ICT Express xxx (xxxx) xxx

Contents lists available at ScienceDirect

ICT Express

journal homepage: www.elsevier.com/locate/icte

Task-adaptive vision experts routing via competency learning guided by predictive uncertainty

Donghyun Han [a], Yuseok Bae [a], Jung Uk Kim [b,*], Hyung-Il Kim [c,d,*]

[a] Visual Intelligence Research Section, Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon 34129, Republic of Korea
[b] School of Computing, Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin 17104, Republic of Korea
[c] School of Electronics and Computer Engineering, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Republic of Korea
[d] Department of Intelligent Electronics and Computer Engineering, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Republic of Korea

ARTICLE INFO

Keywords:
Vision foundation model
Expert routing
Image classification
Multi-task learning
Mixture-of-experts

ABSTRACT

Large-scale pre-trained vision models such as ViT, CLIP, and SAM provide strong foundations for diverse vision tasks, motivating recent Mixture-of-Experts (MoE) approaches that combine multiple experts. However, existing methods often rely on static or implicit routing strategies, limiting adaptability to task semantics and input characteristics. We propose a task-adaptive vision expert routing framework based on competency learning guided by predictive uncertainty. We define expert competency as the relative reduction in predictive uncertainty induced by inter-expert interaction, and formulate expert routing as a learning problem driven by this signal. Our method uses task embeddings derived from textual descriptions to guide expert routing, refines expert features through cross-expert interaction, and aggregates them adaptively into a unified representation. By directly optimizing routing and feature composition using an uncertainty-based competency signal, the model learns how expert collaboration improves task-specific prediction reliability. Extensive experiments on diverse vision tasks demonstrate superior generalization performance and adaptive routing behavior aligned with task semantics.

1. Introduction

Recently, a wide range of large-scale pre-trained models have been released, accelerating progress across various computer vision tasks. Models such as the Vision Transformer (ViT) [1], Contrastive Language-Image Pretraining (CLIP) [2], and Segment Anything Model (SAM) [3] exhibit strong generalization capabilities across diverse tasks and domains. Trained on vast datasets, these models acquire rich and transferable representations that enable robust performance even on downstream tasks with limited supervision. As a result, leveraging pre-trained models has become a central strategy for building effective and data-efficient vision systems.

To further enhance adaptability and performance, recent studies have explored combining multiple pre-trained models using Mixture-of-Experts (MoE) frameworks [4–6]. These approaches dynamically select or fuse models based on the task or input, aiming to exploit the complementary strengths of diverse experts. Recent works have also investigated combining heterogeneous models and modalities to improve robustness and adaptability across diverse tasks [7,8]. As one of the recent works, EAGLE [4] explores different strategies to combine features extracted from multiple encoders, analyzing how feature aggregation impacts multimodal model performance. MoVA [6] uses a large language model (LLM) [9–11] to perform context-aware selection of vision experts and fuses their features via a dedicated adapter. And, AM-RADIO [5] unifies diverse vision foundation models via multi-teacher distillation, combining their strengths into a single, efficient student model for downstream vision tasks.

Despite recent progress, existing methods face two key limitations: (*i*) reliance on static or implicit expert selection, and (*ii*) limited interaction among experts. Most approaches lack explicit signals to assess whether the selected or fused experts effectively satisfy task requirements [12], making adaptive expert routing difficult. Moreover, treating experts independently restricts collaboration and prevents experts from recognizing their contribution after interaction, motivating the need for mechanisms that support both inter-expert communication and self-awareness.

To address these limitations, we propose a novel framework for task-adaptive vision expert routing via competency learning. Our approach introduces a task-guided routing module that explicitly learns task-conditioned expert composition from semantic embeddings, together with a cross-expert feature refinement module that enables information

* Corresponding authors.
E-mail addresses: mpolio2@etri.re.kr (D. Han), baeys@etri.re.kr (Y. Bae), ju.kim@khu.ac.kr (J.U. Kim), hyungil.kim@chonnam.ac.kr (H.-I. Kim).

https://doi.org/10.1016/j.icte.2026.04.007
Received 30 December 2025; Received in revised form 10 April 2026; Accepted 14 April 2026
Available online 16 April 2026

exchange among experts. The key idea is to dynamically route visual inputs through multiple pre-trained vision experts based on task semantics and input context, while learning to construct more reliable and task-effective representations through uncertainty-driven feedback. To this end, we extract task embeddings from textual descriptions, such as dataset and task-level specifications, which capture high-level semantic information about the intended objective. These embeddings guide the routing module to assign adaptive weights to each expert, enabling task-specific expert selection rather than static or implicitly defined routing.

The refined features are aggregated into a unified representation guided by routing weights. During training, we introduce a competency signal defined as the ratio of predictive uncertainty before and after aggregation, encouraging routing strategies that improve prediction reliability. Extensive experiments across a wide range of vision tasks demonstrate that our approach outperforms existing MoE-based and task-specific models, while exhibiting dynamic routing behavior aligned with task semantics.

Our contributions can be summarized in threefold as follows:

- We propose TAVER, a task-adaptive vision expert routing framework that learns expert competency through predictive uncertainty reduction induced by inter-expert interaction, enabling task-conditioned expert routing beyond static or heuristic selection.
- We design a competency learning strategy that leverages an uncertainty-based signal, defined as the ratio of predictive uncertainty before and after expert aggregation, to jointly guide expert routing and representation refinement.
- We validate TAVER on a wide range of vision tasks, demonstrating superior generalization over existing MoE-based and task-specific models, along with interpretable routing behavior aligned with task semantics.

2. Related works

2.1. Large-scale pretrained experts

Large-scale pre-trained vision models have significantly improved generalization across a wide range of computer vision tasks. Vision Transformer (ViT) [1] demonstrated the effectiveness of transformer architectures for image representation learning, while self-supervised approaches such as DINO [13] further enhanced representation without labeled data. Vision-language models like CLIP [2] enable strong zero-shot and cross-modal transfer by aligning visual and textual representations. Recently, the Segment Anything Model (SAM) [3] introduced promptable segmentation capabilities, providing high-quality spatial representations transferable across segmentation tasks. These models offer complementary strengths, motivating approaches that combine multiple vision experts for improved adaptability [14].

2.2. Mixture-of-experts

Mixture-of-Experts (MoE) frameworks aim to leverage multiple models by dynamically selecting or aggregating experts. Traditional MoE approaches [15–17] improve scalability through learned gating mechanisms, while recent works extend this paradigm to large pre-trained vision models. EAGLE [4] analyzes feature aggregation strategies across multiple vision encoders, and AM-RADIO [5] distills diverse foundation models into a unified student representation. MoVA [6] employs language models to perform context-aware expert selection and feature fusion. Beyond classification and segmentation, MoE-style strategies have also been explored in other vision domains such as image restoration [18–20] and enhancement [21], highlighting the flexibility of expert-based architectures across heterogeneous visual tasks [22].

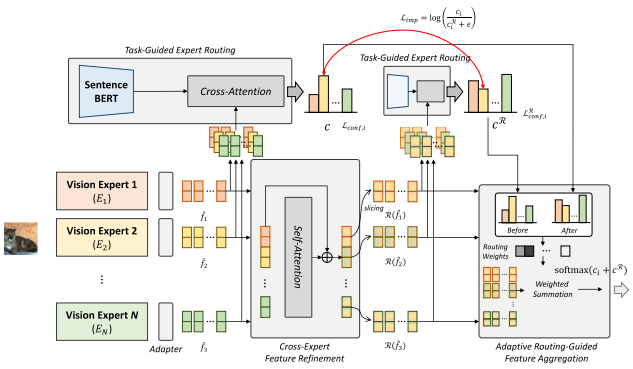


Fig. 1. Overview of the Task-Adaptive Vision Expert Routing (TAVER) framework.

While conventional MoE formulations primarily focus on parameter scalability or sparse activation within homogeneous expert layers, recent vision-oriented approaches increasingly combine heterogeneous, independently pre-trained foundation models. However, most existing methods emphasize expert selection or feature aggregation [12], without explicitly modeling how inter-expert interaction contributes to task-specific prediction reliability.

In contrast, TAVER defines expert competency as the relative reduction in predictive uncertainty induced by inter-expert interaction, and directly optimizes expert routing and feature composition using this signal. This shifts the focus from selecting experts to learning how expert collaboration improves prediction reliability in a task-adaptive manner. (see Fig. 1).

3. Proposed method

3.1. Cross-expert feature refinement

Given an input image I and a set of N pre-trained vision experts $E_1, E_2, \ldots, E_N$, we first extract features from each expert as $f_i = E_i(I)$. Since the experts operate on different input resolutions (e.g., 224×224 for CLIP and $1,024 \times 1,024$ for SAM), the input image is resized accordingly before being fed into each expert. As a result, the extracted features differ in both spatial and channel dimensions and cannot be directly merged.

To unify these heterogeneous features, we introduce a lightweight adapter module for each expert. The adapters align expert features to a shared representation space while enabling task-specific adaptation, with all vision experts kept frozen to preserve their generalization capability. Each adapter applies layer normalization [23] followed by two parallel linear projections, one of which is activated by GeLU [24]. The two projected outputs are summed element-wise, followed by an additional layer normalization, yielding aligned features $\hat{f}_i = \text{Adapter}(E_i(I))$. This design enables minimal yet effective adaptation of expert features for subsequent refinement and routing.

To reduce redundancy among experts and enhance complementary information, we further introduce a cross-expert feature refinement module $\mathcal{R}$. The aligned features $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_N$ are concatenated and used to compute expert-wise attention. Specifically, the concatenated feature is projected into query, key, and value matrices, and a multi-head attention mechanism is applied with the number of heads set to N, allowing each expert to attend to others explicitly. The attention output is added to the original concatenated feature via a residual connection, preserving original information while enhancing refined representations. Finally, the refined feature is split back into N expert-wise segments, producing refined expert features $\mathcal{R}(\hat{f}_i)$ for subsequent routing and aggregation.

This refinement stage enables explicit interaction among heterogeneous experts, allowing complementary cues to be emphasized while suppressing redundant responses. The refined representations $\mathcal{R}(\hat{f}_i)$ therefore provide a context-aware basis for subsequent routing, where the relative competency of each expert is assessed before and after inter-expert interaction.

3.2. Task-guided expert routing

In parallel with feature refinement, we introduce a task-guided expert routing module that enables adaptive composition of expert features conditioned on task semantics. To represent each task, we use a textual description that includes both the high-level task objective (e.g., classification) and dataset-specific properties such as modality or label type. These structured task descriptions are encoded into task embeddings using Sentence-BERT [25]. The resulting embedding serves as the query in a cross-attention mechanism, where the expert features prior to refinement, $\hat{f}_1, \ldots, \hat{f}_N$, are used as key and value. Through this cross-attention, routing decisions are determined by the interaction between the task embedding and the visual representations of each expert, rather than by text information alone.

The attended representations are passed through a feedforward network (linear-GeLU-linear), producing two outputs per expert: (1) a task-specific prediction (e.g., class logits), and (2) a scalar confidence score. Formally, for each expert i, we compute

$$c_i = \sigma\big(\mathrm{FFN}(\mathrm{CrossAttn}(t, \hat{f}_i))\big), \tag{1}$$

where t denotes the task embedding, $\mathrm{CrossAttn}(\cdot)$ represents the cross-attention operation, $\mathrm{FFN}(\cdot)$ is a lightweight feedforward network, and $\sigma(\cdot)$ is the sigmoid function. The confidence score $c_i \in (0, 1)$ reflects the estimated reliability of expert i for the given input and task.

These confidence scores $c_1, \ldots, c_N$ are later used to compute routing weights that determine the relative contribution of each expert in the final representation. In addition to the confidence computed from the original aligned features, the same routing process is applied to the refined features $\mathcal{R}(\hat{f}_1), \ldots, \mathcal{R}(\hat{f}_N)$, producing a second set of confidence scores, $c_1^{\mathcal{R}}, \ldots, c_N^{\mathcal{R}}$. This dual-stage estimation enables comparison of expert reliability before and after inter-expert interaction. The final expert weights are derived by jointly considering both pre- and post-refinement confidence:

$$w_i = \mathrm{SoftMax}(c_i + c_i^{\mathcal{R}}). \tag{2}$$

With $\mathrm{SoftMax}(c_i + c_i^{\mathcal{R}})$, experts with higher combined confidence contribute more, while less reliable ones are down-weighted; this weighting mainly depends on relative confidence, making it less sensitive to moderate miscalibration. This routing mechanism enables the model to assess expert competency in a task-aware manner and supports informed, adaptive expert composition based on the semantic alignment between task intent and expert knowledge.

3.3. Adaptive routing-guided feature aggregation

Given expert-wise routing weights $w_1, \ldots, w_N$ obtained from the task-guided expert routing module, we compute the final representation as a weighted sum of the refined expert features:

$$F_{\text{final}} = \mathrm{LayerNorm}\left(\sum_{i=1}^{N} w_i \cdot \mathcal{R}(\hat{f}_i)\right). \tag{3}$$

This aggregation reflects both the task relevance and reliability of each expert, as captured through the dual-stage confidence estimation. Experts with higher estimated confidence contribute more significantly, while less reliable ones are naturally down-weighted.

Since the routing weights are derived from the combined pre- and post-refinement confidence scores, the aggregation step directly incorporates the estimated competency of each expert after inter-expert interaction. As a result, the final representation integrates not only task alignment but also the relative reliability gained through refinement.

3.4. Training objectives

In this section, we present the loss functions designed to optimize our task-adaptive vision expert routing framework. The training objectives enable effective competency learning and adaptive routing.

Uncertainty-guided Confidence Loss Conventional Mixture-of-Experts approaches typically aggregate expert outputs using fixed or heuristic weights [26], without explicitly aligning expert confidence with task performance. In TAVER, each expert learns a task-conditioned confidence score that reflects its reliability for a given input.

We introduce an uncertainty-guided confidence loss defined as:

$$\mathcal{L}_{\text{conf},i} = \frac{1}{2}\left\{c_i \mathcal{L}_{\text{task}} + \log\left(\frac{1}{c_i + \epsilon}\right)\right\}, \text{ where } \mathcal{L}_{\text{task}} = -\sum_{\forall j \in C} y_j \log \hat{y}_j, \tag{4}$$

where c_i denotes the predicted confidence of expert i, $\mathcal{L}_{\text{task}}$ is the task loss (cross-entropy for classification and semantic segmentation in this work), y_j is the ground-truth label, and $\hat{y}_j$ is the predicted probability. The first term increases the contribution of the task loss when the predicted confidence is high, while the logarithmic regularizer discourages overconfident predictions. This formulation encourages alignment between confidence and actual task performance, improving the reliability of confidence estimates. The uncertainty-guided confidence loss is applied to each expert both before and after Cross-Expert Feature Refinement. For N experts, a total of $2N$ losses are computed and averaged:

$$\mathcal{L}_{\text{total\_conf}} = \frac{1}{2N}\sum_{i=1}^{N}\left(\mathcal{L}_{\text{conf},i} + \mathcal{L}_{\text{conf},i}^{\mathcal{R}}\right), \tag{5}$$

where $\mathcal{L}_{\text{conf},i}^{\mathcal{R}}$ denotes the confidence loss computed from the refined features $\mathcal{R}(\hat{f}_i)$.

Confidence Improvement Loss While the uncertainty-guided confidence loss aligns confidence with task performance, we further encourage effective inter-expert collaboration by promoting confidence improvement after refinement.

To capture confidence changes before and after Cross-Expert Feature Refinement, we define the confidence improvement loss as:

$$\mathcal{L}_{\text{imp},i} = \log\left(\frac{c_i}{c_i^{\mathcal{R}} + \epsilon}\right), \tag{6}$$

where c_i denotes the confidence before refinement and $c_i^{\mathcal{R}}$ denotes the confidence after refinement. This loss encourages relative improvement in confidence after inter-expert interaction, linking competency learning directly to the refinement process without enforcing trivial confidence inflation. The confidence improvement loss is applied independently to each expert and averaged across all experts:

$$\mathcal{L}_{\text{total\_imp}} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{imp},i}. \tag{7}$$

Total Loss The total loss function for training TAVER combines the task-specific loss, the uncertainty-guided confidence loss, and the confidence improvement loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{final\_task}} + \lambda_1 \mathcal{L}_{\text{total\_conf}} + \lambda_2 \mathcal{L}_{\text{total\_imp}}, \tag{8}$$

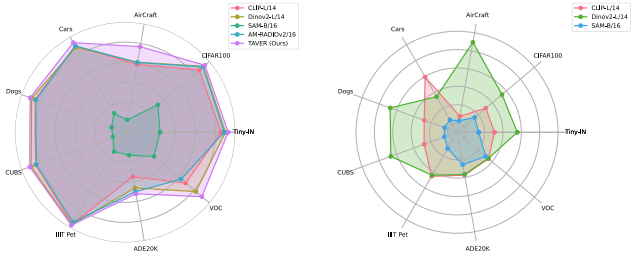
where $\mathcal{L}_{\text{final\_task}}$ is the task-specific loss applied to the final output of the model, and λ_1 and λ_2 control the relative importance of each component. In our experiments, we set $\lambda_1 = 10.0$ and $\lambda_2 = 1.0$.

4. Experiments

Experimental Setup We conduct all experiments with frozen vision experts (CLIP-L/14, DINOv2-L/14, SAM-B/16) and evaluate our model using linear probing to ensure fair comparison with baselines. Specifically, the final representation F_{final} is fed into a single linear

Table 1
Image classification accuracy (%).

Methods	Tiny	C100	A/C	Cars	Dogs	CUBS	Pet	Avg
CLIP-L/14	85.96	86.62	58.75	89.50	85.60	85.03	95.61	83.87
DINOv2-L/14	90.15	92.42	61.00	88.00	89.31	89.86	96.35	86.76
SAM-B/16	24.92	32.70	2.48	11.79	4.42	2.94	12.22	11.62
AM-RADIOv2/16	89.04	91.52	60.70	89.56	84.73	84.39	93.73	84.84
Plain fusion	91.07	92.87	65.16	90.78	89.66	90.03	96.89	88.07
TAVER	**92.87**	**94.08**	**76.79**	**93.32**	**90.85**	**91.29**	**97.16**	**90.85**



(a) Performance Comparison (b) Weight distribution of TAVER

Fig. 2. Performance comparison across various datasets and the corresponding expert weight distributions inferred by TAVER.

Table 2
Average expert routing weights for image classification.

Expert	Tiny	C100	A/C	Cars	Dogs	CUBS	Pet
CLIP-L/14	0.308	0.309	0.077	**0.595**	0.280	0.282	**0.454**
DINOv2-L/14	**0.556**	**0.539**	**0.896**	0.347	**0.675**	**0.668**	0.441
SAM-B/16	0.135	0.151	0.026	0.056	0.044	0.048	0.104

classification head for image classification and a linear segmentation head for pixel-wise prediction, while all backbone parameters remain frozen. For all experiments, we use PyTorch 2.1 [27] and train on NVIDIA RTX A6000 GPUs with AdamW optimizer [28], applying a cosine annealing learning rate schedule [29]. Each expert model receives input resized to its native resolution before being processed. For SAM, we utilize only the frozen image encoder without any prompt-based decoder (e.g., points, boxes, or masks), and evaluate its representation through linear probing. As a baseline, we define *Plain fusion* as the simple average of aligned expert features without task-guided routing, i.e., $F_{\text{plain}} = \frac{1}{N}\sum_{i=1}^{N} \hat{f}_i$, allowing isolation of the performance gain from competency-based routing. For classification tasks, all reported accuracies correspond to standard top-1 accuracy (%), where only the highest-scoring predicted class is considered correct.

Datasets For classification tasks, we evaluate on seven diverse datasets: Tiny ImageNet (Tiny) [30], CIFAR-100 (C100) [31], FGVC-Aircraft (A/C) [32], Stanford Cars (Cars) [33], Stanford Dogs (Dogs) [34], CUB-200-2011 (CUBS) [35], and IIIT Pet (Pet) [36]. For semantic segmentation, we evaluate on ADE20K [37] and Pascal VOC [38] datasets using standard metrics. For each dataset, a structured task description is constructed using task attributes such as task type (e.g., image classification or semantic segmentation), number of classes, and characteristic properties (e.g., fine-grained categories or low-resolution images). For example, the Tiny ImageNet task description includes attributes such as "image classification", "200 classes", and "64 × 64 resolution", which are serialized into a textual prompt and encoded via Sentence-BERT to generate task embeddings for routing. Task descriptions for all datasets are encoded using Sentence-BERT-L [25] to generate embeddings for the task-aware expert router.

4.1. Image classification results

Overall Performance Table 1 presents classification accuracy across diverse image datasets. Our proposed TAVER framework consistently outperforms all baseline methods, achieving an average improvement of 2.78 percentage point (pp) over the plain fusion approach. This improvement is particularly pronounced on the AirCraft dataset (11.63 pp gain), highlighting our model's effectiveness in fine-grained classification tasks. Even the plain fusion of expert features surpasses both individual models and AM-RADIOv2, suggesting that these pre-trained experts provide complementary information. TAVER further enhances performance by dynamically weighting each expert's contribution according to task-specific competency. Fig. 2 visualizes the performance comparison, clearly illustrating how TAVER consistently outperforms other methods across all domains.

To evaluate robustness, we conducted five independent runs on the Stanford Cars dataset under identical linear probing settings. The resulting accuracies ranged from 92.84 to 93.26, with a mean of 93.07 and a standard deviation of 0.21. These results demonstrate consistent performance across different random initializations.

Expert Routing Analysis Table 2 presents the average expert weights across different datasets. The distribution of weights varies significantly depending on the nature of the dataset. DINOv2-L/14 dominates with the highest weights across most datasets, particularly for FGVC-Aircraft (0.896), suggesting its self-supervised visual feature extraction capabilities are especially effective for fine-grained classification tasks. For Stanford Cars and IIIT Pet datasets, CLIP-L/14 is the most crucial expert with weights of 0.595 and 0.454 respectively, indicating that CLIP's image-text alignment pretraining may have learned more useful representations for these common object categories. While SAM-B/16 shows relatively lower weights across most datasets, it still contributes meaningfully, particularly for Tiny ImageNet and CIFAR-100 with weights exceeding 0.1. These results demonstrate that our TAVER framework adaptively combines experts based on both dataset characteristics and individual image properties. These task-dependent weight distributions further suggest that the routing mechanism aligns task semantics with expert-specific representation strengths, grounding textual task descriptions in data-driven visual evidence.

4.2. Ablation studies

Table 3 illustrates the impact of various components in our framework. Adding Cross-Expert Feature Refinement improves performance across most datasets, with notable gains on FGVC-Aircraft (+4.05 pp) and Stanford Cars (+2.16 pp). Incorporating the confidence loss function ($\mathcal{L}_{\text{conf}}$) further enhances performance, particularly on FGVC-Aircraft (+5.31 pp), suggesting that learning to predict expert reliability significantly benefits fine-grained tasks. Finally, adding the confidence improvement loss ($\mathcal{L}_{\text{imp}}$) yields consistent gains across all datasets, with FGVC-Aircraft showing a remarkable total improvement of 11.63 pp over plain fusion. These results demonstrate the complementary contributions of each component to the model's adaptability and performance.

Notably, the performance consistently increases as refinement and competency-related objectives are introduced sequentially. This trend indicates that confidence alignment ($\mathcal{L}_{\text{conf}}$) and refinement-induced confidence improvement ($\mathcal{L}_{\text{imp}}$) contribute progressively to more reliable expert routing and feature composition. The Cross-Expert Feature Refinement serves as a crucial component enabling interaction between experts. Without this mechanism, expert features cannot effectively enhance each other, and confidence changes cannot be measured. The average performance gain of 1.36 pp from this component alone highlights the importance of inter-expert interactions in building stronger feature representations.

Expert Combination Analysis Table 4 reports FGVC Aircraft classification results for different expert combinations. Although SAM-B/16 performs poorly in isolation (2.48), its contribution becomes significant when combined with other experts. In particular, pairing SAM-B/16 with CLIP-L/14 within TAVER achieves 65.79 accuracy, outperforming

Table 3
Ablation study on the impact of each component.

Methods	Tiny-IN	CIFAR100	AirCraft	Cars	Dogs	CUBS	IIIT Pet
Plain fusion	91.07	92.87	65.16	90.78	89.66	90.03	96.89
+ Cross-Expert Feature Refinement	92.08	93.62	69.21	92.94	90.00	90.82	96.84
+ $\mathcal{L}_{\text{conf}}$	92.34	94.05	74.52	93.24	90.80	90.96	96.94
+ $\mathcal{L}_{\text{imp}}$	**92.87**	**94.08**	**76.79**	**93.32**	**90.85**	**91.29**	**97.16**

Table 4
FGVC Aircraft classification performance by expert combination.

Combination strategy of experts		Accuracy
CLIP-L/14		58.75
DINOv2-L/14		61.00
SAM-B/16		2.48
Plain fusion		65.16
TAVER	CLIP + DINOv2	74.10
	CLIP + SAM	65.79
	DINOv2 + SAM	70.29
	CLIP+DINOv2+SAM	76.79

Table 5
Evaluation of segmentation results on the ADE20K and Pascal VOC datasets.

Methods	ADE20K		VOC	
	mIoU	mAcc	mIoU	mAcc
CLIP-L/14	35.22	47.16	68.98	79.23
DINOv2-L/14	46.41	58.52	82.19	89.08
SAM-B/16	13.26	18.21	27.65	37.00
AM-RADIOv2/16	50.22	62.92	84.87	90.37
Plan fusion	51.03	62.94	86.21	92.48
TAVER (Ours)	**52.41**	**66.52**	**90.04**	**94.99**

Table 6
Average weights of each expert at semantic segmentation.

Expert weights	ADE20K	VOC
CLIP-L/14	**0.375**	**0.347**
DINOv2-L/14	0.366	0.344
SAM-B/16	0.259	0.308

CLIP-L/14 alone (58.75) and plain fusion (65.16). These results indicate that TAVER effectively exploits complementary information from diverse experts, including individually weak ones. The improvement observed in combinations including SAM suggests that its structural representations provide complementary cues that are not sufficiently captured by semantic-rich experts alone. While SAM lacks strong categorical discriminativeness when used independently, its spatial sensitivity enhances feature refinement and supports more reliable routing decisions when integrated within the competency learning framework. Furthermore, the highest performance achieved by combining all three experts (76.79) indicates that routing dynamically balances semantic and structural information, rather than relying on a single dominant representation space.

Robustness to Noisy Task Descriptions. To assess sensitivity to task-description corruption, we perturb the task embedding at inference time by mixing the original description with noise. As shown in Table 7, performance decreases gradually as noise increases, yet remains stable under moderate corruption (30%–50%). Even with fully corrupted descriptions (100%), TAVER achieves accuracy comparable to plain fusion, indicating that routing is grounded in image-conditioned visual features rather than relying solely on text.

4.3. Semantic segmentation results

Tables 5 and 6 summarize the performance and expert routing weights for semantic segmentation. TAVER consistently outperforms all baselines in both mIoU and mAcc. On ADE20K, TAVER achieves 52.41 mIoU and 66.52 mAcc, outperforming AM-RADIOv2 by 2.19 pp and 3.60 pp, respectively. Compared to Plain fusion, TAVER improves mIoU by 2.07 pp on ADE20K and 3.42 pp on Pascal VOC, indicating that task-adaptive routing remains beneficial for pixel-level prediction.

The expert routing behavior differs from classification, exhibiting a more balanced distribution across experts. In particular, SAM's contribution increases from an average of 0.09 in classification to 0.259 on ADE20K and 0.308 on VOC, reflecting the importance of spatial cues in segmentation. Although only the frozen SAM image encoder is used, its structural representations complement the semantic features of CLIP and DINOv2, contributing to improved boundary awareness and region consistency. At the same time, the weight gap between CLIP and DINOv2 narrows, indicating complementary contributions for pixel-level prediction. This shift in expert weights suggests that routing is conditioned not only on dataset identity but also on task characteristics, prioritizing spatially informative experts when dense prediction is required.

Table 7
Robustness to noisy task descriptions (Stanford Cars, top-1 accuracy %). Noise is injected at inference time by mixing the original and noise task descriptions.

Setting	0%	30%	50%	100%	Plain
TAVER	93.32	92.78	92.42	89.25	90.78

Table 8
Computational cost comparison.

Model	Params (M)	GFLOPs	Latency (ms)
CLIP-L/14	303.6	155.6	19.5
DINOv2-L/14	304.8	155.6	23.0
SAM-B/16	93.9	974.0	129.3
AM-RADIOv2/16	1021.0	5271.2	136.2
TAVER (Ours)	717.2	1286.3	178.4

4.4. Computational cost analysis

As can be seen in Table 8, TAVER requires 1286.3 GFLOPs and 178.4 ms latency due to executing multiple expert backbones, yet remains substantially more efficient than AM-RADIOv2 (5271.2 GFLOPs). The higher complexity of AM-RADIOv2 is largely attributed to its large unified architecture and high-resolution processing with additional adapter components. Although the total parameter count of TAVER is 717M, only 19.9M parameters are trainable, since all backbone experts remain frozen. This significantly reduces training cost, as optimization is limited to lightweight routing and refinement modules.

5. Conclusion

In this paper, we introduced TAVER, a task-adaptive vision expert routing framework that dynamically combines multiple pre-trained vision experts based on task semantics and input characteristics. By defining expert competency as predictive uncertainty reduction induced by inter-expert interaction, TAVER learns to improve task-specific prediction reliability through uncertainty-guided routing and cross-expert feature refinement. Extensive experiments on diverse image classification and semantic segmentation benchmarks demonstrate that TAVER consistently outperforms existing MoE-based and task-specific

approaches, while exhibiting task-aligned and interpretable routing behavior. These results highlight the effectiveness of competency learning as a principled mechanism for adaptive expert composition in vision systems. While validated on classification and segmentation tasks, the proposed uncertainty-guided routing framework can be extended to broader vision applications, including safety-critical domains such as medical imaging and autonomous systems. Future work includes scalable expert routing strategies and lightweight fine-tuning to further enhance adaptability and efficiency.

CRediT authorship contribution statement

Donghyun Han: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation. **Yuseok Bae:** Writing – review & editing, Project administration, Investigation, Funding acquisition, Conceptualization. **Jung Uk Kim:** Writing – review & editing, Visualization, Validation, Supervision, Investigation, Formal analysis, Conceptualization. **Hyung-Il Kim:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the IITP grant funded by the Korea government (MSIT) (2022-0-00124), the “RISE” program through the Gwangju RISE Center funded by the MOE and Gwangju Metropolitan Government, Republic of Korea (2025-RISE-05-011), and the Basic Science Research Program through the NRF funded by the Ministry of Education (RS-2025-25398164).

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[2] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: Int’L Conf. Machine Learning, ICML, 2021, pp. 8748–8763.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: IEEE/CVF Int’L Conf. Computer Vision, ICCV, 2023, pp. 4015–4026.

[4] M. Shi, F. Liu, S. Wang, S. Liao, S. Radhakrishnan, Y. Zhao, D.-A. Huang, H. Yin, K. Sapra, Y. Yacoob, et al., Eagle: Exploring the design space for multimodal llms with mixture of encoders, 2024, arXiv preprint arXiv:2408.15998.

[5] M. Ranzinger, G. Heinrich, J. Kautz, P. Molchanov, AM-RADIO: Agglomerative vision foundation model reduce all domains into one, in: IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2024.

[6] Z. Zong, B. Ma, D. Shen, G. Song, H. Shao, D. Jiang, H. Li, Y. Liu, Mova: Adapting mixture of vision experts to multimodal context, 2024, arXiv preprint arXiv:2404.13046.

[7] M.M. Alam, M.A. Dini, D.-S. Kim, T. Jun, TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech, ICT Express (2025).

[8] M.A. Altaf, M.Y. Kim, Multiple object detection and tracking in autonomous vehicles: A survey on enhanced affinity computation and its multimodal applications, ICT Express (2025).

[9] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J.E. Gonzalez, I. Stoica, E.P. Xing, Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, 2023, URL https://lmsys.org/blog/2023-03-30-vicuna/.

[10] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., Qwen technical report, 2023, arXiv preprint arXiv:2309.16609.

[11] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, L. Hong, R. Tian, R. Xie, J. Zhou, M. Gerstein, D. Li, Z. Liu, M. Sun, ToolLLM: Facilitating large language models to master 16000+ real-world APIs, 2023, arXiv preprint arXiv:2307.16789.

[12] S.-G. Cheon, H.-J. Shin, S.-H. Bae, Region-aware knowledge distillation between monocular camera-based 3D object detectors, ICT Express (2025).

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: IEEE/CVF Int’L Conf. Computer Vision, ICCV, 2021, pp. 9650–9660.

[14] M. Awais, J. Choi, J. Park, Y.H. Kim, Intelligent data-aided semantic sensing with variational deep embedding, ICT Express 10 (4) (2024) 824–830.

[15] D. Eigen, M. Ranzato, I. Sutskever, Learning factored representations in a deep mixture of experts, 2013, arXiv preprint arXiv:1312.4314.

[16] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017, arXiv preprint arXiv:1701.06538.

[17] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V.Y. Zhao, A. Dai, Z. Chen, Q. Le, J. Laudon, Mixture-of-experts with expert choice routing, in: Annual Conf. Neural Information Processing Systems, NeurIPS, 2022.

[18] H. Guo, T. Dai, Y. Bai, B. Chen, X. Ren, Z. Zhu, S.-T. Xia, Parameter efficient adaptation for image restoration with heterogeneous mixture-of-experts, in: Annual Conf. Neural Information Processing Systems, NeurIPS, 2024.

[19] J. Jiang, Z. Zuo, G. Wu, K. Jiang, X. Liu, A survey on all-in-one image restoration: Taxonomy, evaluation and future trends, IEEE Trans. Pattern Anal. Mach. Intell. 47 (12) (2025) 11892–11911.

[20] Y. Ren, X. Li, B. Li, X. Wang, M. Guo, S. Zhao, L. Zhang, Z. Chen, Moe-diffir: Task-customized diffusion priors for universal compressed image restoration, in: European Conference on Computer Vision, Springer, 2025, pp. 116–134.

[21] M. Liao, H. Dong, X. Wang, K. Ubul, Y. Shao, Z. Yan, GM-MoE: Low-light enhancement with gated-mechanism mixture-of-experts, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2025, pp. 8766–8776.

[22] C. Rao, X. Fang, Y. Zhang, W. Fan, D. Zhou, Cross-domain autonomous driving visual segmentation based on enhanced target data learning, ICT Express 11 (1) (2025) 53–58.

[23] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, arXiv preprint arXiv:1607.06450.

[24] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), 2023, arXiv preprint arXiv:1606.08415.

[25] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Conf. Empirical Methods in Natural Language Processing, EMNLP, 2019, pp. 3982–3992.

[26] J. Puigcerver, C. Riquelme, B. Mustafa, N. Houlsby, From sparse to soft mixtures of experts, 2024, arXiv preprint arXiv:2308.00951.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Annual Conf. Neural Information Processing Systems (NeurIPS), Vol. 32, 2019.

[28] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: Int’L Conf. Learning Representations, ICLR, 2019.

[29] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: Int’L Conf. Learning Representations, ICLR, 2017.

[30] Y. Le, X. Yang, Tiny ImageNet visual recognition challenge, 2015, http://cs231n.stanford.edu/tiny-imagenet-200.zip.

[31] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images, Tech. Rep., University of Toronto, 2009.

[32] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, A. Vedaldi, Fine-grained visual classification of aircraft, in: British Machine Vision Conference, BMVC, 2013.

[33] J. Krause, M. Stark, J. Deng, L. Fei-Fei, Fine-grained visual classification of aircraft, in: IEEE/CVF Int’l Conf. Computer Vision Workshops, ICCVW, 2013.

[34] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel dataset for fine-grained image categorization: Stanford dogs, in: IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops, CVPRW, 2011.

[35] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200–2011 Dataset, Tech. Rep., California Institute of Technology, 2011.

[36] O.M. Parkhi, A. Vedaldi, A. Zisserman, C.V. Jawahar, Cats and dogs, in: IEEE/CVF Conf. Computer Vision and Pattern Recognition, CVPR, 2012.

[37] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ADE20k dataset, in: IEEE/CVF Conf. Computer Vision and Pattern Recognition, CVPR, 2017.

[38] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int’L J. Comput. Vis. (IJCV) (2010).