftarrow importance\_scores[client\_idx]$
    $probabilities \leftarrow \{$importance$[layer]$/sum(importance) $\forall$layer$\}$
    $selected\_layers \leftarrow$ RANDOM_SAMPLE($range(num\_layers), k = num\_layers$)
    Add $selected\_layers$ to $solution$
**end for**
    **return** $solution$

**Algorithm 37** ONLOOKER_BEES

**Require:** Bee solutions.
**Ensure:** Updated bee solutions based on fitness.

$total\_fitness \leftarrow \sum_{s \in bee\_solutions} (\text{CALCULATE\_IMPORTANCE}(s) - \text{CALCULATE\_DIVERSITY}(s))$
**if** $total\_fitness = 0$ **then**
    $total\_fitness \leftarrow 1$ ▷ Prevent division by zero
**end if**
$probabilities \leftarrow \left[ \frac{\text{CALCULATE\_IMPORTANCE}(s) - \text{CALCULATE\_DIVERSITY}(s)}{total\_fitness} \; \forall s \in bee\_solutions \right]$
**for** $i \leftarrow 1$ to |bee_solutions| **do**
    **if** RANDOM(0, 1) $< probabilities[i]$ **then**
        $new\_solution \leftarrow$ PERTURB_SOLUTION($bee\_solutions[i]$)
        **if** DOMINATES($new\_solution$, $bee\_solutions[i]$) **then**
            $bee\_solutions[i] \leftarrow new\_solution$
        **end if**
    **end if**
**end for**

**Algorithm 38** SCOUT_BEES

**Require:** Bee solutions, trial counter, limit of trials.
**Ensure:** Updated bee solutions by replacing abandoned ones.

**for** $i \leftarrow 1$ to |bee_solutions| **do**
    **if** $trial\_counter[i] \geq$ limit **then**
        $bee\_solutions[i] \leftarrow$ INITIALIZE_SOLUTION ▷ Replace with new random solution
        $trial\_counter[i] \leftarrow 0$ ▷ Reset trial counter
    **end if**
**end for**

**Algorithm 39** EMPLOYED_BEES

**Require:** Bee solutions, trial counter.
**Ensure:** Updated bee solutions after local exploitation.

**for** $i \leftarrow 1$ to $num\_bees$ **do**
    $new\_solution \leftarrow$ PERTURB_SOLUTION($bee\_solutions[i]$)
    **if** DOMINATES($new\_solution$, $bee\_solutions[i]$) **then**
        $bee\_solutions[i] \leftarrow new\_solution$
        $trial\_counter[i] \leftarrow 0$
    **else**
        $trial\_counter[i] \leftarrow trial\_counter[i] + 1$
    **end if**
**end for**

**Algorithm 40** DOMINATES

**Require:** Two solutions, `solution1` and `solution2`.
**Ensure:** `True` if `solution1` Pareto-dominates `solution2`, otherwise `False`.

$imp1, div1 \leftarrow$ CALCULATE_IMPORTANCE($solution1$), CALCULATE_DIVERSITY($solution1$)
$imp2, div2 \leftarrow$ CALCULATE_IMPORTANCE($solution2$), CALCULATE_DIVERSITY($solution2$)
**return** $(imp1 \geq imp2$ **and** $div1 \leq div2)$ **and** $(imp1 > imp2$ **or** $div1 < div2)$

**Algorithm 41** PERTURB_SOLUTION

**Require:** Current solution.
**Ensure:** Perturbed solution (neighbor).
$new\_solution \leftarrow$ deep copy of current solution
$client\_idx \leftarrow$ RANDOM_INTEGER$(0, num_clients - 1)$
$num\_layers \leftarrow layers\_per\_client[client\_idx]$
$current\_layers \leftarrow new\_solution[client\_idx]$
$new\_layer \leftarrow$ RANDOM_CHOICE$(range(num_layers) \setminus current_layers)$
Replace one randomly selected layer in $current\_layers$ with $new\_layer$
$new\_solution[client\_idx] \leftarrow current\_layers$
**return** $new\_solution$

**Algorithm 42** CALCULATE_DIVERSITY

**Require:** Solution layer assignments.
**Ensure:** Diversity score.
$layer\_counts \leftarrow \{0$ for all layers$\}$
**for all** $client\_layers \in solution$ **do**
  **for all** $layer \in client\_layers$ **do**
    $layer\_counts[layer] \leftarrow layer\_counts[layer] + 1$
  **end for**
**end for**
$mean \leftarrow$ MEAN$(layer_counts)$
$variance \leftarrow$ VARIANCE$(layer_counts)$
$diversity \leftarrow \sqrt{variance}$
**return** $diversity$

**Algorithm 43** CALCULATE_IMPORTANCE

**Require:** Solution layer assignments.
**Ensure:** Importance score.
$importance \leftarrow 0$
**for all** $(client\_idx, client\_layers) \in solution$ **do**
  **for all** $layer \in client\_layers$ **do**
    $importance \leftarrow importance + importance\_scores[client\_idx][layer]$
  **end for**
**end for**
**return** $importance$

**Algorithm 44** UPDATE_PARETO_ARCHIVE

**Require:** Current Pareto archive, new solution.
**Ensure:** Updated Pareto archive.
$non_dominated \leftarrow \{s \in archive : \neg$DOMINATES$(new_solution, s)\}$
**if** $\neg\exists s \in archive$ :DOMINATES$(s, new_solution)$ **then**
  Add $new\_solution$ to $non\_dominated$
**end if**
**return** $non\_dominated$

**Algorithm 45** PICK_BEST_SOLUTION

**Require:** Pareto-optimal solutions, weights for importance and diversity.
**Ensure:** Best solution based on weighted score.

$best\_solution \leftarrow \emptyset, best\_score \leftarrow -\infty$
**for all** $solution \in pareto\_set$ **do**
  $importance \leftarrow$ CALCULATE_IMPORTANCE($solution$)
  $diversity \leftarrow$ CALCULATE_DIVERSITY($solution$)
  $score \leftarrow weight\_importance \cdot importance - weight\_diversity \cdot diversity$
  **if** $score > best\_score$ **then**
    $best\_solution \leftarrow solution$
    $best\_score \leftarrow score$
  **end if**
**end for**
  **return** $best\_solution$

## B. Convergence Analysis: Full Proofs and Theoretical Motivation

**Lemma B.1.** Based on Assumption 1, we have:

$$\mathbb{E}[F(\theta_{t+1})] - \mathbb{E}[F(\theta_t)] \leq \frac{1}{2\gamma}\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi^l(\theta_t)\right\|^2\right] + \mathbb{E}\left\langle \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t), \theta_{t+1}-\theta_t\right\rangle + \gamma\mathbb{E}\left[\|\theta_{t+1}-\theta_t\|^2\right] \quad (22)$$

**Proof.** Based on $\gamma$-smoothness in Assumption 1, we compute the loss decay as follows:

$$\mathbb{E}[F(\theta_{t+1})] - \mathbb{E}[F(\theta_t)] \leq \mathbb{E}\langle\nabla F(\theta_t), \theta_{t+1}-\theta_t\rangle + \frac{\gamma}{2}\mathbb{E}[\|\theta_{t+1}-\theta_t\|^2] \quad (23)$$

$$= \mathbb{E}\left\langle \nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t) + \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t, \theta_{t+1}-\theta_t)\right\rangle + \frac{\gamma}{2}\mathbb{E}\|\theta_{t+1}-\theta_t\|^2 \quad (24)$$

$$= \mathbb{E}\left\langle \nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t), \theta_{t+1}-\theta_t\right\rangle + \sum_{l\in\mathcal{L}_t}\mathbb{E}\left\langle \nabla\psi_t^l(\theta_t), \theta_{t+1}-\theta_t\right\rangle + \frac{\gamma}{2}\mathbb{E}[\|\theta_{t+1}-\theta_t\|^2]. \quad (25)$$

Now, using Young's inequality,

$$\mathbb{E}\left\langle \nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t), \theta_{t+1}-\theta_t\right\rangle \leq \frac{1}{2\gamma}\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] + \frac{\gamma}{2}\mathbb{E}[\|\theta_{t+1}-\theta_t\|^2]. \quad (26)$$

Plugging it back into the inequality gives:

$$\mathbb{E}[F(\theta_{t+1})] - \mathbb{E}[F(\theta_t)] \leq \frac{1}{2\gamma}\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] + \mathbb{E}\left[\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t), \theta_{t+1}-\theta_t\right] + \gamma\mathbb{E}\left[\|\theta_{t+1}-\theta_t\|^2\right]. \quad (27)$$

Now we analyze $\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right]$ and find its upper bound below:

We decompose the term $\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right]$ using Jensen's inequality as:

$$\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] \leq 2\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l F(\theta_t)\|^2 + 2\|\sum_{l\in\mathcal{L}_t}\nabla_l F(\theta_t) - \nabla_l\psi_t^l(\theta_t)\|^2, \quad (28)$$

For the first term, we get:

$$\mathbb{E}\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l F(\theta_t)\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{l\notin\mathcal{L}_t}\nabla_l F(\theta_t)\right\|^2\right] \tag{29}$$

For the second term, we get:

$$\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l \psi_t^l(\theta_t)\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\sum_{i\in\mathcal{N}}\alpha_{i,t}\nabla_l F_i(\theta_t) - \sum_{l\in\mathcal{L}_t}\sum_{i\in\mathcal{N}}\alpha_{i,t}m_{i,t}^l\nabla_l F_i(\theta_t)\right\|^2\right] \tag{30}$$

$$= \sum_{l\in\mathcal{L}_t}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{N}}\frac{\alpha_{i,t}m_{i,l}^t - \alpha_{i,t}}{\sqrt{\alpha_{i,t}}}\sqrt{\alpha_{i,t}^l}\left(\nabla_l F_{i,t}^l(\theta_t) - \nabla_l F_l(\theta_t)\right)\right\|^2\right] \tag{31}$$

$$\leq \sum_{l\in\mathcal{L}_t}\left[\sum_{i\in\mathcal{N}}\frac{(\alpha_{i,t}m_{i,t}^l - \alpha_{i,t})^2}{\alpha_{i,t}}\right]\sum_{i\in\mathcal{N}}\alpha_i\mathbb{E}\left[\|\nabla_l F_i(\theta_t) - \nabla_l F(\theta_t)\|^2\right] \tag{32}$$

$$\leq \sum_{l\in\mathcal{L}_t}\sum_{i\in\mathcal{N}}\alpha_{i,t}(m_{i,t}^l - 1)^2 k_l^2. \tag{33}$$

where (32) is based on Cauchy-Schwartz inequality and (33) is based on Assumption 3. Now based on this, we prove the convergence.

We derive the value of $\mathbb{E}\left\langle \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t), \theta_{t+1} - \theta_t\right\rangle$ as follows:

$$\mathbb{E}\left\langle \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t), \theta_{t+1} - \theta_t\right\rangle = \mathbb{E}\left[\left\langle \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t), -\eta\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\rangle\right]$$

$$= -\eta\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right]. \tag{34}$$

We get an upper bound for the term $\mathbb{E}\left[\|\theta_{t+1} - \theta_t\|^2\right]$ as follows:

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_t\|^2\right] = \mathbb{E}\left[\left\|\eta\sum_{l\in\mathcal{L}_t}\sum_{i\in\mathcal{N}}\alpha_{i,t}m_{i,t}^l G_{i,t}(\theta_t; B_t)\right\|^2\right] \tag{35}$$

$$\leq \eta^2\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + \eta^2\sigma^2, \tag{36}$$

where (36) is based on Assumption 2.

Based on the result in Lemma B.1, we get:

$$\mathbb{E}[F(\theta_{t+1})] - \mathbb{E}[F(\theta_t)] \leq \frac{1}{2\gamma}\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] - \eta\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + \gamma\eta^2\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + \gamma\eta^2\sigma^2 \tag{37}$$

$$= \frac{1}{2\gamma}\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] - \eta(1-\gamma\eta)\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla\psi_t^l(\theta_t)\right\|^2\right] + \gamma\eta^2\sigma^2. \tag{38}$$

Arranging the terms in (38), we get:

$$\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla\psi_t^l(\theta_t)\right\|^2\right] \leq \frac{1}{\eta(1-\gamma\eta)}\Big[\mathbb{E}[F(\theta_t)] - \mathbb{E}[F(\theta_{t+1})]\Big] + \frac{1}{2\gamma\eta(1-\gamma\eta)}\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi^l(\theta_t)\right\|^2\right] + \frac{\gamma\eta}{(1-\gamma\eta)}\sigma^2. \tag{39}$$

By Jensen's inequality, we have:

$$\mathbb{E}\Big[\|\nabla F(\theta_t)\|^2\Big] = \mathbb{E}\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla\psi_t^l(\theta_t) + \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] \tag{40}$$

$$\leq 2\mathbb{E}\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + 2\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] \tag{41}$$

$$\tag{1}$$

Combining (39) and (41) gives:

$$\mathbb{E}\Big[\|\nabla F(\theta_t)\|^2\Big] \leq \frac{2}{(1-\gamma\eta)}\Big[\mathbb{E}[F(\theta_t)] - \mathbb{E}[F(\theta_{t+1})]\Big] + \Big(\frac{1}{\gamma\eta(1-\gamma\eta)} + 2\Big)\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + \frac{2\gamma\eta}{(1-\gamma\eta)}\sigma^2. \tag{42}$$

Summing both sides of (42) over $t = 0, 1, \ldots, T-1$ and divide by $T$, we get $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\theta_t)\|^2\Big]$:

$$\leq \frac{2}{\eta(1-\gamma\eta)T}\Big[\mathbb{E}[F(\theta_0)] - \mathbb{E}[F(\theta_T)]\Big] + \frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1}{\gamma\eta(1-\gamma\eta)} + 2\Big)\mathbb{E}\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + \frac{2\gamma\eta}{(1-\gamma\eta)}\sigma^2. \tag{43}$$

$$\leq \frac{2}{\eta(1-\gamma\eta)T}\Big[F(\theta^0) - F(\theta^*)\Big] + \frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1}{\gamma\eta(1-\gamma\eta)} + 2\Big)\mathbb{E}\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + \frac{2\gamma\eta}{(1-\gamma\eta)}\sigma^2. \tag{44}$$

$$\leq \frac{2}{\eta(1-\gamma\eta)T}\Big[F(\theta^0) - F(\theta^*)\Big] + \frac{2\gamma\eta}{(1-\gamma\eta)}\sigma^2 + \frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1}{\gamma\eta(1-\gamma\eta)} + 2\Big)\Big(\mathbb{E}\left[\left\|\sum_{l\notin\mathcal{L}_t}\nabla_l F(\theta_t)\right\|^2\right] + \sum_{l\in\mathcal{L}_t}\sum_{i\in\mathcal{N}}\alpha_{i,t}(m_{i,t}^l - 1)^2 k_l^2\Big). \tag{45}$$

This concludes our convergence proof.

**Proof of general case:** Now, we consider the general case where the number of steps per round $\tau > 1$. Below, we analyze the convergence and observe that the impact of $\Big(\mathbb{E}\left[\left\|\sum_{l\notin\mathcal{L}_t}\nabla_l F(\theta_t)\right\|^2\right] + \sum_{l\in\mathcal{L}_t}\sum_{i\in\mathcal{N}}\alpha_{i,t}(m_{i,t}^l - 1)^2 k_l^2\Big)$ is similar to that in Theorem 1.

Let $C' \triangleq 1 - 4\gamma\tau - 8\eta\gamma^2\tau^2(\tau-1) - 32\gamma^3\eta^2\tau^2(\tau-1) > 0$ and $A_t \triangleq \eta + 2\gamma^2\tau(\tau-1)$. With Assumptions 1–3, we have $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\theta_t)\|^2\Big]$ :

$$\leq \frac{2}{\eta\gamma C'T}\Big[F(\theta^0) - F(\theta^*)\Big] + \frac{4A_\tau}{C'}\sigma^2 + \frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1}{\eta\gamma C'} + 2\Big)\Big(\mathbb{E}\left[\left\|\sum_{l\notin\mathcal{L}_t}\nabla_l F(\theta_t)\right\|^2\right] + \sum_{l\in\mathcal{L}_t}\sum_{i\in\mathcal{N}}\alpha_{i,t}(m_{i,t}^l - 1)^2 k_l^2\Big). \tag{46}$$

**Proof.** The term $\mathbb{E}\left\langle \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t), \theta_{t+1} - \theta_t\right\rangle$ in Lemma B.1 denotes the client drift due multiple local gradient updating steps. Using the inequality $\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{\|\mathbf{a}\|^2}{2} + \frac{\|\mathbf{b}\|^2}{2}$, and Assumption 1, the upper bound of this can be derived as

follows:

$$\mathbb{E}\left\langle \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t), \theta_{t+1}-\theta_t \right\rangle = -\eta \sum_{k=0}^{\tau-1} \mathbb{E}\left\langle \left( \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t), \sum_{l\in\mathcal{L}_t}\sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla_l F_i(\theta_{t,k}^l) \right) \right\rangle \tag{47}$$

$$= -\eta \sum_{k=0}^{\tau-1} \mathbb{E}\left\langle \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t), \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t) \right\rangle$$

$$+ \eta \sum_{k=0}^{\tau-1} \mathbb{E}\left\langle \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t), \nabla_l \psi_t^l(\theta_t) - \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla F_i(\theta_{i,t,k}) \right\rangle \tag{48}$$

$$\leq -\frac{\eta\tau}{2} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t) \right\|^2 \right] + \frac{\eta\gamma^2}{2} \sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \alpha_{i,t} m_{i,t}^l (\theta_t - \theta_{i,t,k}) \right\|^2 \right]. \tag{49}$$

The term $\mathbb{E}\left[\|\theta_{t+1}-\theta_t\|^2\right]$ is upper-bounded as follows (using Assumptions 1, 2 and Jensen's inequality):

$$\mathbb{E}\left[\|\theta_{t+1}-\theta_t\|^2\right] \leq \eta^2\tau \mathbb{E}\left[ \left\| \sum_{k=0}^{\tau-1} \sum_{l\in\mathcal{L}_t} \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla_l F_i(\theta_{i,t,k}) \right\|^2 \right] + \eta^2\tau\sigma^2 \tag{50}$$

$$\leq \eta^2\tau \mathbb{E}\left[ \left\| \sum_{k=0}^{\tau-1} \sum_{l\in\mathcal{L}_t} \left( \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla F_i(\theta_{it,k}) - \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla F_i(\theta_t) + \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla F_i(\theta_t) \right) \right\|^2 \right] + \eta^2\tau\sigma^2 \tag{51}$$

$$\leq 2\eta^2\tau \sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla_l F_i^l(\theta_{t,k}) - \sum_{l\in\mathcal{L}_t} \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla F^l(\theta^t) \right\|^2 \right]$$

$$+ 2\eta^2\tau \sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t) \right\|^2 \right] + \eta^2\tau\sigma^2 \tag{52}$$

$$\leq 2\eta^2\gamma^2\tau \sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l (\theta_{i,t,k} - \theta_t) \right\|^2 \right] + 2\eta^2\tau^2 \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t) \right\|^2 \right] + \eta^2\tau\sigma^2 \tag{53}$$

(2)

The upper bound of $\sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \alpha_{i,t} m_{i,t}^l (\theta_t - \theta_{i,t,k}) \right\|^2 \right]$ can be expressed as:

$$\sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \alpha_{i,t} m_{i,t}^l (\theta_t - \theta_{i,t,k}) \right\|^2 \right] \leq \sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l (\theta_t - \theta_{i,t,k}) \right\|^2 \right] \tag{54}$$

$$\leq 8\eta^2\tau^2(\tau-1) \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l \nabla F_i(\theta_t) \right\|^2 \right] + \sum_{l\in\mathcal{L}_t} \sum_{i\in\mathcal{N}} \alpha_{i,t} m_{i,t}^l 4\eta^2\tau^2(\tau-1)\sigma^2 \tag{55}$$

$$= 8\gamma^2\tau^2(\tau-1) \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \nabla_l \psi_t^l(\theta_t) \right\|^2 \right] + 4\eta^2\tau^2(\tau-1)\sigma^2. \tag{56}$$

Let us denote $\sum_{k=0}^{\tau-1} \mathbb{E}\left[ \left\| \sum_{l\in\mathcal{L}_t} \alpha_{i,t} m_{i,t}^l (\theta_t - \theta_{i,t,k}) \right\|^2 \right]$ as $\lambda_1$.

Substituting (49), (53), (56) into (22), we get the following:

$$\mathbb{E}[F(\theta_{t+1})] - \mathbb{E}[F(\theta_t)] \leq \frac{1}{2\gamma}\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] - \frac{\eta\tau}{2}\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right]$$
$$+ (\frac{\eta\gamma^2}{2} + 2\eta^2\gamma^2\tau^2)\lambda_1 + 2\eta^2\tau^2\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + \eta^2\tau\sigma^2 \tag{62}$$
$$= \frac{1}{2\gamma}\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] - \frac{\eta\tau}{2}(1-4\eta\tau)\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right]$$
$$+ \eta^2\tau\sigma^2 + \left(\frac{\eta\gamma^2}{2} + 2\eta^2\gamma^2\tau\right)\lambda_1 \tag{63}$$
$$= \frac{1}{2\gamma}\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] - \frac{\eta\tau}{2}(1-4\eta\tau)\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla\psi_t^l(\theta_t)\right\|^2\right]$$
$$+ \eta^2\tau\sigma^2 + \left(\frac{\eta\gamma^2}{2} + 2\eta^2\gamma^2\tau\right)\left(8\eta^2\tau^2(\tau-1)\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla\psi_t^l(\theta_t)\right\|^2\right] + 4\eta^2\tau^2(\tau-1)\sigma^2\right) \tag{65}$$
$$= \frac{1}{2\gamma}\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] - \frac{\eta\tau}{2}\left(1 - 4\eta\tau - 8\gamma^2\gamma^2\tau(\tau-1) - 32\eta^3\gamma^2\tau^2(\tau-1)\right)\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta\right.\right.$$
$$+ \Big(\eta^2\tau + 2\eta^3\gamma^2\tau^2(\tau-1) + 8\eta^4\gamma^3\tau^3(\tau-1)\Big)\sigma^2. \tag{66}$$

Denoting $C' \triangleq 1 - 4\eta\tau - 8\eta^2\gamma^2\tau(\tau-1) - 32\eta^3\gamma^2\tau^2(\tau-1) > 0$ and $A_t \triangleq \eta + 2\eta^3\gamma^2\tau^2(\tau-1) + 8\eta^4\gamma^2\tau^3(\tau-1)$, we get:

$$\mathbb{E}\left[\left\|\sum_{l\in\mathcal{L}_t}\nabla\psi_t^l(\theta_t)\right\|^2\right] \leq \frac{2}{\eta\tau C'}\Big[\mathbb{E}[F(\theta_t)] - \mathbb{E}[F(\theta_{t+1})]\Big]$$
$$+ \frac{1}{\eta\tau\gamma C'}\mathbb{E}\left[\left\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\right\|^2\right] + \frac{2}{C'}A_\tau\sigma^2. \tag{67}$$

Using (41), we get:

$$\mathbb{E}\Big[\|\nabla F(\theta_t)\|^2\Big] \leq \frac{4}{\eta\tau C'}\Big[\mathbb{E}[F(\theta_t)] - \mathbb{E}[F(\theta_{t+1})]\Big] + \Big(\frac{1}{\eta\gamma\tau C'} + 2\Big)\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] + \frac{4A_\tau}{C'}\sigma^2. \tag{68}$$

Summing both sides over $t = 0, 1, \ldots, T-1$ and dividing by $T$, we get $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\theta_t)\|^2\Big]$:

$$\leq \frac{2}{\eta\tau C'T}\Big[\mathbb{E}[F(\theta_0)] - \mathbb{E}[F(\theta_T)]\Big] + \frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1}{\eta\gamma C'\tau} + 2\Big)\mathbb{E}\left[\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l\psi_t^l(\theta_t)\|^2\right] + \frac{4A_\tau}{C'}\sigma^2 \tag{69}$$
$$\leq \frac{2}{\eta\tau C'T}\Big[F(\theta_0) - F(\theta^*)\Big] + \frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1}{\eta\gamma C'} + 2\Big)(\|\nabla F(\theta_t) - \sum_{l\in\mathcal{L}_t}\nabla_l F(\theta_t)\|^2 + \|\sum_{l\in\mathcal{L}_t}\nabla_l F(\theta_t) - \nabla_l\psi_t^l(\theta_t)\|^2) + \frac{4A_\tau}{C'}\sigma^2. \tag{70}$$

This concludes our proof of generalized version of the theorem.

# C. Experimental Setup and Implementation Details

## C.1. Proposed dataset: Ultra-MedVQA (Task 3)

In this section, we detail the process of constructing our Ultra-MedVQA dataset. To maximize the utilization of real medical images, we compile a large-scale medical classification dataset and generate question-answer pairs based on the inherent attributes of the data using the ChatGPT API. Broadly, the construction process involves four main steps:

1. **Preparation of Original Dataset:** To create a comprehensive VQA benchmark, we gathered 10 diverse medical classification datasets covering 9 distinct imaging modalities consisting a total of 707,962 samples [4, 5, 9, 15, 30, 33, 38, 58, 61–64]: Chest X-Ray (117,976 samples), Retinal Optical CT (109,309 samples), Colon Pathology (107,180 samples), Dermatoscope (10,015 samples), Fundus Camera (1,600 samples), Ultrasound (780 samples), Blood cell Microscope (17,092 samples), Kidney cortex Microscope (236,386 samples), Abdominal CT (107,624 samples). It represents 12 different human anatomical regions: Colon, Lung, Skin, Eye, Breast, Kidney, Blood, Femur, Heart, Liver, Pancreas, and Spleen. We use different modality specific datasets as individual clients. Accordingly, we have 9 clients as shown in the Fig. 8 of the main paper. We split the data into training (80%) and testing datasets (20%) in each client.
2. **Question-Answer Template Design:**
   To transform the collected datasets into a question-answer (QA) format, we convert the original classification attributes into QA pairs. This process begins by constructing QA templates for each dataset. On one hand, category information naturally lends itself to QA pair construction. For instance, for the Chest X-Ray dataset, which contains 14 disease categories, we design a QA template like: "Q: What is the specific diagnosis for the lung in this image?; A: Pneumothorax."
   On the other hand, by further analyzing the dataset, we create QA pairs based on additional attributes such as imaging modality and anatomical region. For example, in the Colon Pathology dataset, questions like "What is the modality of the image?" or "What is the abnormal tissue/anatomy in the picture?" are crafted to evaluate modality recognition and tissue/anatomy localization.
   In summary, all QA pairs fall into six distinct question types: Modality Recognition ($\approx$ 10%), Anatomy Identification ($\approx$ 20%), Disease Diagnosis ($\approx$ 39%), Disease Grading ($\approx$ 1%), Tissue Identification ($\approx$ 20%), and Other Biological Attributes ($\approx$ 10%).
3. **Question-Answer Refinement:** To enhance the diversity of our dataset, we utilize ChatGPT-4o to rephrase the questions in each item, altering their expression style and syntactic structure while retaining their original semantic meaning.
4. **Manual Double Checking:** To maintain data quality, we performed additional inspections to ensure the accuracy and reliability of our Ultra-MedVQA dataset.

## C.2. Other Datasets

### C.2.1 VQA Task 1

In this FL scenario, we include five MedVQA clients as shown in Tab. 1 and Fig. 1, with number of samples per class in the clients ranging from 3.90 to 48.03. Here, each client includes one MedVQA dataset that combines different imaging modalities such as CT, MRI, X-Ray, etc.

**(a) SLAKE:** SLAKE [35] combines semantic labels with a structured medical knowledge base. The images are sourced from three open-source datasets [31, 49, 60], and annotated by experienced physicians. To gather questions, experienced doctors either selected predefined questions or rewrote them, ensuring a balanced representation across question types. In this work, we utilize the English subset of the dataset, comprising 642 images and 7033 question–answer pairs. The answers are categorized into 209 classes using GPT-4 [41] and then manually revised to ensure correctness and consistency. We utilize the original partitioning with 4919 samples for training, 1061 for validation, and 1053 for testing.

**(b) VQA-RAD:** VQA-RAD [34] is a radiology-specific dataset introduced in 2018. It features a balanced collection of images from MedPix[1], covering head, chest, and abdomen. The images were provided to clinicians and asked to generate both free-form and template-based questions. Our dataset version consists of 315 images and 2248 question–answer pairs. We categorize the answers into 461 classes using the aforementioned procedure and utilize the partitioning with 1799 samples for training and 449 for testing.

**(c) VQA-Med-2019:** VQA-Med-2019 [7], the second edition of VQA-Med, was introduced during ImageCLEF 2019 challenge. Drawing inspiration from VQA-RAD, VQA-Med-2019 addressed four prevalent question categories: modality, plane, organ system, and abnormality. We categorized the answers into 308 classes following the same process. The total number of image samples was 4200 while the number of QA pairs was 15292. We utilize the partitioning with 14792 samples for training and 500 for testing.

**(d) VQA-Med-2020:** VQA-Med-2020 [3], the third edition of VQA-Med, was released as part of the ImageCLEF 2020 challenge. The images were also collected from MedPix dataset which comprised 36 imaging modalities, 16 planes and 10 organ systems. The QA pairs were generated using previously established patterns. The questions in

[1] medpix.nlm.nih.gov

Table 1. Overview of VQA Datasets

| Dataset | # Images | # QA | Source of images and content | # Classes | Question Category |
|---|---|---|---|---|---|
| **Task 1** | | | | | |
| SLAKE | 642 | 7033 Train:4919 Val:1061 Test:1053 | Medical Segmentation Decathlon, NIH Chest X-ray, CHAOS (Chest X-rays/CTs, Abdomen CTs/MRIs, Head CTs/MRIs, Neck CTs, Pelvic cavity CTs) | 209 | Anatomy, Position, Knowledge Graph, Abnormality, Modality, Plane, Quality, Color, Size, Shape |
| VQA-RAD | 315 | 2248 Train:1799 Test:449 | MedPix (Head axial single-slice CTs or MRIs, Chest X-rays, Abdominal axial CTs) | 461 | Modality, Plane, Anatomy, Abnormality, Object/Condition, Positional Reasoning, Color, Size, Attribute, Other, Counting, Other |
| VQA-Med 2019 | 4200 | 15292 Train:14792 Test:500 | MedPix database (36 modalities, 16 planes, and 10 organ systems) | 308 | Modality, Plane, Anatomy, Abnormality |
| VQA-Med 2020 | 1000 | 1000 Train:800 Test:200 | MedPix database | 187 | Abnormality |
| VQA-Med 2021 | 1000 | 1000 Train:800 Test:200 | MedPix database | 133 | Abnormality |
| **Task 2** | | | | | |
| CT Modality | 978 | 1980 Train:1584 Test:396 | Chest CT Scan [55], Covid CT [56], and SARS-CoV-2 CT-scan [50] | '16 | Modality, Anatomy, Abnormality |
| US Modality | 10855 | 10991 Train:8793 Test:2198 | RadImageNet [39] | 16 | Modality, Anatomy, Abnormality |
| OCT Modality | 3791 | 4646 Train:3717 Test:929 | OCT & X-Ray 2017 [32] (where we consider only OCT images) and Retinal OCT-C8 [2] | 19 | Modality, Anatomy, Abnormality |
| Fundus Modality | 4986 | 5311 Train:4249 Test:1062 | 8 fundus datasets: ACRIMA [18], DeepDRiD [37], Diabetic Retinopathy [57], DRIMDB [47], JSIEC [10], OLIVES [44], PALM2019 [19], Yangxi [36] | 58 | Modality, Anatomy, Abnormality |
| Microscopy Modality | 2969 | 3399 Train:2719 Test:680 | 5 datasets: BioMediTech [40], Blood Cell [1], HuSHeM [48], ALL Challenge [21], and MHSMA [27] | 27 | Modality, Anatomy, Abnormality |
| Histopathology Modality | 2012 | 2281 Train:1825 Test:456 | 4 datasets: BreakHis [51], NLM-Malaria Data [54], CRC100k [29], and MAlig Lymph [42] | 22 | Modality, Anatomy, Abnormality |
| Dermatoscopy Modality | 5897 | 6679 Train:5343 Test:1336 | 7 different skin datasets: Fitzpatrick [20], ISBI2016 [22], ISIC2018 [14], ISIC2019 [17], ISIC2020 [46], Monkeypox Skin Image [25], and PAD-UFES-20 [43] | 36 | Modality, Anatomy, Abnormality |
| X-Ray Modality | 5752 | 7245 Train:5796 Test:1449 | 11 X-Ray datasets: Knee Osteoarthritis [12], RUS CHN [2], Pulmonary Chest Shenzhen [26], Chest X-Ray PA [6], CoronaHack [16], Covid-19 tianchi [53], Covid19 heywhale [13], COVIDx CXR-4 [59], MIAS [52], Mura [45], and Pulmonary Chest MC [26] | 41 | Modality, Anatomy, Abnormality |

this dataset specifically addressed abnormalities. We categorized the answers into 187 classes. The total number of image and QA pair samples in the publicly available validation and test sets amounted to 1000. We divided these into training and test samples using an 80:20 ratio. The number of samples per class is very low thereby making the task highly challenging.

**(e) VQA-Med-2021:** VQA-Med-2021 [8] was introduced during the ImageCLEF 2021 challenge, following the same foundational principles as VQA-Med-2020. The validation and test sets were publicly available, newly curated, and reviewed by medical professionals. We categorized the ab-

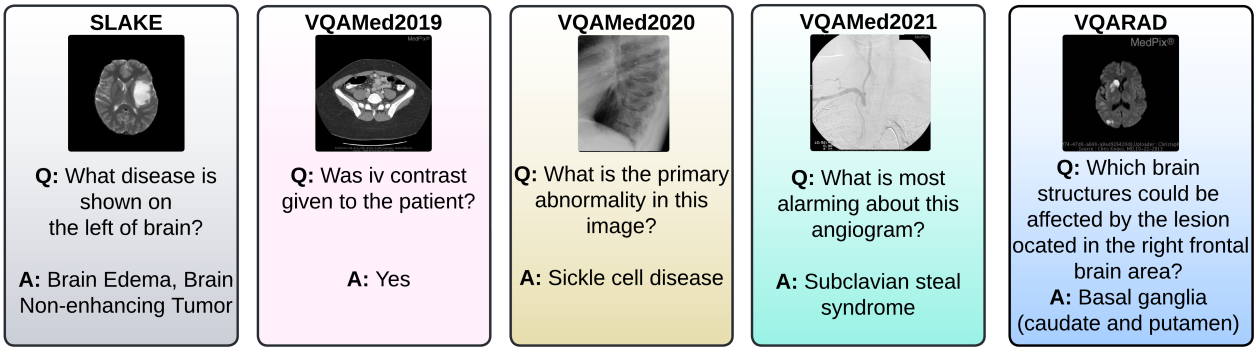


Figure 1. Sample VQA triplets from different clients in Task 1

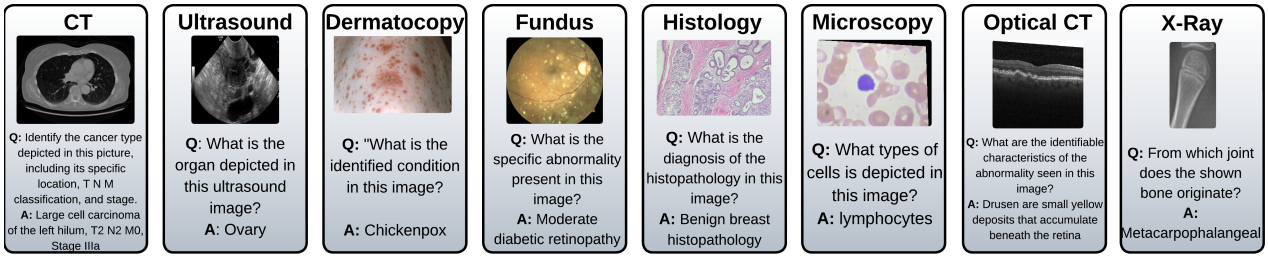


Figure 2. Sample VQA triplets from different clients in Task 2

normalities into 133 classes. Like VQA-Med-2020, the total number of samples, combining all datasets, amounted to 1000 which were divided into training and test samples using an 80:20 split. The limited number of samples per class in this dataset significantly increases the difficulty of the task.

### C.2.2 VQA Task 2

In this scenario, we create eight modality-specific medical imaging clients as shown in Tab. 1 and Fig. 2. The modalities are: CT, Ultrasound, Dermatoscopy, fundus, histology, microscopy, optical CT, and X-Ray. For each client, we combine multiple medical imaging datasets related to the same modality but varying in terms of anatomical regions and abnormalities. We design this setup to mimic real-world settings where different medical clinics might possess different modalities based on the types of medical tests and scans.

**(a) Client 1 (CT):** This client includes 3 CT datasets: Chest CT Scan [55], Covid CT [56], and SARS-CoV-2 CT-scan [50]. There are a total of **16 possible answers** in the answer pool (separated by comma here): Stage Ib, Squamous cell carcinoma of the left hilum T1 N2 M0 Stage IIIa, Large cell carcinoma of the left hilum, Adenocarcinoma of the left lower lobe T2 N0 M0 Stage Ib, COVID-19 infection, Yes, No, Stage IIIa, Chest region, Lungs, CT, Large cell carcinoma of the left hilum T2 N2 M0 Stage IIIa, Stage IIIa of Squamous cell carcinoma of the left hilum, Adenocarcinoma of the left lower lobe, Chest, Squamous cell carcinoma of the left hilum.

**(b) Client 2 (US):** It includes Ultrasound images from RadImageNet [39]. The answer pool has **16 different answers** (separated by comma here) : portal vein, gallbladder, bladder, uterus, thyroid nodule, thyroid, common bile duct, pancreas, liver, ovary, kidney, Ultrasound, spleen, inferior vena cava, aorta, fibroid.

**(c) Client 3 (OCT):** This includes 2 optical CT datasets: OCT & X-Ray 2017 [32] (where we consider only OCT images) and Retinal OCT-C8 [2]. There are **19 possible answers** in the answer list (separated by comma here): The image displays swelling and fluid accumulation in the macula due to Diabetic Macular Edema (DME), The image shows signs of damage to the blood vessels in the retina caused by diabetes, Drusen are small yellow deposits that accumulate beneath the retina, Optical Coherence Tomography (OCT), No, Yes, Macular Hole (MH), Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD) causes progressive damage to the macula leading to vision loss in the center of the visual field, The condition is characterized by the accumulation of fluid in the central retina, This is a normal oct image, There is a small hole in the macula which is the central part of the retina, Age-related Macular Degeneration (AMD), Central Serous Retinopathy (CSR), Drusen, Diabetic Macular Edema (DME), No abnormality detected (normal), The

choroidal neovascularization appears as abnormal blood vessels growing beneath the retina, Choroidal Neovascularization (CNV).

**(d) Client 4 (Fundus images):** This client includes 8 fundus datasets: ACRIMA [18], DeepDRiD [37], Diabetic Retinopathy [57], DRIMDB [47], JSIEC [10], OLIVES [44], PALM2019 [19], Yangxi [36]. There are a total of **58 answer categories** in the answer pool (separated by comma here): Disc swelling and elevation, Diabetic retinopathy level 2, Transverse eye axis, it's a outlier retinal image, Severe hypertensive retinopathy, Vessel tortuosity, Preretinal hemorrhage, VKH disease, Branch retinal vein occlusion (BRVO), Severe diabetic retinopathy, Blur fundus without proliferative diabetic retinopathy (PDR), right eye, It's normal: glaucoma negative, it's a good retinal image, Blur fundus with suspected PDR, Macular hole, Fundus imaging, Proliferative diabetic retinopathy, Silicon oil in the eye, Retinal fundus imaging, Congenital disc abnormality, Myelinated nerve fiber, In this image there are no apparent abnormalities. It represents a normal or fundus of high myopia, Tessellated fundus, Fibrosis, left eye, pathologic myopia, The imaging modality used for this image is fundus photography, it's a bad retinal image, No diabetic retinopathy, fundus photography, Mild diabetic retinopathy, Retinal photography, Massive hard exudates, Cotton-wool spots, Central retinal vein occlusion (CRVO), Rhegmatogenous retinal detachment, Epiretinal membrane, Vitreous particles, Retinitis pigmentosa, Vertical eye axis, Laser spots, Maculopathy, Bietti crystalline dystrophy, Fundus neoplasm, Yellow-white spots-flecks, Normal, Moderate diabetic retinopathy, Large optic cup, Glaucoma positive, Chorioretinal atrophy-coloboma, Central serous chorioretinopathy (CSC), Retinal artery occlusion, Pathological myopia, Peripheral retinal degeneration and break, Dragged disc, Color fundus photography, Diabetic retinopathy level 3.

**(e) Client 5 (Microscopy):** This client includes 5 microscopy datasets: BioMediTech [40], Blood Cell [1], HuSHeM [48], ALL Challenge [21], and MHSMA [27]. There are **27 answer classes** in the answer pool (separated by comma here): The head appears normal, neutrophils, microscopy, No the tail appears to be normal, Microscopy, Yes, the tail appears to be abnormal, monocytes, Epithelioid cells, No, the vacuole appears to be normal, Amorphous, Pyriform, Yes, the vacuole appears to be abnormal, Cobblestone cells, Fusiform cells, Retinal pigmented epithelium (RPE), Hematologic Malignancies, Acute lymphoblastic leukemia, It is abnormal, Phase-contrast microscopy, lymphocytes, Mixed cells of several classes (Fusiform, Epithelioid, Cobblestone), eosinophils, Sperm, The head appears abnormal, Tapered, No, the acrosome appears to be normal, Normal.

**(f) Client 6 (Histopathology):** It includes 4 histopathological datasets: BreakHis [51], NLM-Malaria Data [54], CRC100k [29], and MAlig Lymph [42]. There are **22 possible answers** in the answer pool (separated by comma here): Adipose tissue, Follicular Lymphoma, Histopathology, No, Yes, Mucus, Normal colonic mucosa, Cancer cells, Malignant breast histopathology, Hematoxylin & eosin (H&E) stained histological image, Colorectal adenocarcinoma epithelium, Cancer-associated stroma, Chronic Lymphocytic Leukemia, Benign breast histopathology, histopathology, Smooth muscle, Lymphocyte, Malaria infection, Background of histological image, Debris, Microscopy, Mantle Cell Lymphoma.

**(g) Client 7 (Dermatoscopy):** It includes 7 different skin datasets: Fitzpatrick [20], ISBI2016 [22], ISIC2018 [14], ISIC2019 [17], ISIC2020 [46], Monkeypox Skin Image [25], and PAD-UFES-20 [43]. There are a total of **36 possible answers** in the answer pool (separated by comma here): Malignant melanoma, Malignant epidermal, Skin, Actinic Keratosis, Benign melanocyte, Genodermatoses, Monkeypox, Vascular lesion, Squamous cell carcinoma, Dermoscopy, Malignant cutaneous lymphoma, Actinic keratosis, Nevus, Dermoscopic imaging, Inflammatory Benign keratosis, Yes, Seborrheic Keratosis, Benign epidermal, Malignant dermal, Benign condition, Malignant, Basal Cell Carcinoma, Melanocytic nevus, Benign dermal, Dermatofibroma, Cowpox, Melanoma, Measles, Smallpox, Chickenpox, No, Benign image, Squamous Cell Carcinoma, Malignant condition, Basal cell carcinoma.

**(h) Client 8 (X-Ray):** It includes 11 X-Ray datasets: Knee Osteoarthritis [12], RUS CHN [2], Pulmonary Chest Shenzhen [26], Chest X-Ray PA [6], CoronaHack [16], Covid-19 tianchi [53], Covid19 heywhale [13], COVIDx CXR-4 [59], MIAS [52], Mura [45], and Pulmonary Chest MC [26]. In total, there are **41 possible answers** in the answer pool: Spiculated masses, Radius, Architectural distortion, First Metacarpophalangeal, Abnormal lung, COVID-19 positive, Abnormality present, No abnormality detected, manifestation of tuberculosis, Calcification, Lungs, The lungs appear healthy and normal, It's NORMAL, Lung, Proximal Interphalangeal, Viral Pneumonia, COVID-19 pneumonia, Mammography, Ulna, Pneumonia, Breast tissue, Well-defined/circumscribed masses, Middle Interphalangeal, No it's normal, COVID-19, No the image appears normal, Viral pneumonia, No, First Distal Interphalangeal, Musculoskeletal system, COVID: Lungs will be affected, Metacarpophalangeal, First Proximal Interphalangeal, The diagnosis is normal lung, X-ray, COVID-19 negative, Other: ill-defined masses, Chest X-ray, Asymmetry, Distal Interphalangeal, Chest.

### C.2.3 Datasets for Disease Classification Tasks 4 and 5

We use two different datasets in this study, *viz.*, MIMIC-CXR and Open-I. MIMIC-CXR is a comprehensive dataset comprising 227,835 imaging studies conducted on 65,379 patients who visited the Beth Israel Deaconess Medical Center Emergency Department from 2011 to 2016. The dataset includes a total of 377,110 images, with most studies typically containing both frontal and lateral views. Only frontal views have been utilized in this work. Additionally, the dataset provides semi-structured free-text radiology reports written by practicing radiologists at the time of routine clinical care.

The Open-I dataset, also known as Indiana University (IU) X-ray dataset, contains 7,466 images out of which 3,851 are paired with diagnostic radiology reports. We selected a total of 3,547 frontal view image-report pairs from this dataset.

The class distribution of the datasets is shown in Figs. 3 and 4. As evident from the figures, the class distribution of the datasets is widely different from each other. MIMIC CXR only shows mild imbalance whereas Open-I shows severe imbalance. Fig. 5 shows 24 sample Chest X-Ray images from the MIMIC CXR dataset. As evident from the figure, the dataset exhibits significant variability in terms of image quality, positioning, and patient characteristics. This variability makes it challenging to develop a robust and generalizable model that can handle diverse imaging conditions. Besides, the available disease labels are slightly noisy as they are extracted based on a natural language processing tool called Chexpert labeler from the text radiology reports. Fig. 6 shows the sample reports of MIMIC CXR that consist of a number of sections each - examination, indication, comparison, findings, and impression. Only the "Findings" section of the report were used in this study.

24 sample Chest X-Ray images from the Open-I dataset have been shown in Fig. 7. As evident, the images are remarkably different from that of the MIMIC dataset. While MIMIC CXR is primarily derived from a clinical database of intensive care unit (ICU) patients, Open-I includes images from different clinical contexts, not necessarily limited to ICU patients. However, the number of samples in the dataset is limited which makes it harder to train deeper models on this dataset without overfitting. Table 2 shows the findings section of 17 randomly chosen sample reports along with their labels. As observed from the table, the length of the reports can vary widely depending on the patient case and radiologist and can correspond to one or more disease categories.

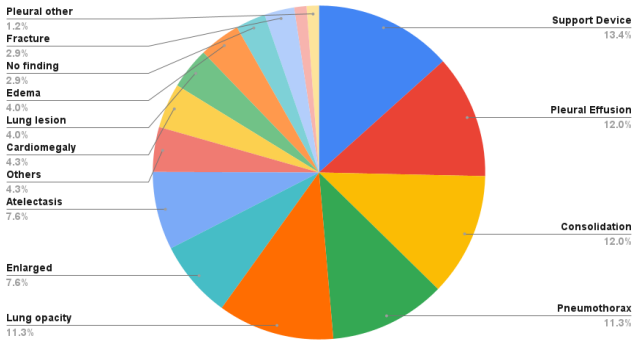


Figure 3. Class proportions (in terms of percentage) in MIMIC Chest X-Ray dataset

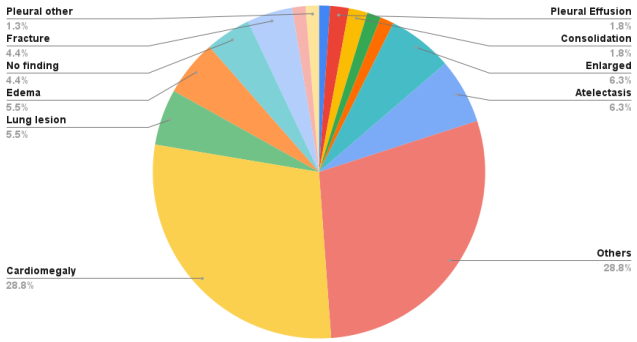


Figure 4. Class proportions (in terms of percentage) in Open-I Chest X-Ray dataset

## C.3. Training and Implementation Details

We fix the initial learning rate $\eta = 0.0001$. We use batch size $B = 16$ for ALBEF and ViLT whereas $B = 4$ for LLaVA-1.5-7b and BLIP-2-7b. We use the AdamW optimizer and a learning rate scheduler with linear decay following [11]. We also use a weight decay of 0.01 with a total of 30 communication rounds for federated fine-tuning including $10\%$ warmup rounds [11]. Each client has task-specific linear classification heads. Each experiment is conducted for three runs and the average value is reported.

For genetic algorithm, the population size is set as 50, number of generations as 20, mutation rate as 0.5. For Particle Swarm Optimization, the population size is kept 50, the number of iterations is kept 20, inertia weight is 0.5, cognitive constant is 1.5, social constant is 1.5. For simulated annealing, initial temperature is 100, final temperature is 1, cooling rate is 0.95, and number of iterations is 20. In ant/bee colony optimization, number of ants/bees is kept 50, number of iterations is 20, pheromone evaporation coefficient is 0.1, pheromone deposit constant is 1.0, initial pheromone is 0.1, influence of pheromone ($\alpha$) = 1.0 and influence of importance scores ($\beta$) = 2.0. All these hyperparameters have been tuned based on comprehensive grid search.

Table 2. The findings of sample reports from Open-I dataset along with the corresponding labels

| Reports | Labels |
|---|---|
| Stable appearance of the right aortic XXXX. Normal heart size. No pneumothorax, pleural effusion or suspicious focal airspace opacity. | No Finding |
| The heart, pulmonary XXXX and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There is a region of left upper lobe perihilar opacity identified. | Lung Opacity |
| The cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size and contour. There is a XXXX-A-XXXX terminating at the caval atrial junction, without evidence of pneumothorax. There is no focal airspace disease. There are small calcified nodules in the superior segment of the right lower lobe, XXXX old granulomatous infection. There are no acute bony findings | Pneumonia |
| No focal consolidation, pneumothorax, or pleural effusions. Stable calcified granulomas. Cardiomediastinal silhouette demonstrates mild tortuosity of the thoracic aorta and atherosclerotic calcifications of the aortic XXXX. No acute osseous abnormality identified. | Enlarged Cardiomediastinum |
| In the interval, consolidation and atelectasis have developed in the right lower lobe. Costophrenic XXXX blunted on the right. Left lung clear. Heart size normal. | Consolidation, Atelectasis |
| Chest Comparison: There is a 2.6 cm diameter masslike density over the lingula partial obscuration left cardiac XXXX. There may be some ill-defined opacity in the right mid and lower lung zone. No pleural effusion is seen. The heart is borderline enlarged. The aorta is dilated and tortuous. Arthritic changes of the spine are present. Pelvis and left hip There is an impacted and rotated fracture through the neck of the femur on the left. No pelvic fracture is seen. Arthritic changes are present in the lower lumbar spine. Large amount of stool and XXXX obscures portions of the pelvis. Femur The femoral images do not XXXX the area of the hip fracture. The remaining portions of the femur appear to be intact with no fracture or destructive process. Extensive atherosclerotic vascular disease throughout the superficial femoral artery is present. Left knee There is osteoporosis and mild arthritic changes. No fracture is seen. No dislocation is identified. Severe atherosclerotic changes of the superficial femoral and popliteal artery are seen. | Cardiomegaly, Lung Lesion |
| No heart size is normal. The lungs are clear. No nodules or masses. Bilateral nipple shadows seen overlying the anterior 6th ribs. Minimal fibrosis in the right apex, may be due to XXXX radiation treatment. | Pleural other |
| Stable postoperative changes with midline sternotomy XXXX and myocardial revascularization. Cardiac size remains mildly enlarged but stable. There is mild vascular congestion. Small bilateral pleural effusions are present, which are XXXX. | Cardiomegaly, Pleural Effusion |
| Prominent hiatal hernia as before. Anticipated senescent changes of mediastinum. Opacity seen XXXX on lateral XXXX XXXX involving both right middle lobe and lingula compatible with some bronchiectasis and chronic inflammatory change. There may be some chronic indolent infection here associated with some chronic consolidation. Perhaps some slight progression, but overall XXXX change since prior examination. On lateral view, the posterior lung bases are grossly clear. No effusions or CHF. | Enlarged Cardiomediastinum, Lung Opacity, Consolidation |
| The lungs are hyperinflated with mildly coarsened interstitial markings consistent with chronic lung disease. No focal consolidation, pneumothorax, or effusion identified. The mediastinal silhouette is stable and within normal limits for size. There is redemonstration without significant change in right hilar calcified lymph XXXX. The bony structures of the thorax demonstrate degenerative changes of the right shoulder and a XXXX right humerus consistent with distal humeral amputation. No acute bony abnormality identified. | Lung Opacity |
| No comparison chest x-XXXX XXXX lungs. Lucency left chest compatible with relatively large pneumothorax and collapse of substantial portion of left lung. No substantial mediastinal shift seen. Right lung grossly clear. | Pneumothorax |
| Stable right-sided subclavian central venous catheter with tip approximating the SVC. Stable right suprahilar opacity, compatible with history of right upper lobe mass. Elevation of the right hemidiaphragm. Right-sided pneumothorax noted measuring approximately 1.8 cm from the the right apex. Stable postsurgical changes left axilla. Degenerative changes thoracic spine. Stable streaky opacities right base. XXXX opacity right midlung, question fluid level, incompletely evaluated, no recent XXXX for comparison. | Lung Opacity, Pneumothorax, Support Devices |
| There is stable, mild enlargement of the cardiac silhouette. Stable mediastinal silhouette. There are low lung volumes with bronchovascular crowding. Scattered XXXX opacities in the right lung base XXXX representing foci of subsegmental atelectasis with scattered airspace opacities in the medial left lower lobe. No pleural effusion. Degenerative changes of the thoracic spine possibly consistent with DISH. | Cardiomegaly, Lung Opacity, Atelectasis, Pneumothorax |
| There is a minimally displaced fracture of the right lateral 7th rib. There is a small right pleural effusion with associated atelectasis of the right lower lobe. There appears to be a healing fracture of the posterolateral right 8th rib. There is questionable cortical defect involving the sternum seen XXXX on lateral view. XXXX would be XXXX to evaluate this finding. As the small right-sided pleural effusion is visible on both PA and lateral views. There is a XXXX left-sided pleural effusion as well. The left lung appears grossly clear. Heart size and pulmonary XXXX appear normal. There is a mild scoliosis involving the thoracic spine. | Atelectasis, Pleural Effusion, Fracture |
| On the right there is marked narrowing of the hip joint space uniformly throughout. Osteophyte formation is present with some sclerosis and subchondral cyst formation vertically along the superior acetabulum and femoral head. I do not see evidence for fracture or destructive process. AP view of the femur shows no femoral XXXX destructive process or other significant abnormality. For of the Left hip shows near-complete obliteration of the joint space with severe subchondral sclerosis and cystic formation in both the superior acetabulum and superior aspect of the femoral head. No fracture or destructive process is identified. Surgical markers were XXXX in the images and left hip for the purpose of surgical planning. PA and lateral chest show the lungs to be clear. There may be some hyperinflation. No pleural effusion is identified. The heart is normal in size. There are calcified mediastinal lymph XXXX. The skeletal structures appear normal. | Support Devices |
| Chest: 2 images. Heart size is normal. Mediastinal contours are maintained. There is a mild pectus excavatum deformity. The lungs are clear of focal infiltrate. There is no evidence for pleural effusion or pneumothorax. No convincing acute bony findings. Right shoulder: 3 images. There has been XXXX and screw fixation of the midshaft right clavicle. The lateral most screw is fractured. This is age-indeterminate as no prior studies are available for comparison. Otherwise, the surgical XXXX appears intact. The humeral head is seen within the glenoid, without evidence for dislocation. No bony fractures are seen. The visualized right ribs appear intact. Right clavicle: 2 images. No clavicle fracture is seen. Once again noted is the surgical fixation XXXX, with fracture of the lateral most fixation screw. | Enlarged Cardiomediastinum, Fracture, Support Devices |
| Chronic bilateral emphysematous changes. The heart size and mediastinal silhouette are within normal limits for contour. The lungs are clear. No pneumothorax or pleural effusions. The XXXX are intact. Stable splenic artery embolism coils. | Support Devices |
| The lungs are clear. Heart and pulmonary XXXX appear normal. The pleural spaces are clear and mediastinal contours are normal. Nodular density overlying the anterior left 4th rib XXXX represents a healing rib fracture. | Lung lesion, Fracture |

# D. Results and Discussions

## D.1. Comparison of Random Vs last few layers fine-tuning

We begin by evaluating the performance of adapters placed in the final layers of the 32-layer LLaVA-1.5-7b vision-

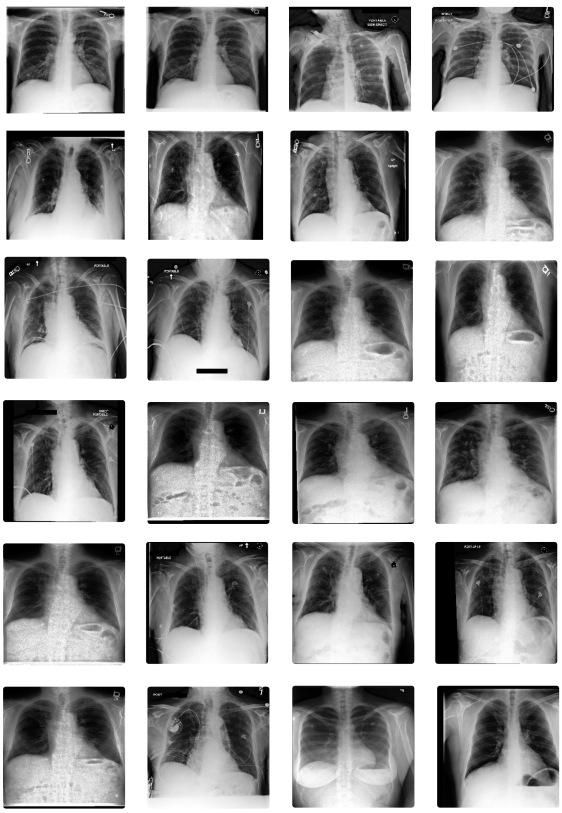

Figure 5. Sample Chest X-Ray images from MIMIC CXR dataset

language model (VLM). The experiments are conducted on the NIH Open-I dataset, addressing the task of multi-label disease detection from Chest X-Ray images and corresponding radiology reports. Our Federated Learning setup comprises four clients created by non-IID partitioning of the dataset using a Dirichlet coefficient $\gamma = 0.5$.

To ensure a comprehensive evaluation, we adopt four distinct adapter configurations: Pfeiffer, Houlsby [24], Compacters [28], and Parallel adapters [23]. Pfeiffer and Houlsby adapters represent traditional architectures within the adapter framework, differing primarily in how they modify intermediate layer representations. Compacters and Parallel adapters, on the other hand, introduce innovative approaches, focusing on parameter efficiency (Compacters) or exploring parallel rather than sequential adapter integration. By selecting a diverse set of adapter configurations, we explore a broad spectrum of design paradigms, from traditional layers (Pfeiffer, Houlsby) to advanced, compact, and efficient alternatives (Compacters, Parallel adapters). This diversity ensures our findings are not overly specific to any one adapter type.

We compare the performance of placing adapters in the last K layers against randomly inserting K adapters throughout LLaVA to determine whether end-layer placement offers a performance advantage. For both approaches, we progressively reduce the number of adapters (K) from 32 to 4 in increments of 4. Results, averaged over three random seeds, are presented in Figures 8–13.

For Pfeiffer (Figures 8–9) and Houlsby adapters (Figure 10), the performance of placing adapters in the last K layers closely matches that of random placement, particularly when a larger number of adapters is used. In contrast, Compacters (Figure 11) and Parallel adapters (Figure 12) show some sensitivity to layer-specific information, especially in the later layers, slightly favoring last-layer placement over random insertion in certain cases.

Overall, as illustrated in Figure 13, our results challenge the initial expectation that end-layer adapter placement would consistently outperform random placement. From a theoretical standpoint, the later layers of a model typically capture task-specific representations, while earlier layers learn general features. Thus, inserting adapters in the last few layers is intuitively expected to yield better task-specific performance. However, our findings suggest that random adapter placement can effectively distribute the task adaptation burden across layers, achieving performance comparable to targeted placement—particularly with Pfeiffer and Houlsby adapters. This observation indicates that the last few layers might not be as critical as hypothesized when compared to random placement.

While the performance differences are generally subtle, occasional fluctuations are observed across clients. For instance, in Compacter and Parallel adapter tuning (Figures 11 and 12), Clients 1 and 3 exhibit minor variations, likely influenced by specific data distributions or task complexities.

In summary, the results indicate that structured end-layer adapter placement may not be the most optimal strategy for adapter integration. This observation motivates our exploration of a more effective mechanism for adapter selection in vision-language models.

## D.2. Explanation of results in Tab.2 of main paper

The table 2 presents the performance comparison of various adapter layer selection strategies on Vision-Language Models (VLMs) across six tasks, measured in terms of accuracy for Tasks 1–3 and F1-score for Tasks 4–6 (as it involves multi-label classification). The experiments are evaluated under two distinct resource allocation settings: **homogeneous resources across clients** and **heterogeneous resources across clients**. The heterogeneous setting simulates **device heterogeneity** by varying the number of trainable layers per client, reflecting practical federated learning scenarios where client resources differ.

**FINAL REPORT**
EXAMINATION: CHEST (PORTABLE AP)

INDICATION: ___ year old woman with CNS lymphoma and AMS. Now with new fever.
// Eval fever

TECHNIQUE: Chest single view

COMPARISON: ___

FINDINGS:

Right Port-A-Cath in place. Elevated right hemidiaphragm, stable. Bibasilar opacities, mildly more prominent on the right, likely atelectasis.
Pneumonitis cannot be excluded in the appropriate clinical setting. There may
be tiny right pleural effusion

IMPRESSION:

More prominent bibasilar opacities, likely atelectasis; pneumonitis cannot be excluded in the appropriate clinical setting, particularly on the right.

**FINAL REPORT**
CHEST RADIOGRAPH

INDICATION: History of Whipple surgery. Abdominal pain.

COMPARISON: ___.

FINDINGS: As compared to the previous radiograph, the lung volumes have minimally decreased, likely as a result of the known widespread fibrotic lung parenchymal process. No newly appeared parenchymal opacities. No pleural effusions. No pulmonary edema. Moderate cardiomegaly with enlargement of the left ventricle and tortuosity of the thoracic aorta.

**FINAL REPORT**
EXAMINATION: CHEST (AP AND LAT)

INDICATION: ___F with cough, mild SOB

COMPARISON: ___.

FINDINGS:

AP upright and lateral views of the chest provided. Underpenetration due to body habitus somewhat limits assessment. There is no focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.

IMPRESSION:

No acute intrathoracic process.

**FINAL REPORT**
EXAMINATION:
Chest: Frontal and lateral views

INDICATION: History: ___F with SOB, s/p fall. hx of SOB // PNA?

TECHNIQUE: Chest: Frontal and Lateral

COMPARISON: ___

FINDINGS:

There are relatively low lung volumes. Mild pulmonary vascular congestion is seen. Right ___- and infrahilar opacity is nonspecific, could relate to prominent pulmonary vasculature, but underlying consolidation due to pneumonia
or aspiration not excluded. The cardiac silhouette is enlarged. There is prominence of the main pulmonary artery which may relate to underlying pulmonary hypertension. No large pleural effusion or pneumothorax is seen.

IMPRESSION:

Relatively low lung volumes. Mild pulmonary vascular congestion. Right ___- and infrahilar opacity is nonspecific, could relate to prominent pulmonary vasculature, but underlying consolidation due to pneumonia or aspiration not excluded.

**FINAL REPORT**
HISTORY: ___-year-old female with fever.

COMPARISON: Prior exam dated ___.

FINDINGS:

PA and lateral views of the chest were provided. There is subsegmental linear atelectasis in the left lower lobe. No definite consolidation effusion or pneumothorax is seen. The heart and mediastinal contours appear normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm. Mild degenerative changes in the mid T-spine.

IMPRESSION:

No definite signs of pneumonia. Left lower lobe linear atelectasis.

**FINAL REPORT**
PA AND LATERAL CHEST OF ___

No prior studies for comparison.

FINDINGS: Heart size, mediastinal and hilar contours are normal. Linear atelectasis is present in the left lower lobe. Additionally, patchy opacities are present in the right mid and lower lung most prominent in the right infrahilar region. There are likely small pleural effusions bilaterally. Drainage catheter seen in the upper abdomen is incompletely imaged on this study.

IMPRESSION:
1. Patchy and linear opacities in the right mid and lower lung are most likely due to atelectasis. If clinical suspicion for pneumonia persists, followup radiograph may be helpful.

2. Probable small bilateral pleural effusions.

**FINAL REPORT**
INDICATION: ___-year-old female with chest pain. Evaluate for pneumonia.

TECHNIQUE: AP and lateral views of the chest.

COMPARISON: Multiple priors with direct comparison made to study from ___

FINDINGS:

The lungs are well inflated and clear. There is persistent widening of the mediastinum due to mediastinal lipomatosis. The cardiomediastinal silhouette
and hilar contours are unchanged. There is no pleural effusion or pneumothorax.

IMPRESSION:

No acute cardiopulmonary process.

**FINAL REPORT**
EXAMINATION: CHEST (PA AND LAT)

INDICATION: History: ___F with cough, orthopnea

TECHNIQUE: Chest PA and lateral

COMPARISON: Chest radiograph ___

FINDINGS:

Heart size remains mildly enlarged. The mediastinal and hilar contours are unchanged. Pulmonary vasculature is not engorged. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.

IMPRESSION:

No acute cardiopulmonary abnormality.

Figure 6. Sample reports from (Multimodal) **CLIENT 1**. In this work, we only use the "FINDINGS" section as radiology report (text modality).

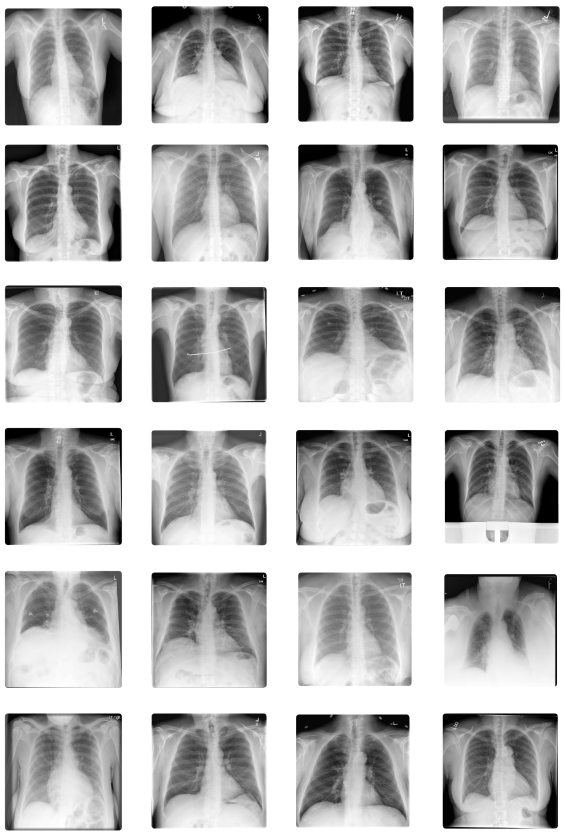

Figure 7. Sample Chest X-Ray images from Open-I dataset. As evident, the images are diverse and contain several artifacts. The images are also notably different from MIMIC-CXR.

#### D.2.1 Homogeneous Setting

In the homogeneous resource scenario, where all clients are allocated the same number of trainable layers ($L = 4$), the $F^3OCUS$ **method consistently outperforms all other strategies** across tasks.

- The mean scores for $F^3OCUS$ (e.g., $72.52$ on Tasks 1–3 and $72.16$ on Tasks 4–6) surpass those of the second-best performing method (LNTK), which achieves mean scores of $69.33$ and $69.15$, respectively. This shows that leveraging multi-objective meta-heuristic strategy, balancing client-specific layer selection with global convergence, leads to superior overall performance. Its ability to adaptively distribute layer selection achieves higher F1-scores in Tasks 4–6, where precision and recall are crucial.
- Other notable methods like FishMask and GradFlow also perform well, with FishMask demonstrating robustness in Tasks 1–3, particularly on ViLT.

#### D.2.2 Heterogeneous Setting

In the heterogeneous setting, performance variations across methods become more apparent due to client-specific resource constraints.

- Device heterogeneity introduces significant challenges, as **only $F^3OCUS$ is able to adapt effectively across all tasks**, achieving the best scores for every model type (ViLT, LlAVA, and BLIP). This highlights the robustness of $F^3OCUS$ in addressing non-uniform resource constraints.
- The detailed heterogeneity settings (e.g., finetuning 6 layers for some clients and 2 layers for others) simulate real-world scenarios where some clients have more computational power or larger datasets than others. Methods that fail to adapt to these constraints (e.g., SNIP, RGN) show lower mean scores, particularly in Tasks 4 and 5.

#### D.2.3 Task-Specific Insights

- **Tasks 1–3 (Accuracy)**: Accuracy improves with more trainable layers in clients with richer computational resources. $F^3OCUS$ achieves the highest accuracy across all three tasks due to its efficient use of client heterogeneity. FishMask and GradFlow exhibit competitive performance, especially for BLIP, suggesting their suitability for moderately heterogeneous setups.
- **Tasks 4–6 (F1-Score)**: These tasks benefit from precise layer selection strategies, as F1-score accounts for both precision and recall. $F^3OCUS$ demonstrates its strength in handling the trade-off between local and global performance by achieving consistently higher F1-scores. LNTK performs well in Tasks 4 and 6 but struggles slightly in Task 5 due to its reliance on principal eigenvalue computations, which might not adapt well to clients with fewer trainable layers.

#### D.2.4 Method-Specific Insights

- $F^3OCUS$: Dominates across all metrics due to its balanced multi-objective optimization, outperforming LNTK by $2.6\%$ in mean accuracy and $3.0\%$ in mean F1-score.
- **LNTK**: Performs strongly in homogeneous settings but slightly lags in heterogeneous setups, highlighting its limitations in dynamic environments.
- **FishMask and GradFlow**: Reliable performers, particularly in homogeneous settings, but lose ground in heterogeneous environments due to their less adaptive layer selection strategies.
- **Magnitude and SNIP**: While lightweight and efficient, these methods fail to leverage client-specific heterogeneity, resulting in lower performance across tasks.

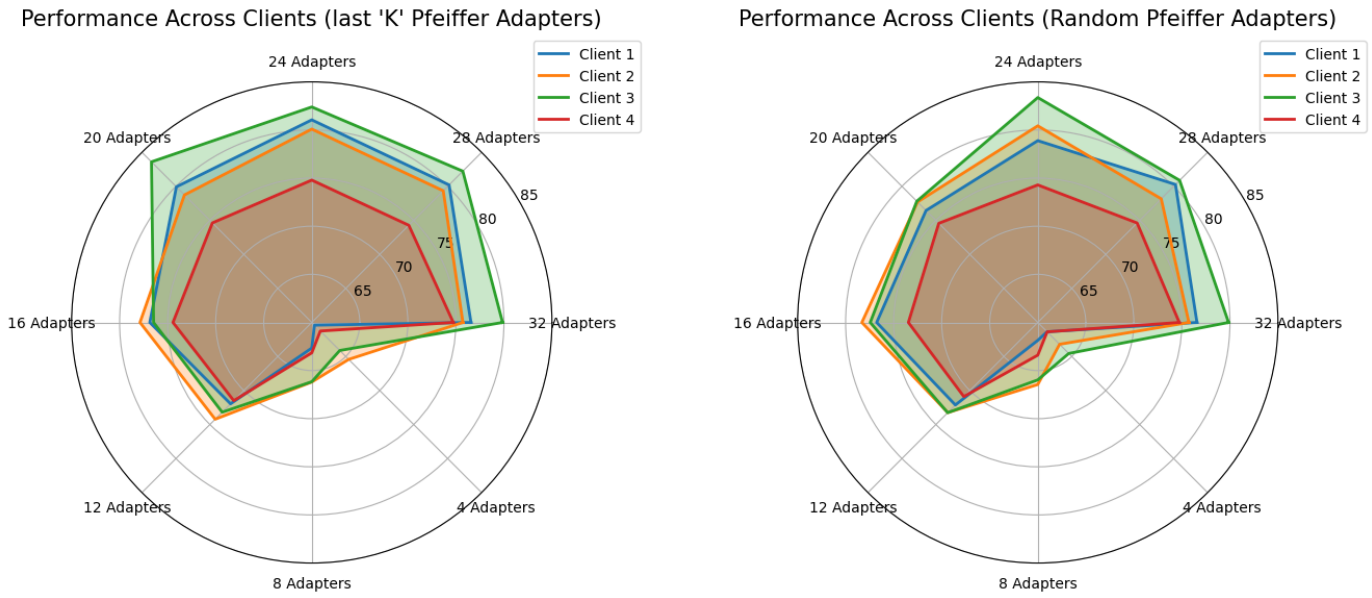


Figure 8. Comparison of parameter-efficient fine-tuning of 32-layered LLaVA-1.5-7b with last 'K' and random 'K' Pfeiffer adapters on Open-I dataset in terms of F1 score.

#### D.2.5 Insights on Device Heterogeneity

The **heterogeneous setting** demonstrates the practical relevance of federated learning in real-world scenarios, where clients operate under varying resource constraints:

- **Tasks with more finetuned layers per client (e.g., Tasks 1 and 6)**: Clients with 6 trainable layers achieve higher accuracy/F1-scores, emphasizing the importance of adapting to client capabilities.
- **Tasks with fewer finetuned layers (e.g., Task 3)**: Methods like $F^3OCUS$ that balance layer allocation across clients maintain superior performance even when some clients operate with limited resources.

### D.3. Scalability Analysis

We increase the number of clients from 10 to 100 in steps of 10 for MIMIC-CXR dataset to demonstrate the scalability of the proposed method. Figure 14 shows that while there is a slight decrease in performance of both ViLT and BLIP-2-7b, overall performance is consistent and stable. In all cases, we subsample 10 clients.

### D.4. Additional experiments on heterogeneous device settings

In the main paper, we systematically design the device heterogeneity in such a way that the average number of layers per client is closer to 4, which is the number of selected layers per client for homogeneous settings, while having inter-client diversity. This is deliberately designed so that we can compare the homogeneous and heterogeneous settings.

Here, we investigate two additional experimental scenarios with varying device heterogeneity where we randomly select the number of clients to avoid any biases in layer selection. In Tab. 3, we report the performance of our algorithm and SOTA methods where we randomly select the number of layers across the clients in each task within 1 and 6. We further increase the range to 1-12 layers and report the performance in Tab. 4. In both the cases, LNTK is observed to outperform the SOTA methods. Additionally, $F^3OCUS$ is observed to improve the performance by around $3\%$ over LNTK which is similar to the performance improvement reported in the Tab. 2 of the main paper and demonstrates the importance of server-level refinement of layer selection.

### D.5. Qualitative analysis

To qualitatively investigate the performance of our proposed method, we plot the t-SNE feature visualizations of each client for Task 2 in Figs. 15 and 16. A closer look into these two plots reveal greater separability achieved by $F^3OCUS$ than the baselines. For the baseline methods, the feature embeddings are scattered and show poor clustering. There is significant overlap between clusters, indicating poor inter-class separability. These methods struggle to create distinct representations for different answers, reflecting suboptimal learning. On the other hand, $F^3OCUS$ demonstrates significantly improved clustering compared to the other two methods. The clusters are tighter, with less overlap between different answers. The separation between different clusters is distinct, showing

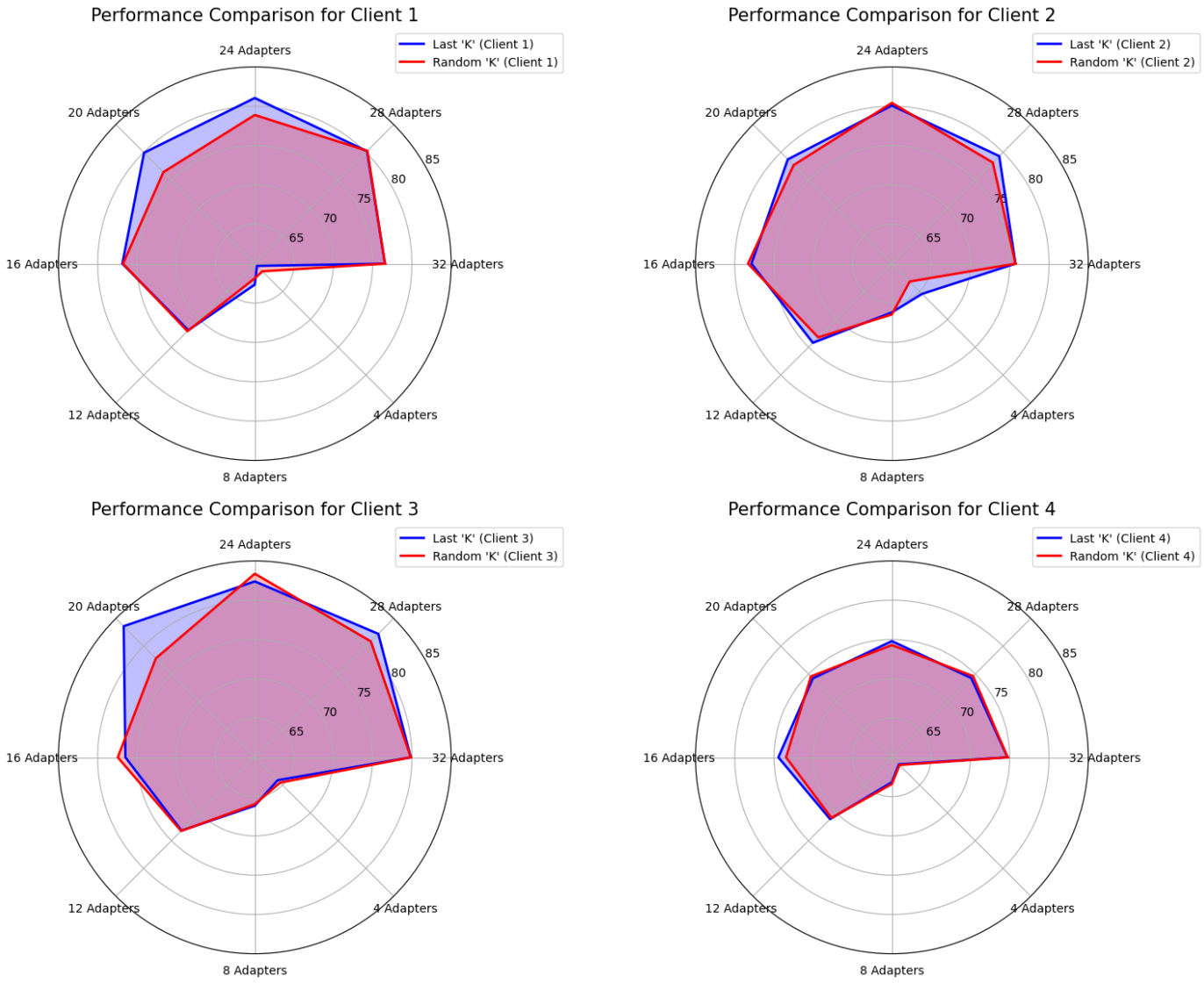


Figure 9. Clientwise Comparison of parameter-efficient fine-tuning of 32-layered LLaVA-1.5-7b with last 'K' and random 'K' Pfeiffer adapters on Open-I dataset in terms of F1 score.

better inter-answer discriminability. This indicates that $F^3OCUS$ effectively learns answer-specific features, using multi-objective meta-heuristics optimization in the federated learning setup. Overall, dermatology and ultrasound clients are observed to be more challenging than the other clients. This is possibly because of the noisy and imbalanced skin cancer dataset (in dermatology client) as well as limited availability of breast ultrasound images (in ultrasound clients).

Furthermore, in order to closely analyze the performance of each SOTA method and compare their feature separability, we further plot the corresponding t-SNE visualizations for a randomly chosen client (Microscopy) in Fig. 17. It shows that the inter-answer discriminability is particularly low in Federated dropout (*i.e.*, random), Last K-layer finetuning, weight magnitude-based selection, SNIP and RGN, whereas discriminability is better in FishMask, GradFlow, GraSP, SynFlow, Fedselect, and SPT. $F^3OCUS$ is observed to be have the highest separability and tighter clusters among all.

## E. Further Discussions and Clarifications

**1. Server-Level Optimization Cost and Scalability**: As shown in Table 5. the meta-heuristic optimization runs efficiently on CPUs with negligible time overhead per round except Genetic Algorithm. The Pareto archive is dynamically updated to retain only non-dominated solutions, re-

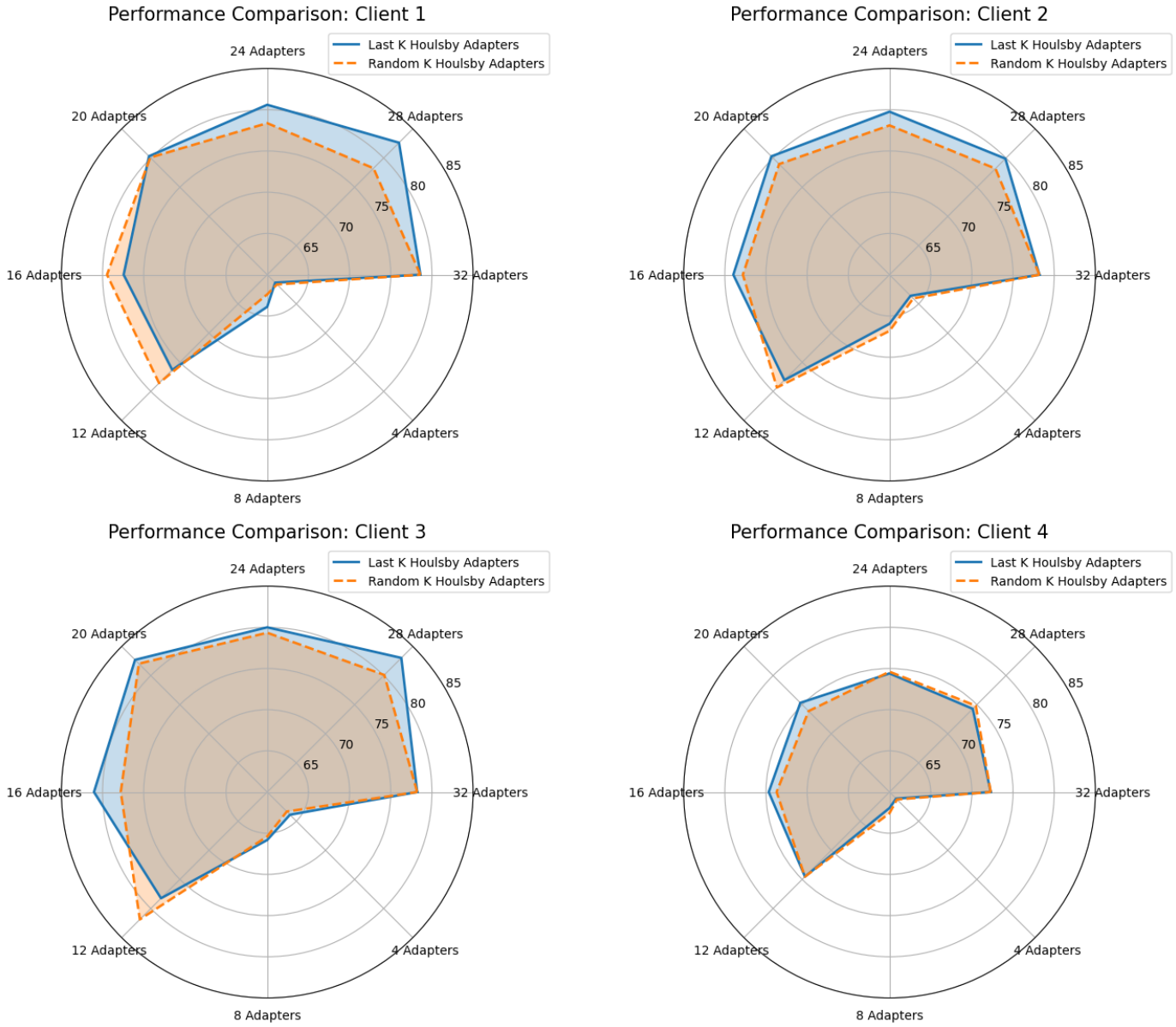


Figure 10. Clientwise Comparison of parameter-efficient fine-tuning of 32-layered LLaVA-1.5-7b with last 'K' and random 'K' Houlsby adapters on Open-I dataset in terms of F1 score.

ducing cost of dominance checks. To ensure scalability while maintaining performance for more clients, we subsample 10 clients/round following previous works.

**2. Dominance of Top Eigenvalue - Empirical evidence**: We provide the intuition behind using principal eigenvalue in the main paper, drawing on spectral bias. Additionally, the layerwise eigenvalue spectrum in Fig. 18 highlights that the principal eigenvalue is 3–6 orders of magnitude larger than the second-largest eigenvalue (Eg: marked in red for 2 layers), showing its spectral dominance.

**3. Most effective meta-heuristic algorithm**: NSGA and MOPSO are particularly suited for our multi-objective optimization. NSGA efficiently explores the Pareto front, ensuring diverse and optimal trade-offs between maximizing importance and promoting layer selection diversity. MOPSO balances exploration and exploitation dynamically, making it scalable for large foundation models. We present the performance of all meta-heuristic algorithms across 6 tasks in Table 6 which shows that MOPSO achieves the best performance closely followed by NSGA. While SA and ABC offer lightweight alternatives, they struggle with high-dimensional search spaces (as reflected in Tab. 6) and

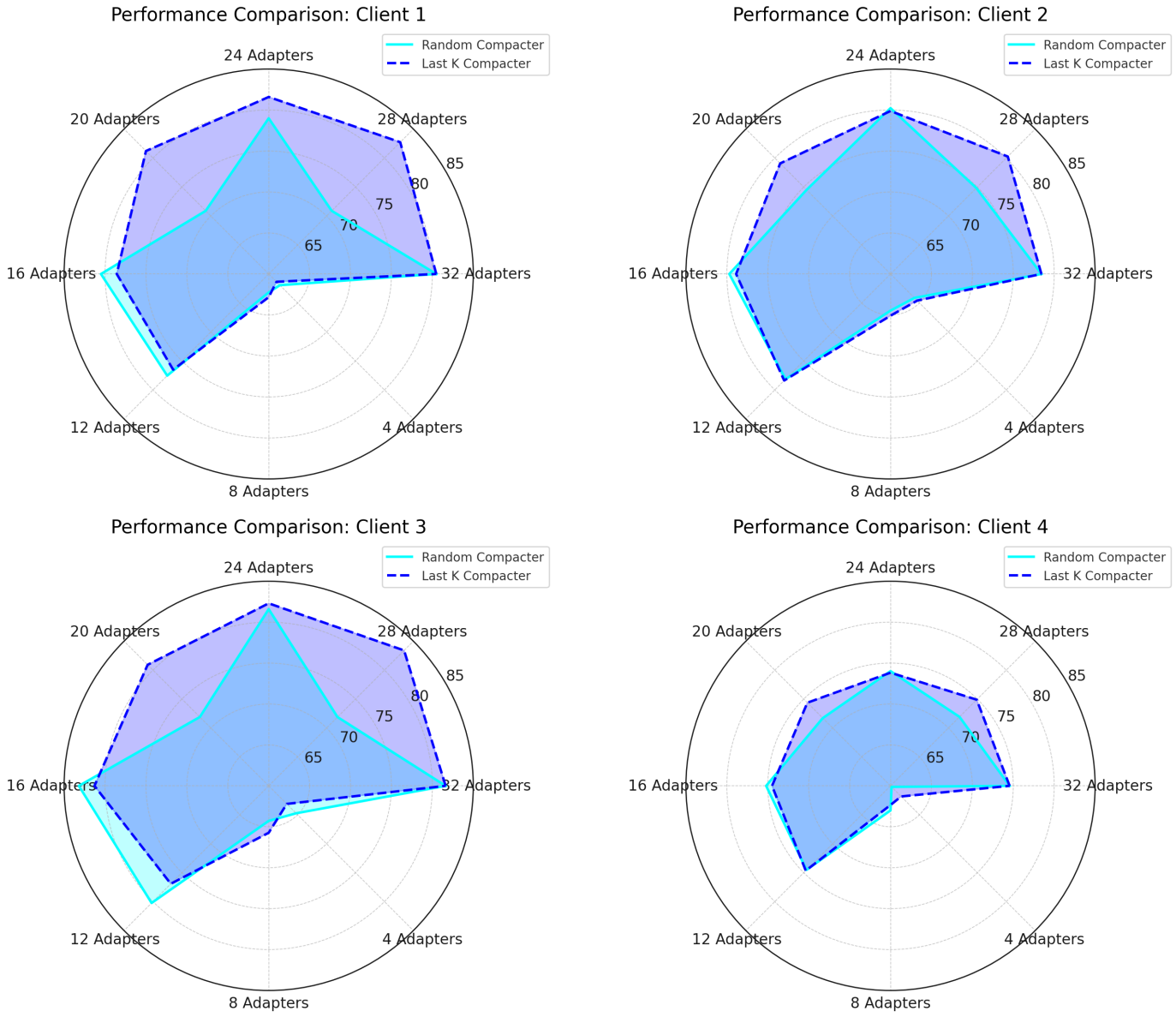


Figure 11. Clientwise Comparison of parameter-efficient fine-tuning of 32-layered LLaVA-1.5-7b with last 'K' and random 'K' Compacters on Open-I dataset in terms of F1 score.

also take more number of iterations to converge in practice.

**4. NTK Eigenvalue decomposition computation**: Our NTK-based method improves standard FedAvg (with random dropout to match # of parameters) by around 5% and 7% with 12-layered ViLT and 32-layered LLaVA (Tab. 2 in main paper) with added time complexity of **0.07s** and **5.64s** per round respectively. For CT client (Task 2) with 4 fine-tunable layers of LLaVA, FedAvg takes $137.53 \pm 1.02s$ while $F^3OCUS$ takes $151.40 \pm 1.25s$ per round. Eigenvalue computation takes only **0.1s** for ViLT and **2.9s** for LLaVA per layer on **single CPU for full decomposition** with *torch.linalg.eigvalsh* whereas only **0.006s** for ViLT and **0.18s** for LLaVA per layer **for only principal Eigenvalue** with *scipy.sparse.linalg.eigsh* (see Table 7).

**5. Adaptive Fine-Tuning for Device Heterogeneity** : Our method adaptively selects layers for fine-tuning based on client device constraints. Given a model architecture and input specifications, it determines the maximum number of tunable layers (K) by analyzing memory requirements with a safety margin using **Algorithm 46**. This client-specific K value then guides layer selection through client (via NTK) and server-side (via meta-heuristic) optimization as men-

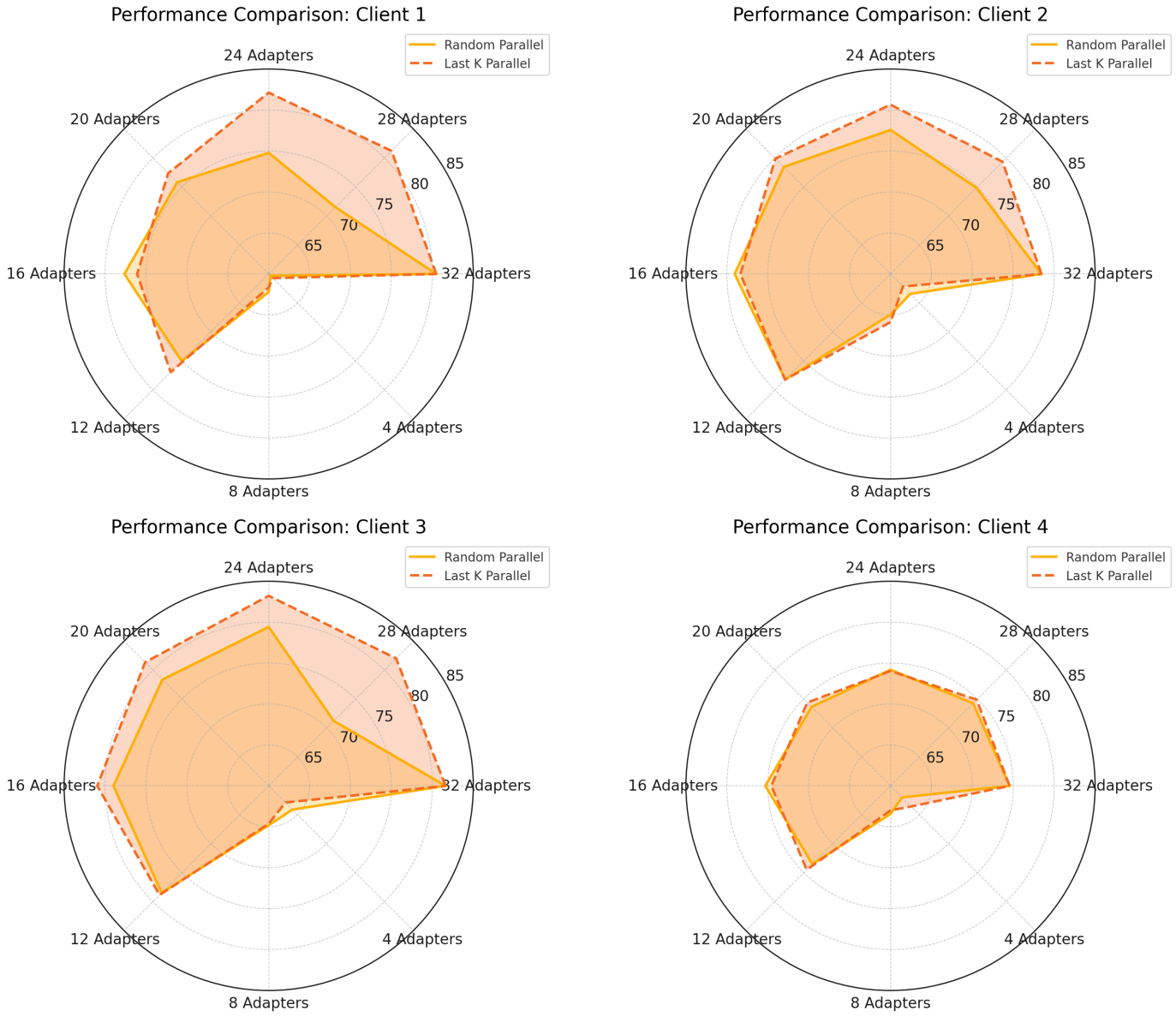


Figure 12. Clientwise Comparison of parameter-efficient fine-tuning of 32-layered LLaVA-1.5-7b with last 'K' and random 'K' Parallel adapters on Open-I dataset in terms of F1 score.

tioned in lines 454-465 for different clients based on our FL set up. *Eg:* If 6 clients have: Tesla V100 (32GB), A100 (40GB), Quadro GV100 (32GB), A6000 (48GB), A40 (45GB), and 2x RTX A4500 (20GBx2), our Algorithm 46 estimates K as 2, 4, 2, 6, 6, 4 layers respectively for fine-tuning LLaVA on CT client based on GPU/image/batch size.

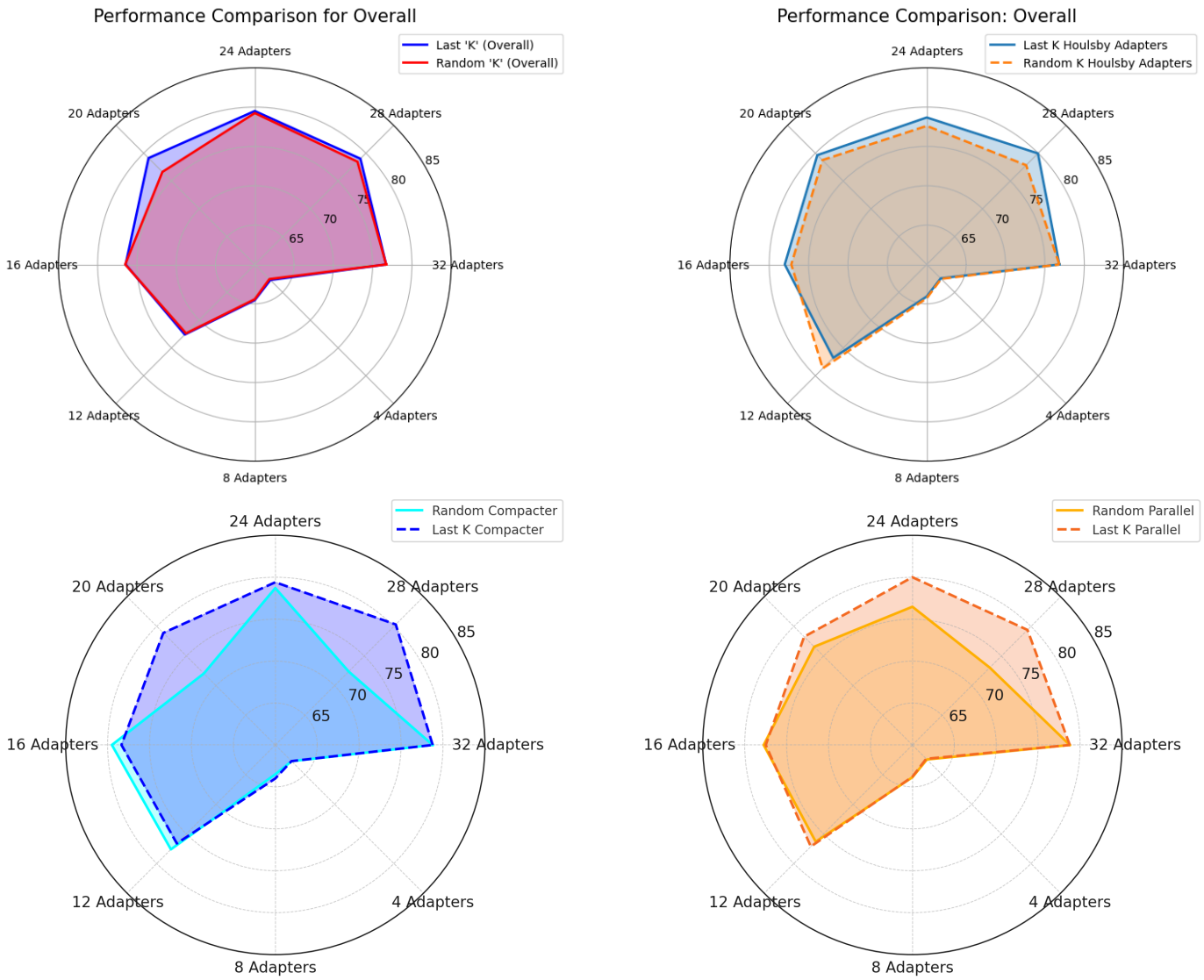


Figure 13. Overall Comparison of different adapter variants for fine-tuning 32-layered LLaVA-1.5-7b with last 'K' and random 'K' adapters on Open-I dataset in terms of F1 score.

**Algorithm 46** Compute Fine-Tunable Layers

**Input:** model, input_size (including batch size), safety_margin
**Output:** Number of fine-tunable layers.
available_memory ← `get_available_gpu_memory()` × (1 - safety_margin)
finetunable_layers ← 0
**for** layers in model() **do**
    layer_memory ← `estimate_memory(layers,input_size)`
    **if** layer_memory > available_memory **then break**
    **end if**
    available_memory ← available_memory - layer_memory
    finetunable_layers ← finetunable_layers + 1
**end for**
**return** finetunable_layers

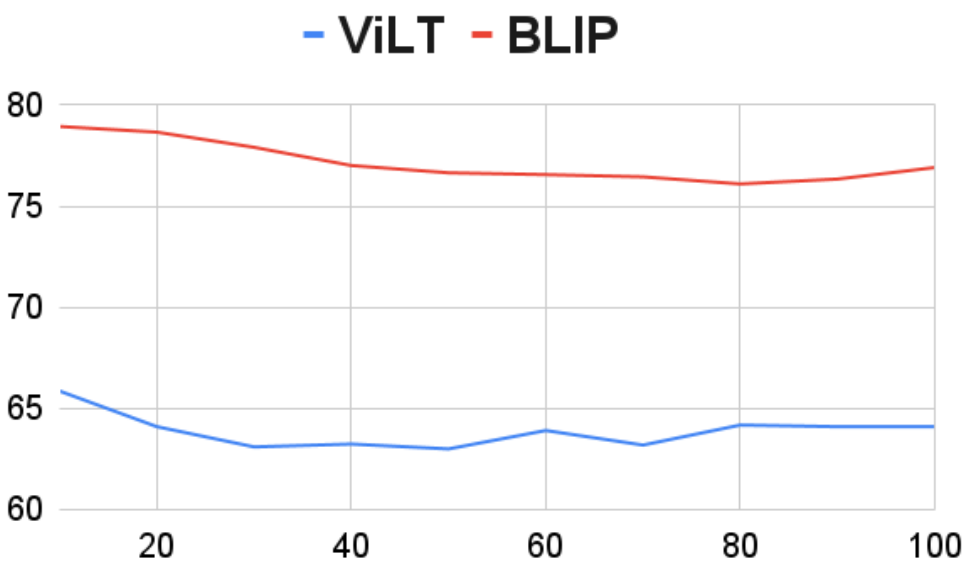


Figure 14. Scalability analysis of $F^3OCUS$ with two different architectures: ViLT and BLIP-2-7b on MIMIC-CXR dataset. We vary the number of clients (on the X-axis) from 10 to 100 in gaps of 10 and show the F1-score (on the Y-axis)

Table 3. Performance Table on VLM layer selection with heterogeneous resources across clients with **randomly chosen number of layers between 1 and 6 in different clients (to ensure unbiased evaluation)**

| **Fine-tuning** | **Task 1** | | **Task 2** | | **Task 3** | | **Task 4** | | **Task 5** | | **Task 6** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ViLT** | **BLIP** | **ViLT** | **BLIP** | **ViLT** | **BLIP** | **ViLT** | **BLIP** | **ViLT** | **BLIP** | **ViLT** | **BLIP** |
| FD | 32.65 | 33.27 | 73.22 | 69.08 | 71.85 | 69.69 | 49.63 | 57.70 | 56.82 | 68.22 | 76.54 | 82.53 |
| Last | 33.38 | 34.30 | 72.46 | 67.97 | 71.69 | 68.90 | 54.59 | 57.31 | 58.66 | 65.73 | 77.72 | 81.52 |
| Magnitude | 31.17 | 30.70 | 70.79 | 69.50 | 70.86 | 71.37 | 52.78 | 58.04 | 57.36 | 66.47 | 77.39 | 82.27 |
| FishMask | 34.94 | 36.19 | 75.18 | 73.17 | 74.07 | 73.85 | 54.30 | 62.21 | 60.78 | 72.69 | 79.35 | 82.54 |
| GradFlow | 34.12 | 36.93 | 75.39 | 72.34 | 74.36 | 74.49 | 54.20 | 61.74 | 61.01 | 71.92 | 80.39 | 81.84 |
| GraSP | 34.73 | 36.22 | 76.22 | 71.69 | 73.48 | 73.85 | 53.29 | 61.44 | 60.77 | 71.16 | 79.59 | 83.15 |
| SNIP | 30.84 | 34.83 | 75.11 | 70.43 | 73.57 | 71.75 | 52.08 | 58.38 | 61.10 | 69.01 | 77.59 | 81.54 |
| RGN | 32.49 | 35.97 | 75.98 | 70.84 | 72.77 | 71.44 | 53.82 | 61.39 | 58.63 | 70.28 | 77.48 | 81.01 |
| Synflow | 34.53 | 37.48 | 76.28 | 71.58 | 74.11 | 73.36 | 54.90 | 61.98 | 60.69 | 72.02 | 78.36 | 82.22 |
| Fedselect | 34.68 | 36.31 | 74.76 | 70.04 | 73.01 | 72.52 | 53.87 | 60.79 | 59.79 | 71.84 | 78.83 | 82.94 |
| SPT | 34.10 | 36.57 | 75.58 | 73.23 | 73.77 | 74.23 | 54.15 | 62.16 | 61.13 | 72.89 | 78.93 | 82.54 |
| LNTK | 36.84 | 39.02 | 78.35 | 75.89 | 76.37 | 76.29 | 56.39 | 65.63 | 63.71 | 74.83 | 83.17 | 85.96 |
| $F^3OCUS$ | **40.23** | **42.12** | **83.67** | **79.63** | **79.02** | **78.22** | **59.58** | **69.03** | **65.87** | **78.18** | **85.84** | **89.12** |

Table 4. Performance Table on VLM layer selection with heterogeneous resources across clients with randomly chosen number of layers between 1 and 12 in different clients (to ensure unbiased evaluation)

| **Fine-tuning** | **Task 1** | | **Task 2** | | **Task 3** | | **Task 4** | | **Task 5** | | **Task 6** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ViLT** | **BLIP** | **ViLT** | **BLIP** | **ViLT** | **BLIP** | **ViLT** | **BLIP** | **ViLT** | **BLIP** | **ViLT** | **BLIP** |
| FD | 33.48 | 34.39 | 75.46 | 70.62 | 73.52 | 70.24 | 51.25 | 57.72 | 58.10 | 68.29 | 80.22 | 84.88 |
| Last | 34.76 | 34.78 | 78.02 | 68.01 | 73.62 | 69.28 | 52.10 | 57.71 | 59.71 | 67.20 | 80.85 | 81.93 |
| Magnitude | 33.10 | 30.77 | 73.32 | 70.63 | 72.22 | 70.43 | 53.02 | 58.24 | 59.14 | 66.97 | 79.53 | 83.34 |
| FishMask | 36.40 | 36.55 | 78.75 | 74.11 | 74.47 | 75.37 | 54.55 | 63.17 | 61.34 | 72.97 | 81.53 | 83.66 |
| GradFlow | 35.81 | 36.84 | 78.38 | 73.60 | 74.92 | 73.48 | 53.94 | 63.24 | 61.52 | 72.02 | 81.69 | 83.45 |
| GraSP | 36.26 | 36.16 | 76.16 | 73.97 | 74.13 | 73.85 | 53.49 | 61.47 | 61.22 | 71.93 | 82.22 | 84.81 |
| SNIP | 31.28 | 35.85 | 75.38 | 73.57 | 75.56 | 72.86 | 53.03 | 59.27 | 62.88 | 70.26 | 80.70 | 83.58 |
| RGN | 33.99 | 35.28 | 77.78 | 75.03 | 74.24 | 72.43 | 53.63 | 63.30 | 61.04 | 73.00 | 78.76 | 82.22 |
| Synflow | 35.30 | 36.60 | 77.66 | 75.59 | 74.33 | 74.99 | 53.58 | 61.80 | 62.13 | 72.92 | 83.02 | 83.60 |
| Fedselect | 36.94 | 36.73 | 79.58 | 74.68 | 74.60 | 72.78 | 53.25 | 62.55 | 60.27 | 72.78 | 79.99 | 84.31 |
| SPT | 33.94 | 36.88 | 77.19 | 74.93 | 75.11 | 73.39 | 54.86 | 62.34 | 61.93 | 72.63 | 81.84 | 84.06 |
| LNTK | 38.47 | 39.02 | 82.55 | 78.71 | 77.27 | 76.85 | 56.82 | 65.78 | 64.98 | 76.31 | 84.79 | 88.67 |
| $F^3OCUS$ | **41.88** | **41.70** | **86.63** | **81.28** | **80.64** | **79.11** | **61.77** | **69.47** | **66.44** | **80.15** | **87.76** | **90.67** |

Table 5. Time and Memory Usage on Server for 10 clients with varying population size (P), *i.e.* candidate solutions every iteration

| | ABC | | ACO | | MOPSO | | SA | NSGA | |
|---|---|---|---|---|---|---|---|---|---|
| | P=25 | P=50 | P=25 | P=50 | P=25 | P=50 | | P=25 | P=50 |
| Time (s) | 0.15 | 0.33 | 0.19 | 0.41 | 0.17 | 0.40 | 0.01 | 5.91 | 35.51 |
| Peak Memory (MB) | 135.13 | | 135.78 | | 136.02 | | 135.82 | 136.92 | |

Table 6. Comparison of meta-heuristic methods for different tasks

| **Algorithm** | **Task 1** | **Task 2** | **Task 3** | **Task 4** | **Task 5** | **Task 6** | **Overall** |
|---|---|---|---|---|---|---|---|
| NSGA | **39.85** | **78.78** | 74.86 | **76.70** | 77.00 | **87.53** | 72.12 |
| ABC | 38.02 | 78.70 | 74.43 | 75.22 | 75.19 | 86.25 | 71.97 |
| ACO | 37.90 | 78.60 | 74.32 | 74.64 | 76.46 | 85.99 | 71.98 |
| SA | 37.79 | 77.71 | 74.10 | 74.27 | 75.69 | 85.45 | 70.83 |
| MOPSO | 39.67 | 78.38 | **75.01** | 76.13 | **77.26** | 86.92 | **72.23** |

Table 7. Dimensions and time for ViLT and LLaVA adapter

| **Model** | **Metric** | **Encoder Weight** | **Encoder Bias** | **Decoder Weight** | **Decoder Bias** | **Entire Adapter** |
|---|---|---|---|---|---|---|
| ViLT | Dimension | $48 \times 48$ | $48 \times 48$ | $768 \times 768$ | $768 \times 768$ | - |
| | (Full) Time (s) | $0.00006 \pm 0$ | $0.00005 \pm 0$ | $0.0518 \pm 0.0008$ | $0.0508 \pm 0.0005$ | $0.10271 \pm 0.00094$ |
| | (Top 1) Time (s) | $0.00002 \pm 0$ | $0.00002 \pm 0$ | $0.0031 \pm 0.0002$ | $0.0031 \pm 0.0001$ | $0.00624 \pm 0.0002$ |
| LLaVA | Dimension | $256 \times 256$ | $256 \times 256$ | $4096 \times 4096$ | $4096 \times 4096$ | - |
| | (Full) Time (s) | $0.0017 \pm 0$ | $0.0014 \pm 0.00005$ | $1.4802 \pm 0.0123$ | $1.4404 \pm 0.0148$ | $2.925 \pm 0.0166$ |
| | (Top 1) Time (s) | $0.0006 \pm 0$ | $0.0005 \pm 0$ | $0.0882 \pm 0.0015$ | $0.0871 \pm 0.0013$ | $0.1764 \pm 0.0022$ |

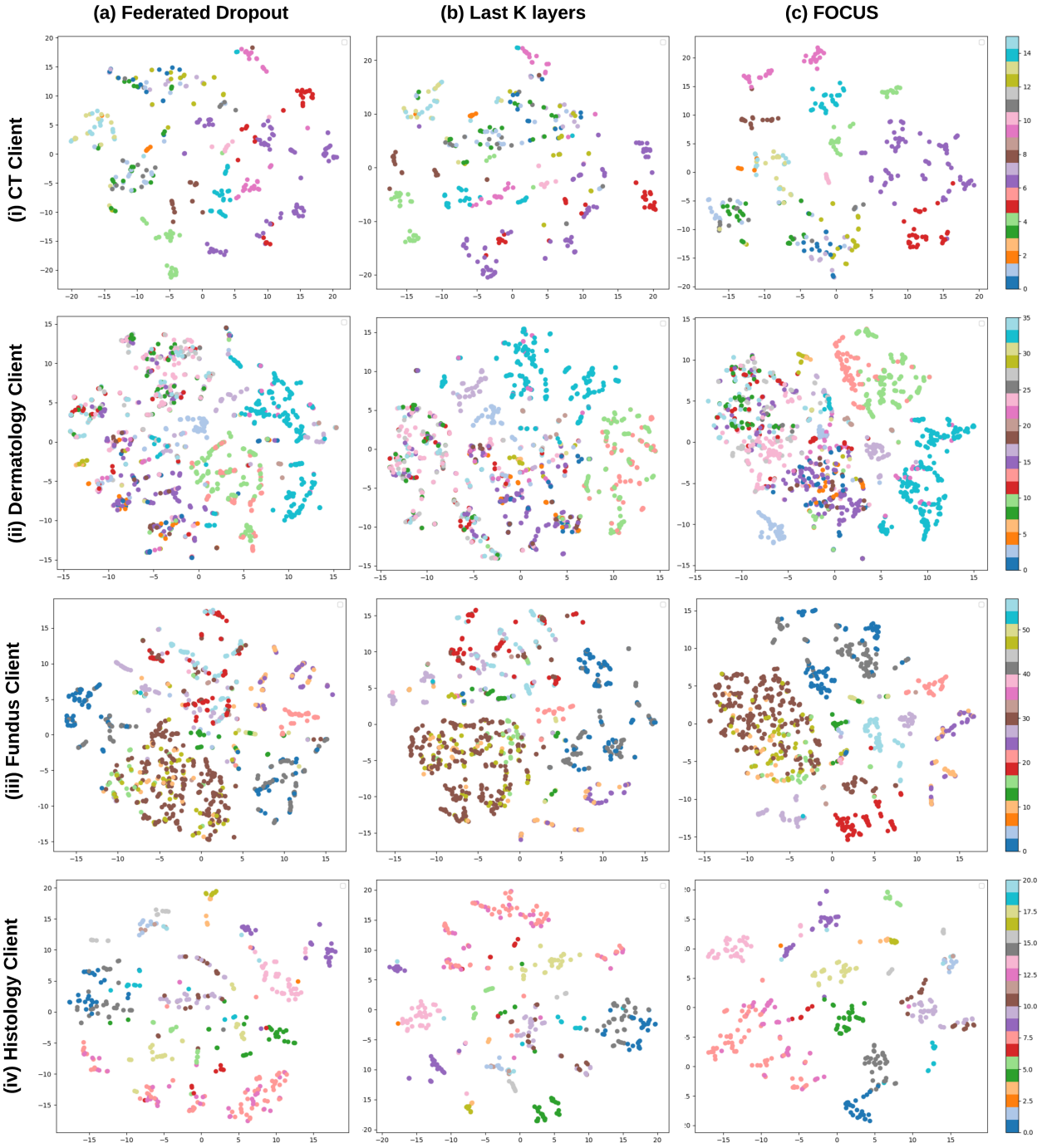


Figure 15. t-SNE feature embedding visualization for first four modality-specific clients of Task 2. (a) The first column denotes Federated Dropout. (b) The second column denotes fine-tuning last K layers. (c) The third column denotes our proposed method, $F^3OCUS$

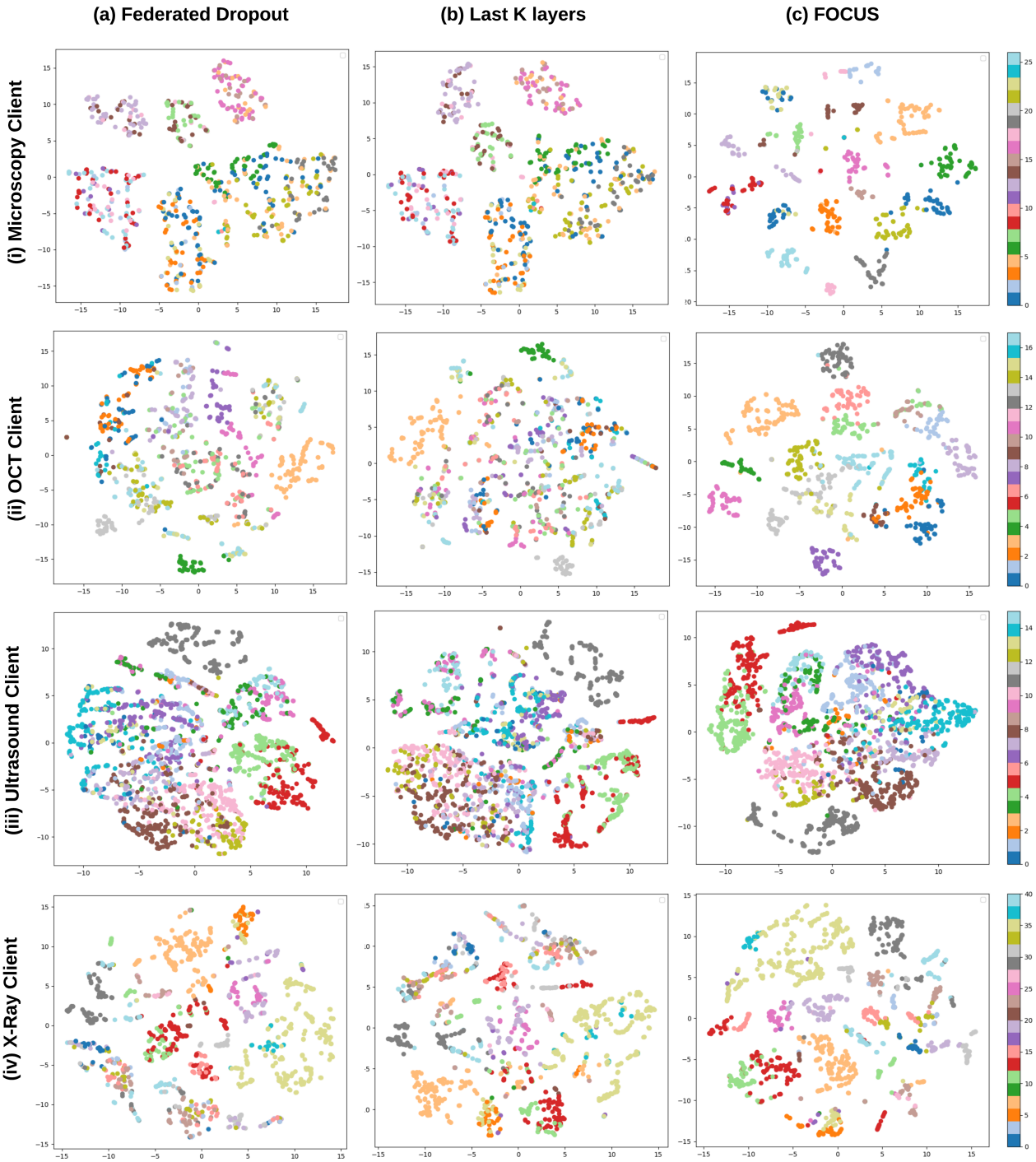


Figure 16. t-SNE feature embedding visualization for last four modality-specific clients of Task 2. (a) The first column denotes Federated Dropout. (b) The second column denotes fine-tuning last K layers. (c) The third column denotes our proposed method, $F^3OCUS$

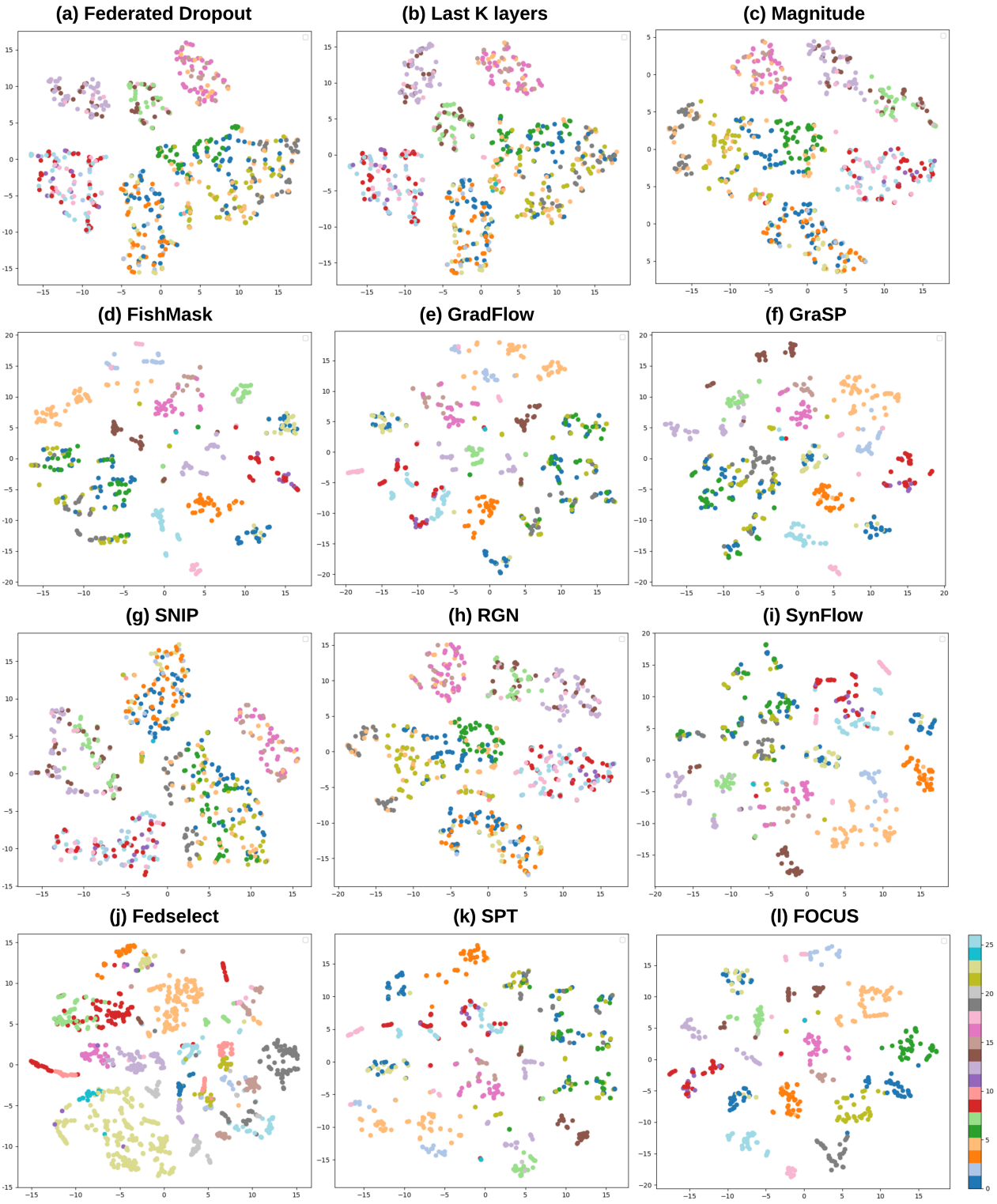


Figure 17. t-SNE feature embedding visualization for different layer selection methods on Microscopy client of Task 2.

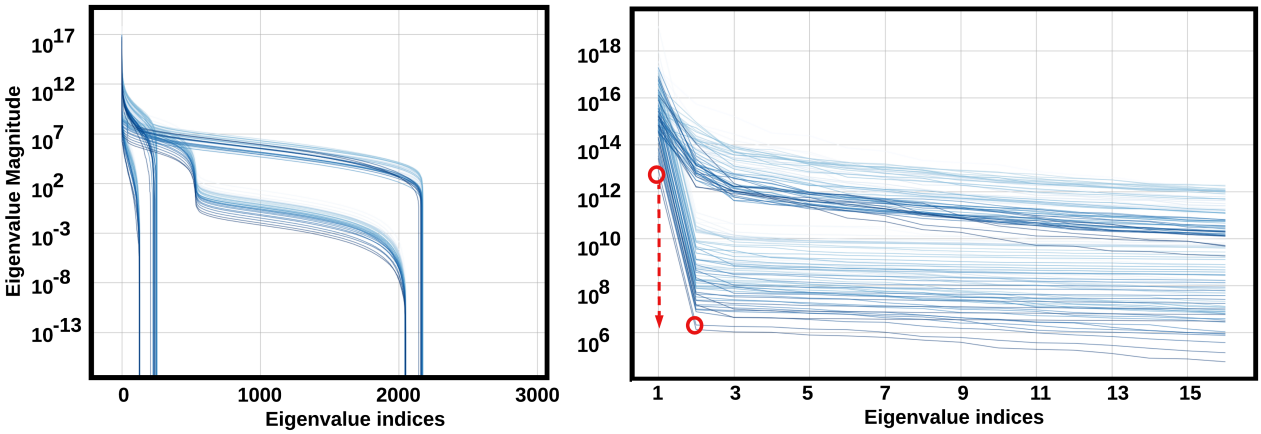


Figure 18. Layerwise eigenvalue spectrum visualization for LLaVA-1.5, with different shades of blue representing different layers. The left plot shows the full eigenvalue distribution, while the right focuses on the first 16 eigenvalues for a detailed view.

## References


[1] Blood cell images. https://www.kaggle.com/datasets/paultimothymooney/blood-cells, 2023. 23, 25

[2] X-ray hand small joint classification dataset (based on bone age scoring method rus-chn), 2023. 23, 24, 25

[3] Abeed S. Ben Abacha, Vivek V. Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *CLEF 2020 Working Notes*. 22

[4] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020. 22

[5] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. 22

[6] Amanullah Asraf and Zabirul Islam. COVID19, Pneumonia and Normal Chest X-ray PA Dataset, 2021. 23, 25

[7] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*, 2019. 22

[8] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*, 2021. 23

[9] Patrick Bilic, Patrick Ferdinand Christ, et al. The liver tumor segmentation benchmark (lits). *CoRR*, abs/1901.04056, 2019. 22

[10] Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature communications*, 12(1):4828, 2021. 23, 25

[11] Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11285–11293, 2024. 26

[12] Pingjun Chen. Knee Osteoarthritis Severity Grading Dataset, 2018. 23, 25

[13] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020. 23, 25

[14] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 23, 25

[15] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 22

[16] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020. 23, 25

[17] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild, 2019. 23, 25

[18] Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M Mossi, and Amparo Navea. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online*, 18:1–19, 2019. 23, 25

[19] Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. *Scientific Data*, 11(1):99, 2024. 23, 25

[20] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021. 23, 25

[21] Anubha Gupta and Ritu Gupta. Isbi 2019 c-nmc challenge: Classification in cancer cell imaging. *Select Proceedings*, 2, 2019. 23, 25

[22] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 23, 25

[23] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022. 28

[24] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 28

[25] Towhidul Islam, Mohammad Arafat Hussain, Forhad Uddin Hasan Chowdhury, and BM Riazul Islam. A web-scraped skin image database of monkeypox, chickenpox, smallpox, cowpox, and measles. *biorxiv*, pages 2022–08, 2022. 23, 25

[26] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6): 475, 2014. 23, 25

[27] Soroush Javadi and Seyed Abolghasem Mirroshandel. A novel deep learning method for automatic assessment of human sperm images. *Computers in biology and medicine*, 109: 182–194, 2019. 23, 25

[28] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 28

[29] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, 2018. 23, 25

[30] Jakob Nikolas Kather, Johannes Krisam, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):1–22, 2019. 22

[31] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 22

[32] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018. 23, 24

[33] Daniel S. Kermany, Michael Goldbaum, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122 – 1131.e9, 2018. 22

[34] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 22

[35] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 22

[36] Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. *Plos one*, 14(9):e0222025, 2019. 23, 25

[37] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6), 2022. 23, 25

[38] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012. 22

[39] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5): e210315, 2022. 23, 24

[40] Loris Nanni, Michelangelo Paci, Florentino Luciano Caetano dos Santos, Heli Skottman, Kati Juuti-Uusitalo, and Jari Hyttinen. Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium. *PLoS One*, 11(2), 2016. 23, 25

[41] R OpenAI et al. Gpt-4 technical report. *ArXiv*, 2303, 2023. 22

[42] Nikita V Orlov, Wayne W Chen, David Mark Eckley, Tomasz J Macura, Lior Shamir, Elaine S Jaffe, and Ilya G Goldberg. Automatic classification of lymphoma images with transform-based global features. *IEEE Transactions on Information Technology in Biomedicine*, 14(4), 2010. 23, 25

[43] Andre GC Pacheco and Renato A Krohling. The impact of patient clinical information on automated skin cancer detection. *Computers in biology and medicine*, 116:103545, 2020. 23, 25

[44] Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022. 23, 25

[45] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017. 23, 25

[46] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021. 23, 25

[47] Uğur Şevik, Cemal Köse, Tolga Berber, and Hidayet Erdöl. Identification of suitable fundus images using automated quality assessment methods. *Journal of biomedical optics*, 19(4):046006–046006, 2014. 23, 25

[48] Fariba Shaker, S Amirhassan Monadjemi, Javad Alirezaie, and Ahmad Reza Naghsh-Nilchi. A dictionary learning approach for human sperm heads classification. *Computers in biology and medicine*, 91:181–190, 2017. 23, 25

[49] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 22

[50] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv*, pages 2020–04, 2020. 23, 24

[51] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopatholog-

ical image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015. 23, 25

[52] John Suckling. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series, 1994*, pages 375–378, 1994. 23, 25

[53] Aliyun Tianchi. Covid-19 image dataset: 3 way classification - covid-19, viral pneumonia, normal. https://tianchi.aliyun.com/dataset/93853, . 23, 25

[54] Aliyun Tianchi. Nlm - malaria data. https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html, . 23, 25

[55] Aliyun Tianchi. Chest ct-scan images dataset. https://tianchi.aliyun.com/dataset/93929, . 23, 24

[56] Aliyun Tianchi. Covid ct dataset. https://tianchi.aliyun.com/dataset/106604, . 23, 24

[57] Aliyun Tianchi. Diabetic retinopathy arranged - retina images with class labels for classification. https://tianchi.aliyun.com/dataset/93926, 2023. 23, 25

[58] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, page 180161, 2018. 22

[59] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1), 2020. 23, 25

[60] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 22

[61] Xiaosong Wang, Yifan Peng, et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 3462–3471, 2017. 22

[62] X. Xu, F. Zhou, et al. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE Transactions on Medical Imaging*, 38(8):1885–1898, 2019.

[63] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195, 2021.

[64] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 22