Supplementary Material for
DeAR: Fine-Grained VLM Adaptation by Decomposing Attention Head Roles

A Implementation Details

This section provides a comprehensive overview of the implementation details used for all experiments to ensure full reproducibility. Our code is built upon PyTorch and will be made publicly available.

A.1 Dataset and Dataloader Configuration

For all datasets, we follow a consistent data loading and augmentation pipeline:

- **Image Pre-processing:** Input images are resized to 224×224 pixels. For training, we apply standard augmentations including `random_resized_crop` with a scale range of $(0.5, 1.0)$ and `random_flip`. The interpolation method used is bicubic. Finally, all images are normalized using the standard CLIP pixel mean $[0.4814, 0.4578, 0.4082]$ and standard deviation $[0.2686, 0.2613, 0.2757]$.
- **Dataloader:** For training, we use a batch size of 16. For evaluation, the batch size is set to 250. We use 8 worker threads for data loading to maximize efficiency.

A.2 Training and Optimization

Our training setup is designed for stable and efficient convergence:

- **Optimizer:** We use the AdamW optimizer for all learnable parameters.
- **Learning Rate Schedule:** The initial learning rate is set to 1×10^{-3}. We employ a cosine annealing learning rate scheduler over a total of 10 epochs.
- **Warmup:** A 1-epoch constant warmup period is used at the beginning of training, with a warmup learning rate of 1×10^{-5}. This helps stabilize the model in the early stages of training.
- **Training Precision:** We use Automatic Mixed Precision (AMP) to accelerate training and reduce memory consumption.

A.3 Model and DeAR Configuration

Backbone and Layer Selection. All experiments are conducted using the official pre-trained CLIP model with a ViT-B/16 backbone. We strategically inject our learnable tokens into the deep layers (Layers 9 to 12) of the encoders, corresponding to $J = 9$ in the main text. This design choice is guided by prior findings demonstrating that functional specialization in Vision Transformers predominantly emerges in the final few layers. We further validated this through mean-ablation on our specific ViT-B/16 backbone.

Vision-Text Asymmetry. While the update mechanism is symmetric across modalities, the token definitions differ by design. Both the vision and text encoders insert learnable tokens at layers L_{9-12} with the same β-controlled flow (Eq. 5 & 6 in the main text) to prevent semantic drift. However, we intentionally exclude explicit Attribute Tokens from the text side. In CLIP-based classification, text features act as fixed anchors (via templates like "a photo of a {class}") for semantic alignment. Keeping the text side simplified prevents over-complication while maintaining the integrity of these semantic anchors.

DeAR Hyperparameters.

- **Token Injection:** We use a learnable prompt/token length of $K = 5$, selected via Density-Based Clustering Validation (DBCV).
- **Context Information Control:** The hyperparameter β which controls the context information flow is set to 0.9.
- **Loss Weights:** The balancing hyperparameters for our composite loss function are set to $\lambda_{\text{reg}} = 1.0$ and $\lambda_{\text{fusion}} = 0.7$. All hyperparameters were fixed across all 11 datasets after initial tuning on a validation set.

Computational Cost. During training, the regularization term $\mathscr{L}_{\text{reg}}$ requires one additional forward pass through the frozen backbone, which increases the overall training cost by approximately 35%. We find this trade-off highly acceptable given the significant zero-shot generalization gains achieved. Crucially, during inference, the computational overhead is exactly zero, as the final adapted model retains the standard ViT architecture without requiring auxiliary modules. All experiments were conducted on a server equipped with three NVIDIA RTX 4090 GPUs.

Table 1: Summary of key hyperparameters used in our experiments.

Hyperparameter	**Value**
Dataloader & Input	
Batch Size (Train / Test)	16 / 250
Input Resolution	224×224
Augmentations	`RRCrop, RandomFlip, Normalize`
Optimizer	
Optimizer Name	AdamW
Learning Rate (Initial)	1×10^{-3}
LR Scheduler	Cosine Annealing
Max Epochs	10
Warmup Epochs	1
DeAR Specific	
Backbone	ViT-B/16
Attribute Token Length (K)	5
Injection Layers	9, 10, 11, 12
Knowledge Anchoring β	0.9
Regularization λ_{reg}	1.0
Fusion Loss λ_{fusion}	0.7

B Functional Roles of Attention Heads

To build the Role-Based Attention Mask for our DeAR framework, we systematically analyzed the top descriptive phrases for each attention head in the later layers (9-12) of the ViT-B/16 backbone.

Offline Concept Entropy Analysis. It is important to note that our framework does **not** rely on downstream task descriptions or labels for head discovery. Following TEXTSPAN, the Concept Entropy is computed via a one-time, offline analysis using a generic text corpus on the pre-trained backbone. The identified head roles are intrinsic to the model, kept frozen, and universally transferred to all downstream tasks without the need for re-discovery.

Q, K, V Masking Mechanism. To strictly enforce functional separation, we implement the role isolation directly within the self-attention mechanism. Specifically, a mask matrix (containing $-\infty$ values) is applied to the Attention matrix before the Softmax operation. This definitively blocks the newly inserted Attribute Tokens from attending to (or being attended by) the original tokens associated with non-target heads, fully preserving the pre-trained backbone's innate generalization flow.

Conceptual Clusters and Role Assignment. Our unsupervised analysis using TEXTSPAN and HDBSCAN identified **12 distinct conceptual clusters**: *Location, Object, Color, Shape, Texture, Animal, Human, Text, Number, Style, Emotion, and Action*. We classify each head into one of four functional roles:

1. **Core Attribute Heads:** Heads focusing on the five selected core concepts (*Color, Shape, Texture, Object, Location*). These heads are assigned specific attribute tokens via Eq. 9.
2. **Other Specialized Heads:** Heads focusing on specific distinct concepts that are not part of the core set (e.g., *Animal, Human*). To prevent interference, they are isolated from the new attribute tokens (using Eq. 8).
3. **Generalization Heads:** Heads focusing on abstract concepts or global composition (e.g., *Style*). These are critical for zero-shot robustness and are also isolated.
4. **Mixed Heads:** Heads with no single dominant focus, allowing unrestricted attention flow.

Table 2 details the classification for all heads in Layers 9-12.

Table 2: **Detailed functional role classification for Attention Heads in Layers 9–12 of ViT-B/16.** We categorize each head into four roles: **Core Attribute**, **Other Specialized**, **Generalization**, and **Mixed**.

Head	Role (Concept)	Representative Phrases
		Layer 9
(9, 0)	**Core Attribute** (Location)	*Photo taken in Galápagos Islands, Photo taken in Okavango Delta, Bustling cityscape at night*
(9, 1)	**Core Attribute** (Location)	*Picture taken in Bhutan, Photo taken in the Rub' al Khali, An image of Andorra*
(9, 2)	**Mixed**	*A network of veins, Eyes, Photograph taken in a rustic barn, A wizard's hat*
(9, 3)	**Core Attribute** (Location)	*Aerial view of a snowy landscape, Picture taken in the Swiss chocolate factories, Australian coral reef*

Continued on next page

Table 2 – continued from previous page

Head	Role (Concept)	Representative Phrases
(9, 4)	**Other Specialized** (Emotion)	*Sarcastic raised eyebrow, Intrigued facial expression, A photo of a young person*
(9, 5)	**Generalization** (Style)	*Cinematic framing, Dramatic chiaroscuro photography, Retro-style poster design*
(9, 6)	**Mixed**	*Futuristic biotechnology, Artwork featuring zebra stripe motifs, Surreal digital collage*
(9, 7)	**Core Attribute** (Texture)	*Delicate ceramic patterns, Close-up of a textured synthetic rubber, Ethereal double exposure*
(9, 8)	**Other Specialized** (Text)	*Artwork featuring Morse code typography, Bold graffiti, Film noir-inspired tones*
(9, 9)	**Core Attribute** (Location)	*Bustling cityscape at night, Picture taken in a city park, Serene countryside sunrise*
(9, 10)	**Generalization** (Style)	*Detailed illustration of a celestial body, Stark minimalism, Geometric tessellation*
(9, 11)	**Mixed**	*An image of three subjects, A photo with the letter T, detailed reptile close-up*
		Layer 10
(10, 0)	**Core Attribute** (Object)	*Picture with a single domesticated animal, A trunk, A whisker, An irregular heptagon*
(10, 1)	**Core Attribute** (Color)	*Earthy color tones, Pop art colors, Image with a pink color, Film noir-inspired tones*
(10, 2)	**Core Attribute** (Texture)	*Photo of a furry animal, Close-up of a textured synthetic fabric, Marbleized design*
(10, 3)	**Core Attribute** (Location)	*Sunlit meadow path, Crumbling and abandoned building, Vibrant city alley*
(10, 4)	**Other Specialized** (Human)	*Image with a team of subjects, An image of two subjects, Crowded and bustling scene*
(10, 5)	**Core Attribute** (Object)	*An image of a Fashion Designer, A photograph of a small object, Striking fashion stance*
(10, 6)	**Core Attribute** (Location)	*Gritty urban street scene, Serene meadow landscape, Breathtaking canyons*
(10, 7)	**Generalization** (Style)	*Playful reflections, Cubist still life painting, Stark minimalism, Timeless classic artwork*
(10, 8)	**Core Attribute** (Location)	*Picture taken in a city park, Busy airport terminal, Urban alleyway*
(10, 9)	**Core Attribute** (Object)	*An image of a dish, A photo of a young person, An image of fish, Colorful hot air balloons*
(10, 10)	**Core Attribute** (Shape)	*Image with an owl, A regular octagon, Image with a futuristic time travel device*
(10, 11)	**Generalization** (Style)	*Intricate pencil drawing, Unexpected symmetry, Futuristic transportation*
		Layer 11

Continued on next page

Table 2 – continued from previous page

Head	Role (Concept)	Representative Phrases
(11, 0)	**Core Attribute** (Object)	*A bookmark, A laptop, A bowl, A skirt, A jacket, A phone, A regular octagon*
(11, 1)	**Core Attribute** (Color)	*A platinum silver color, A gold color, A photo with the letter F, Image with a pink color*
(11, 2)	**Other Specialized** (Animal)	*Playful animals, Flowers, An image with dogs, Image with a butterfly, An image with seagulls*
(11, 3)	**Core Attribute** (Location)	*Picture snapped in the Alaskan mountains, Photo taken in Bangkok Thailand, Ocean*
(11, 4)	**Core Attribute** (Texture)	*Photo of a furry animal, A bookmark, Close-up of a textured synthetic wood, A skirt*
(11, 5)	**Core Attribute** (Texture)	*Artwork featuring Morse code typography, Delicate embroidery, Kaleidoscopic patterns*
(11, 6)	**Core Attribute** (Location)	*Cultural exhibition, Photograph taken in a cozy cafe, Picture snapped in the Greek islands*
(11, 7)	**Other Specialized** (Action)	*Joyful toddlers, Hands in an embrace, A paw, An image with pedestrians*
(11, 8)	**Generalization** (Style)	*Anime style image, A high-resolution image, Artwork featuring abstract fractal patterns*
(11, 9)	**Core Attribute** (Object)	*A necklace, A sock, A megaphone, A jacket, A fork, A belt, Precise clock mechanism*
(11, 10)	**Core Attribute** (Location)	*Tranquil boating on a lake, Peaceful rural farmland, Secluded beach cove*
(11, 11)	**Core Attribute** (Location)	*Bustling city nightlife, Secluded forest cabin, Energetic music festival crowd*
		Layer 12
(12, 0)	**Core Attribute** (Location)	*Picture taken in the Scotland countryside, Photo taken in Barcelona Spain, Photo taken in Tokyo*
(12, 1)	**Generalization** (Style)	*Reflections, Central focal point, Motion freeze, Captivating city pulse, Dramatic skies*
(12, 2)	**Generalization** (Style)	*Ephemeral glimmers, Whispering horizons, Symmetry disrupted, Unconventional beauty*
(12, 3)	**Mixed**	*Picture with multiple wild animals, An image of a Chef de Cuisine, Picture with cars*
(12, 4)	**Other Specialized** (Human)	*Playful siblings, Image with a five people, A photo of a woman, A group photo*
(12, 5)	**Core Attribute** (Shape)	*Photo of a reptile, Image with a seagull, A scalene triangle, A snail, Herringbone pattern*
(12, 6)	**Core Attribute** (Location)	*Photo taken in Namib Desert, Photo taken in the Alaskan mountains, Scottish moors*
(12, 7)	**Core Attribute** (Object)	*A stick, A cup, A bonnet, A shoelace, A bottle, A belt, A jacket, A puddle*
(12, 8)	**Mixed**	*A laptop, A rug, A shelf, A bag, Picture taken in an art gallery, Urban subway station*

Continued on next page

Table 2 – continued from previous page

Head	Role (Concept)	Representative Phrases
(12, 9)	**Other Specialized** (Number)	*An image of the number 10, An image of the number 7, The number fifteen*
(12, 10)	**Core Attribute** (Color)	*An image with cold green tones, Image with a red color, A charcoal gray color*
(12, 11)	**Other Specialized** (Text)	*A photo with the letter J, A photo with the letter K, A swirling eddy, A photo with the letter C*

C Complete Experiment Results

Our DeAR framework consistently outperforms previous approaches across all benchmark datasets.

C.1 Statistical Robustness and Variance

To ensure the statistical reliability of our improvements, we evaluate DeAR alongside a strong baseline (MMRL++) across multiple random seeds ($\{1, 2, 3\}$). As presented in Table 3, DeAR exhibits extremely low variance across independent runs and consistently outperforms the baseline on both the average of the 11 base-to-novel datasets and the challenging ImageNet-1K benchmark. This validates that the performance gains are robust and directly attributable to the proposed role-based masking strategy rather than initialization artifacts.

Table 3: **Robustness Analysis.** Results are reported as Mean $\pm$ Standard Deviation across 3 independent seeds.

	Average (11 Datasets)			ImageNet-1K		
Method	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**
MMRL++	85.53	78.32	81.77	**77.63**	71.50	74.44
DeAR (Ours)	**85.62**$_{\pm 0.21}$	**79.77**$_{\pm 0.11}$	**82.71**$_{\pm 0.18}$	77.52$_{\pm 0.39}$	**71.84**$_{\pm 0.17}$	**74.73**$_{\pm 0.23}$

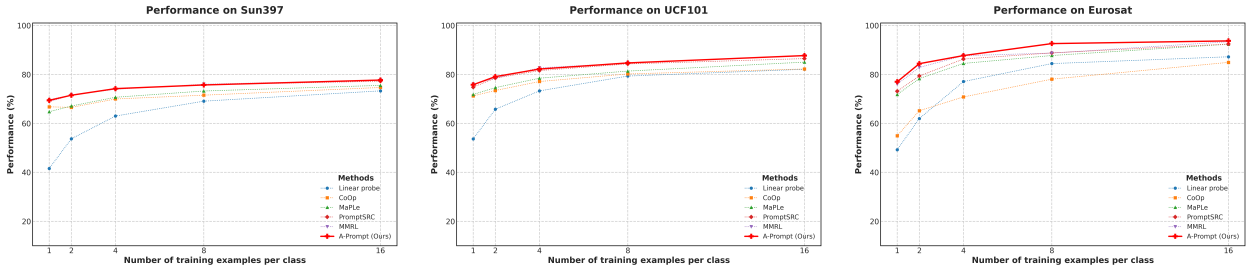
D Prompt Templates

Table 4 summarizes the specific prompt templates used for each dataset in our experiments. These templates are manually designed to better match the characteristics of each dataset.

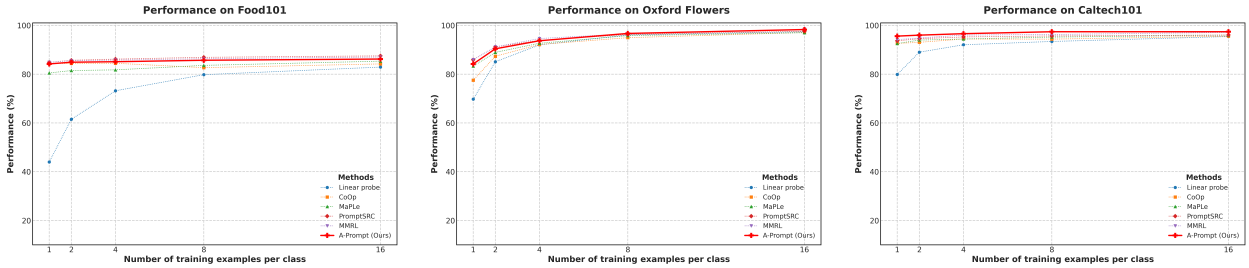
E Additional Ablation Studies

Task-Adaptive Fusion Logic and DTD Discrepancy. A key component of DeAR is its Task-Adaptive Fusion strategy. The purpose of $\mathscr{L}_{\text{fusion}}$ is to encourage the model to prioritize the generalization feature ($\mathbf{f}_{\text{cls}}$) as the primary decision-maker, utilizing the specialized attributes ($\mathbf{f}_{\text{attr}}$) solely for fine-grained refinement. Without this loss constraint, the model risks over-relying on learned attributes, which degrades zero-shot capability.

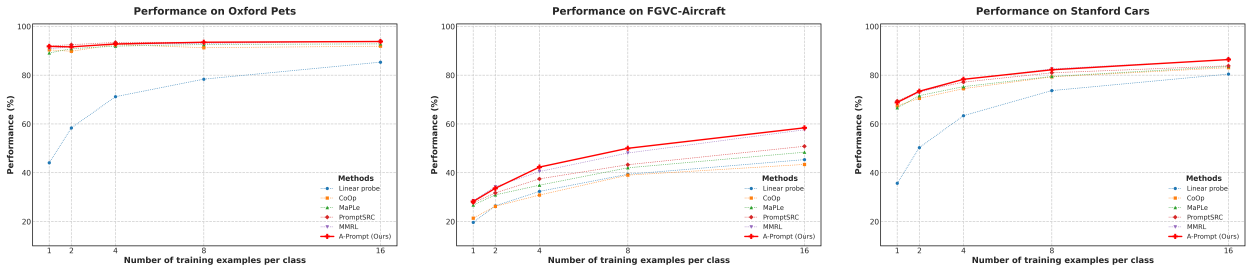
(a) Average (11 datasets) (b) ImageNet (c) DTD



(d) SUN397 (e) UCF101 (f) EuroSAT



(g) Food101 (h) Oxford Flowers (i) Caltech101



(j) Oxford Pets (k) FGVC Aircraft (l) Stanford Cars

Figure 1: **Few-Shot Performance Comparison.** Comparison of DeAR with previous methods across 11 datasets. DeAR consistently achieves superior performance in few-shot settings.

It is worth noting the behavioral difference on the Describable Textures Dataset (DTD). In the base-to-novel setting, predictions utilize the full Task-Adaptive Fusion ($\mathbf{f}_{\text{attr}} + \mathbf{f}_{\text{cls}}$), which heavily benefits texture-centric datasets like DTD. Conversely, in the cross-dataset evaluation protocol, predictions rely solely on the decoupled general-purpose feature ($\mathbf{f}_{\text{cls}}$) for strict fairness across methods. This discrepancy actually validates that our attribute tokens successfully capture critical fine-grained features that $\mathbf{f}_{\text{cls}}$ alone might miss.

As shown in Table 5, our proposed Task-Adaptive Fusion significantly outperforms a standard decoupled

Table 4: Prompt templates used for each dataset.

Dataset	Prompt Template
OxfordPets	a photo of a {}, a type of pet.
OxfordFlowers	a photo of a {}, a type of flower.
FGVCAircraft	a photo of a {}, a type of aircraft.
DescribableTextures	{} texture.
EuroSAT	a centered satellite photo of {}.
StanfordCars	a photo of a {}.
Food101	a photo of {}, a type of food.
SUN397	a photo of a {}.
Caltech101	a photo of a {}.
UCF101	a photo of a person doing {}.
ImageNet	a photo of a {}.
ImageNetSketch	a photo of a {}.
ImageNetV2	a photo of a {}.
ImageNetA	a photo of a {}.
ImageNetR	a photo of a {}.

strategy (+1.23% HM).

Table 5: Ablation study on the inference strategy. Average accuracy (%) on the base-to-novel benchmark.

Inference Strategy	Base	Novel	HM
Decoupled Inference	**85.94**	78.50	82.05
Task-Adaptive Fusion (Ours)	**85.94**	**79.73**	**82.72**

Threshold Sensitivity and Role Ablation. The Role-Based Mask uses Concept Entropy quantiles to separate specialized heads (bottom 20% entropy) from generic heads (top 20% entropy). To demonstrate the stability of this heuristic, we conduct a sensitivity analysis on ImageNet-1K (Table 6, Left). DeAR maintains a stable performance plateau across a broad threshold range (Top 10% to 30%), indicating it does not require fragile hyperparameter tuning.

Furthermore, accurately identifying these roles is paramount. As shown in Table 6 (Right), randomly shuffling the assigned roles results in a severe performance drop (from 74.73% to 70.85%), conclusively proving the necessity of the proposed entropy-based role decomposition.

Table 6: **Sensitivity Analysis and Ablation Studies on ImageNet-1K.**

Sensitivity (Entropy Threshold %)		**Role & Attribute (K) Ablation**	
Threshold Q.	**IN-1K HM (±Std)**	**Strategy**	**IN-1K HM (±Std)**
Top 10%	74.35 ± 0.21	Random Role Shuffle	70.85 ± 0.25
Top 20%	$\mathbf{74.73 \pm 0.23}$	**DeAR ($K = 5$)**	$\mathbf{74.73 \pm 0.23}$
Top 30%	74.28 ± 0.20	Concept Cluster $K = 3$	73.85 ± 0.29
Top 40%	71.10 ± 0.24	Concept Cluster $K = 7$	74.01 ± 0.33

Sensitivity to "Core Attribute" Selection. Our default design selects $K = 5$ core attributes. Table 6 (Right) also explores adjusting the granularity of these concept clusters. While $K = 3$ or $K = 7$ slightly underperform the optimal $K = 5$, they still substantially exceed baseline performances, demonstrating the framework's robustness to attribute granularity.

Visualization of Learned Task Priors. To further understand how DeAR adapts to different domains, we visualize the learned fusion weights α_k (averaged over heads of the same attribute type) across all 11 datasets in Figure 2.

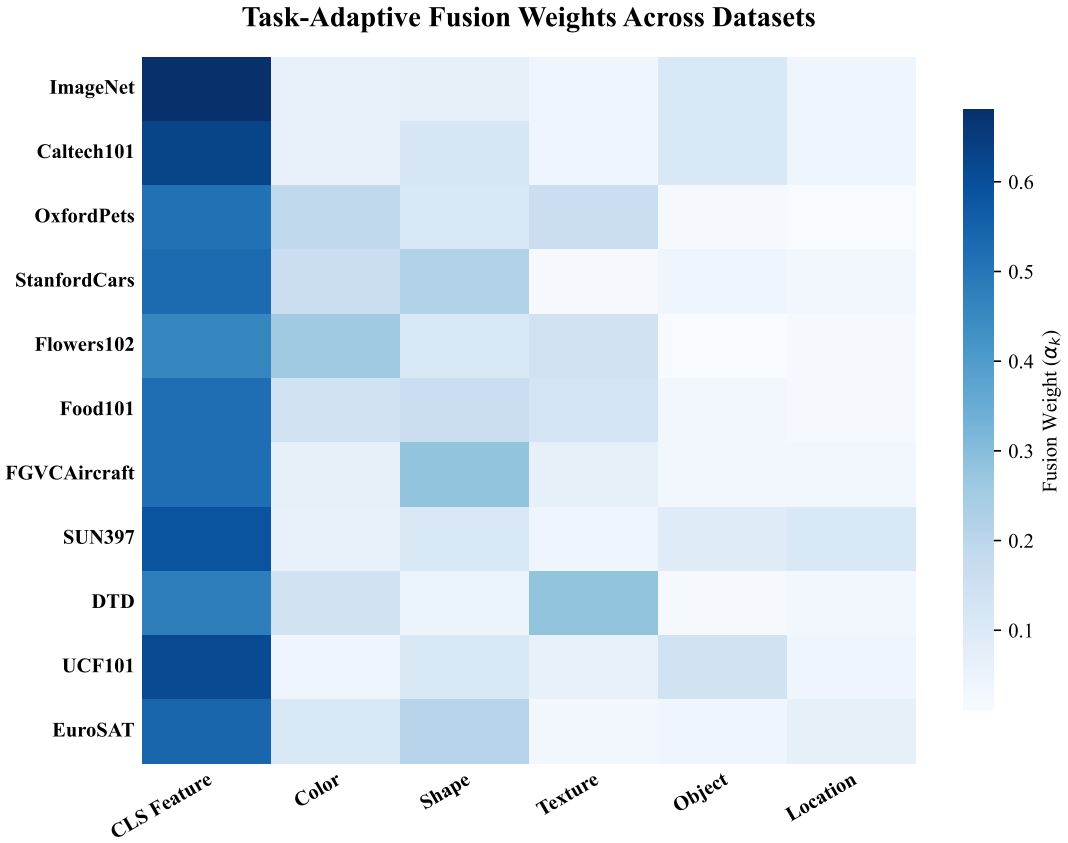


Figure 2: **Visualization of Task-Adaptive Fusion Weights Across Datasets.** The heatmap displays the magnitude of the learned fusion weights (α_k) for the global [CLS] feature and the five specialized attribute tokens. Darker blue indicates higher importance.

The heatmap reveals highly interpretable patterns:

- **DTD (Texture Dataset):** Shows a significantly higher weight for the **Texture** token.
- **OxfordPets & Flowers102:** Exhibit balanced but elevated attention to **Color** and **Shape**.
- **FGVCAircraft & StanfordCars:** Show reduced reliance on Color but maintain attention on **Shape**.
- **ImageNet:** Maintains a strong reliance on the global **[CLS] Feature**.

F Code

Code is available at our GitHub repository.