Discovering Attention Head Interactions in Vision Transformers

Supplementary Material

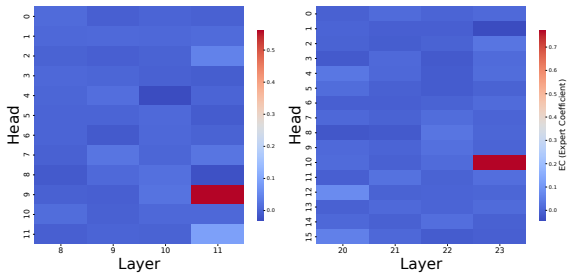

Figure 4. Layer–head distribution of EC on MNIST for ViT-B/16 (left) and ViT-L/14 (right).

Table 7. Interfering head experiments on Waterbirds dataset. Random heads removal is reported as mean ± std over five trials.

Model	Condition	$\text{Acc.}_{\text{worst}}$	Acc._{avg}
ViT-B/16	$\mathcal{M}$	47.04	76.60
	$\mathcal{M}_{\setminus R}$	49.90±2.47	76.16±1.26
	$\mathcal{M}_{\setminus A}$	**60.12**	**83.32**

10. Head Caption

This section provides additional examples of head captions that illustrate the semantic tendencies and visual focus of different attention heads in Vision Transformers. Each caption corresponds to a single attention head and summarizes the type of visual concepts or patterns that the head responds to most strongly. The examples in Table 8 demonstrate the diversity of head behaviors, covering texture sensitivity, object semantics, color preference, and shape awareness et al.

11. Visual evidence

This section provides additional visualization examples that illustrate the spatial behavior of attention heads associated with four relation types. Each relation is presented in a separate figure and includes several representative cases to show the typical attention distribution patterns.

Figure 5 displays examples for the expertise relation. Heads in this group tend to produce localized, concentrated responses on task-relevant regions, such as salient object parts or contours.

Figure 6 presents the collaboration relation. The examples show heads attending to complementary spatial areas; when considered together, these responses cover coherent object semantics more completely than any single head.

Figure 7 focuses on the interference relation. These heads often emphasize background or context regions that are not directly related to the target, suggesting a potential source of prediction noise.

Figure 8 summarizes the redundancy relation. The visualizations reveal overlapping attention patterns across similar regions, indicating limited additional contribution beyond other heads with comparable focus.

All figures are produced with the same visualization settings to ensure a consistent comparison across relations.

12. Additional Analyses on DeiT and Swin Transformers

To examine whether the interaction patterns observed in ViT generalize to other transformer-based architectures, we conducted additional analyses on the DeiT [41] and Swin [28] models. Both models, including head-caption generation and activation visualization, follow the setting of Balasubramanian et al. [3]. Corresponding visualization examples for these two models are shown in Figure 9 and Figure 10. Representative textual descriptions of attention heads are provided for Swin in Table 9 and for DeiT in Table 10.

These analyses suggest that similar relational tendencies may also emerge in other transformer architectures, indicating a potential generality of the proposed taxonomy.

13. Experiment setting

In experiments of main paper, we use publicly available pretrained OpenCLIP-ViT [39] models for zero-shot inference without any additional training or fine-tuning. Specifically, we adopt the ViT-B/16 and ViT-L/14 variants.

These models are directly used for encoding visual and textual inputs, and no model parameters are updated throughout our experiments.

All datasets used in this work are publicly available and can be accessed online.

14. Layer-wise Distribution Analysis

To provide a more holistic view of attention head behaviors, we analyze all attention heads in the last four layers.

Figure 4 shows the layer–head distribution of EC for ViT-B/16 and ViT-L/14.

15. Interfering Head Experiments

We provide additional quantitative results (Table 7) for interfering head identification.

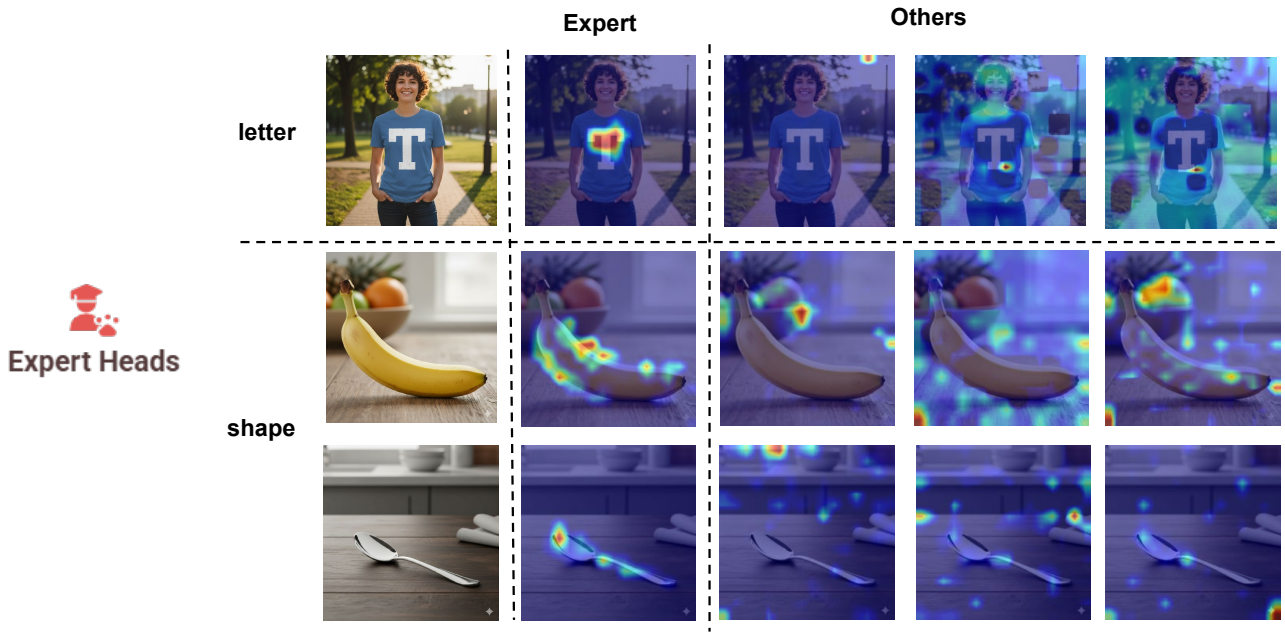

Figure 5. Visualization examples of *Expert Heads*. Each row shows representative image samples where the highlighted head exhibits strong and localized activation on task-relevant regions, compared with other heads. The first group ("letter") demonstrates heads focusing on textual or symbolic patterns, while the second ("shape") shows attention concentrated on object geometry. Other heads display diffuse or background-oriented activations, indicating weaker specialization.

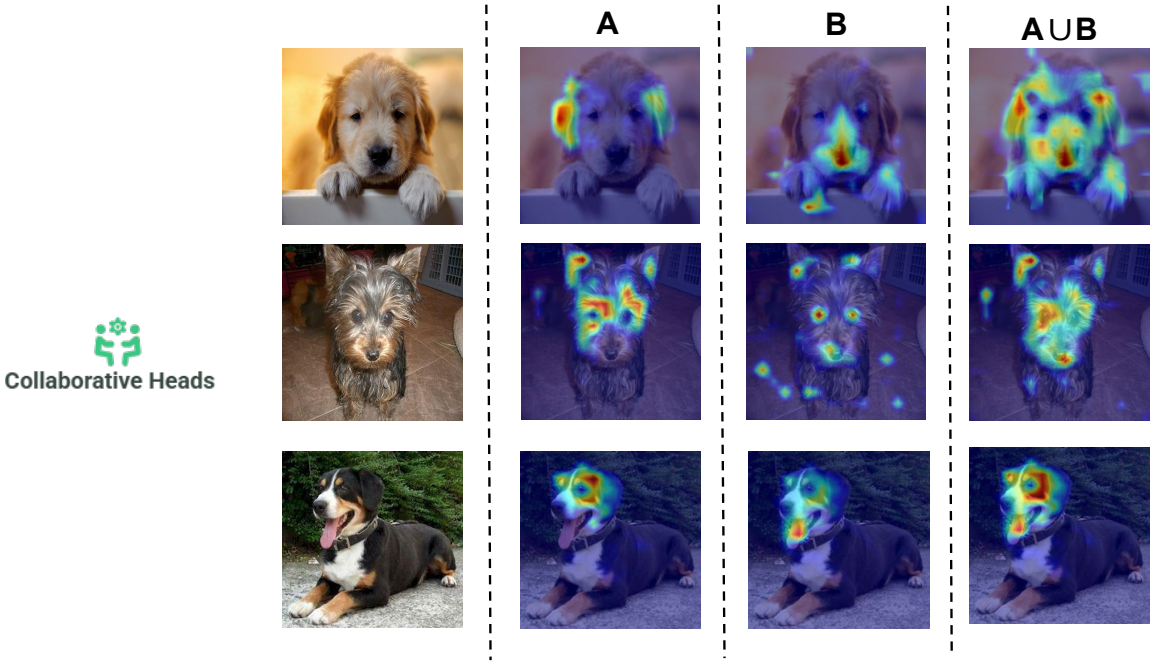

Figure 6. Visualization examples of *Collaborative Heads*. Each row presents representative image samples along with attention maps produced by two groups of heads (A and B) and their joint activation ($A \cup B$). The two groups typically attend to different yet complementary regions of the same object—for example, one group focuses on a dog's ears while the other emphasizes its eyes. When their activations are combined, the resulting attention becomes more complete and coherent, suggesting that such head groups provide complementary cues that jointly enhance the representation.

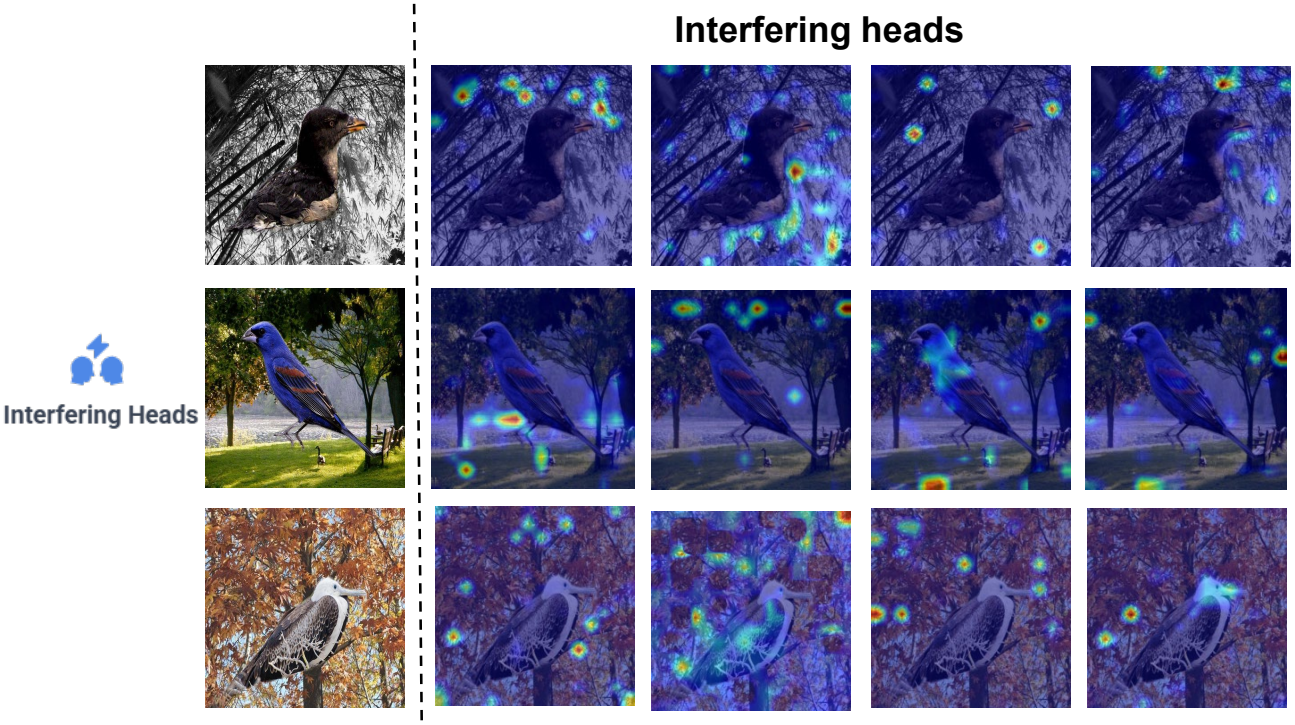

Figure 7. Visualization examples of *Interfering Heads*. Each row shows image samples where the highlighted heads attend to background or context regions that are unrelated to the main object, such as trees, grass, or surrounding textures. These heads tend to distribute attention diffusely over non-discriminative areas, introducing noise that may hinder correct classification.

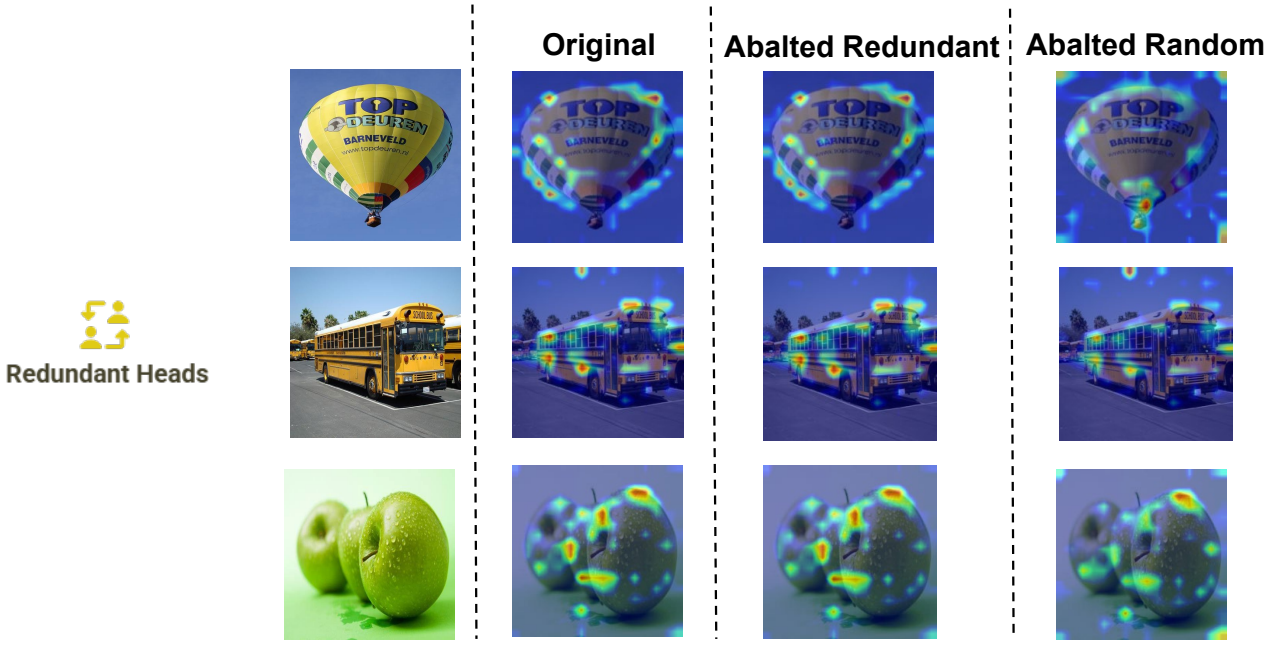

Figure 8. Visualization of redundant heads. Each row corresponds to one example image. The first column shows the original image, the second column shows the attention visualization from the original model, the third column shows the visualization after ablating redundant heads, and the fourth column shows the result after randomly ablating an equal number of heads. Ablating redundant heads leads to minimal changes in attention distribution, while random ablation introduces noticeable distortions, indicating that redundant heads contribute overlapping information.

Table 8. Representative head captions illustrating the semantic focus of different attention heads in ViT/B-16.

Layer	Head	Head Caption
8	0	The attention head is likely to be selective about objects or scenes with a vintage or antique aesthetic.
8	1	The attention head is highly selective to areas with cracked textures, regardless of the color tone.
8	2	The attention head is sensitive to textures, particularly the detailed and intricate textures of the eye's veins.
8	3	The attention head is selective towards images that combine elements of nature with human-made structures.
8	4	The attention head is highly selective towards the heart shape.
8	5	The attention head is highly selective towards images with vibrant and contrasting colors.
8	6	The attention head is strongly activated by elongated or distinct shapes such as sticks, bird silhouettes.
8	7	The attention head is selective about areas with high color contrast and vividness.
8	8	The attention head is highly selective towards images with a combination of bright and contrasting colors.
8	9	The attention head is selective about strong contrast, such as bright lights against a dark background.
8	10	The attention head is selective to circular shapes in various contexts, including different sizes, colors, and textures.
8	11	The attention head is highly sensitive to areas with strong contrast in lighting, such as dramatic sunrises/sunsets...
9	0	The head is selective towards elements that combine natural and artificial aspects
9	1	The attention head is selective about images with strong visual contrasts, whether it be in color, texture, or lighting.
9	2	The attention head is highly selective towards objects with rich and intricate textures, such as the fur of animals.
9	3	The attention head is consistently selective about areas with natural light and sunlight effects.
9	4	The head is selective about feature a combination of bright and vivid colors along with human subjects.
9	5	The attention head is likely to be selective about images that contain a clear focal point or subject.
9	6	The attention head is selective about vibrant and intense colors.
9	7	The attention head is selective about high contrast or brightness.
9	8	The attention head is likely selective to scenes with a high density of human activity and movement.
9	9	The attention head is selective about regions with high color contrast and vivid colors, such as the colorful of birds.
9	10	The attention head is highly selective for images containing animals, particularly birds.
9	11	The attention head is consistently drawn to elements that convey a sense of motion or speed.
10	0	The head is selective about objects with distinct shapes like circular, hexagonal, triangular, etc
10	1	The attention head is highly selective towards metallic colors like platinum silver and gold
10	2	The attention head is highly selective about animals, including mammals like monkeys and dolphins
10	3	The attention head is highly selective towards natural landscapes with expansive views
10	4	The attention head is highly selective towards objects with vibrant and varied colors
10	5	The attention head is sensitive to intricate and detailed patterns, such as lion's fur, the lace pattern.
10	6	The attention head is selective vibrant and contrasting colors.
10	7	The attention head is selective about soft and warm textures.
10	8	The attention head is selective towards images that contain vibrant and contrasting colors
10	9	The attention head is highly selective towards objects with a shiny or reflective surface
10	10	The attention head is highly selective towards scenes with a serene and tranquil atmosphere
10	11	The attention head is highly sensitive to vibrant and contrasting colors
11	0	The head is highly selective of natural landscapes with a sense of vastness and openness
11	1	The attention head is highly sensitive to areas with significant contrast in lighting and color
11	2	The attention head is selective about areas with high visual contrast, including dramatic lighting and shadow effects.
11	3	The attention head is likely selective about images with strong contrasts between light and dark areas
11	4	The attention head is highly selective towards human interactions and activities, including playing games...
11	5	The attention head is highly selective towards distinct animal patterns and textures such as the giraffe's spots...
11	6	The attention head is highly selective towards mountainous regions.
11	7	The attention head is selective about objects with reflective surfaces, such as the megaphone...
11	8	The attention head is highly selective towards elements with intricate textures...
11	9	The attention head is highly selective towards numbers...
11	10	The attention head is sensitive to areas with high color contrast...
11	11	The attention head is highly selective towards letters...

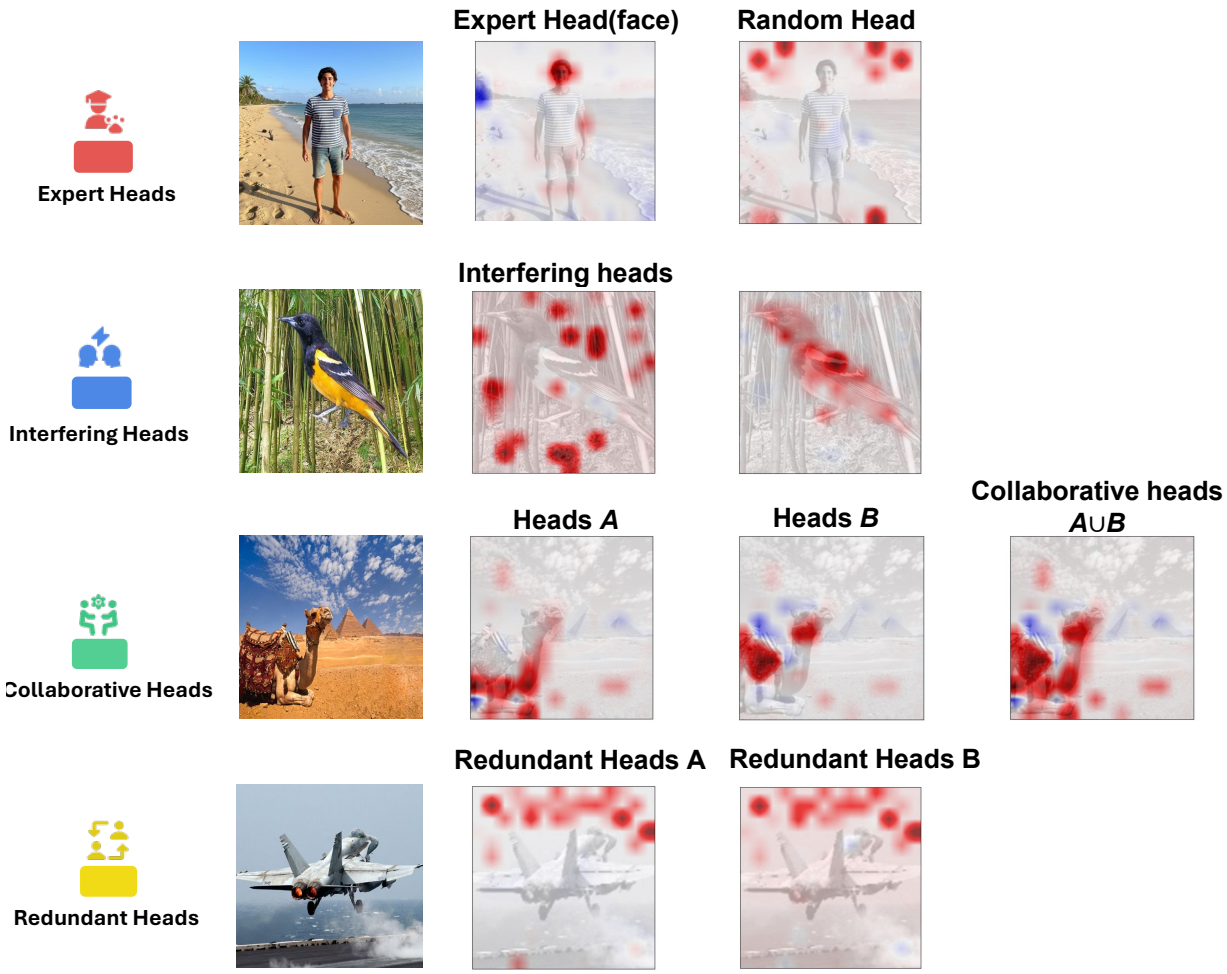

Figure 9. Visualization examples of Swin-B

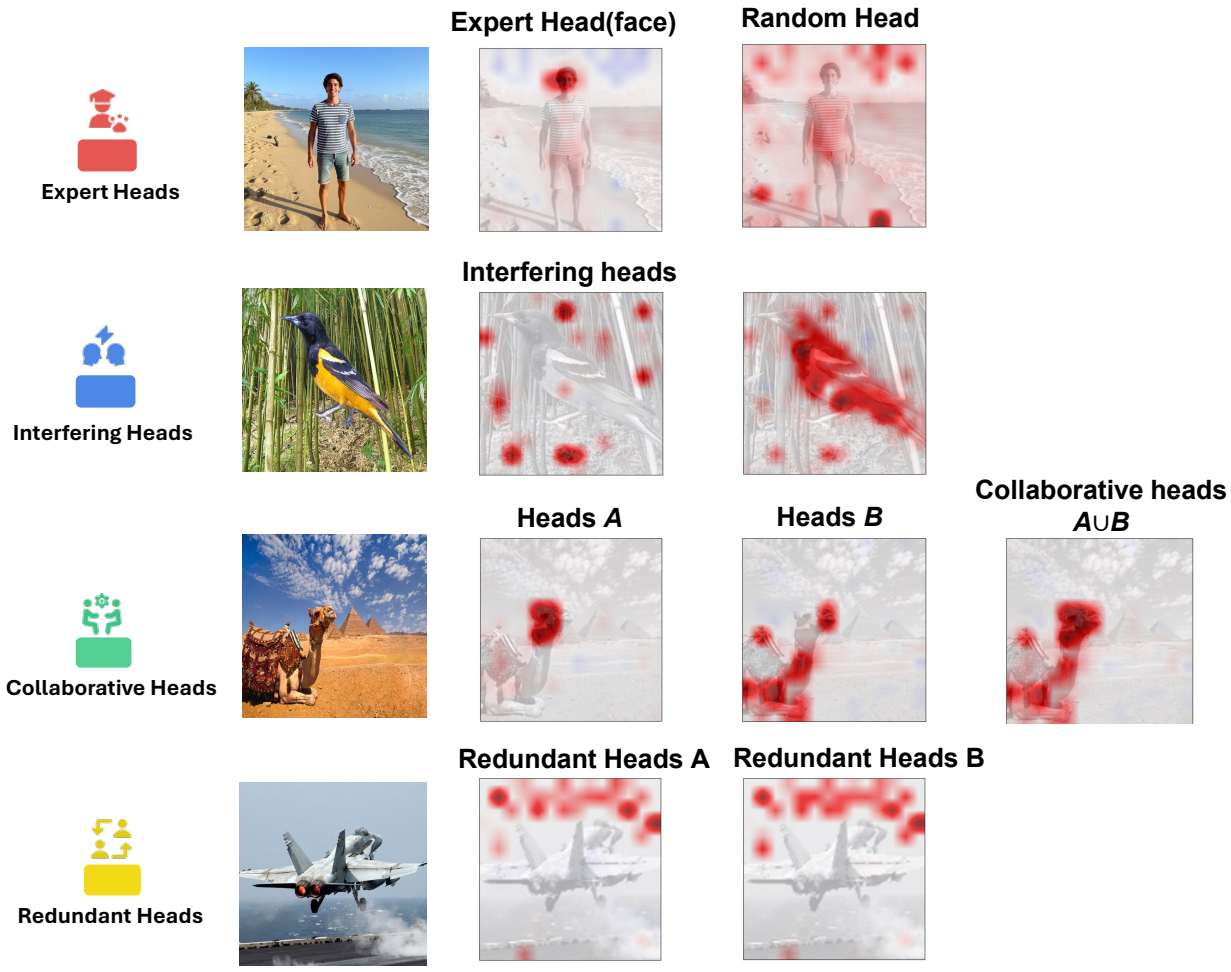

Figure 10. Visualization examples of DeiT-B

Table 9. Representative head captions illustrating the semantic focus of different attention heads in Swin-B.

Stage	Layer	Head	Head Caption
2	18	0	Prefers lively scenes with strong global color style and prominent foreground subjects in both nature and cities.
2	18	1	Focuses on human interactions and emotional portraits within diverse outdoor and urban environments.
2	18	2	Attends to mid scale subjects and linear structures in scenic contexts that mix human, animal.
2	18	3	Emphasizes textured close views and reflective rustic scenes that combine animals, people and patterned surfaces.
2	18	4	Prefers compact scenes with a salient central subject framed by nature, warm interiors or strong color accents.
2	18	5	Captures scenic views and silhouettes with strong tonal contrast across mountains, deserts, cities and performances.
2	18	6	Specializes in global color palettes and dense patterns with pronounced chromatic contrast and fine detail.
2	18	7	Focuses on animals and small objects highlighted by depth of field, bokeh and motion rich backgrounds.
2	18	8	Attends to graphic patterns and textures with text, repetitive motifs and minimal architectural forms.
2	18	9	Prefers elongated or sinuous forms and spiky textures set against vivid and high contrast natural scenes.
2	18	10	Emphasizes dynamic scenes and crowds under stylized lighting across deserts, coasts, gardens and ornate spaces.
2	18	11	Integrates natural, architectural and market scenes with moderate lighting and rich environmental detail.
2	18	12	Highlights salient foreground subjects in busy urban or commercial environments with vivid fine detail.
2	18	13	Focuses on expressive poses or focal figures framed by strong composition and atmospheric lighting.
2	18	14	Prefers calm human and landscape scenes across diverse locations with a cohesive tonal atmosphere.
2	18	15	Captures broad scenic views where mood and lighting dominate across natural landscapes and urban skylines.
3	1	0	Prefers salient foreground objects with rich texture.
3	1	1	Focuses on animals and human presence embedded in streets, paths, weathered facades.
3	1	2	Emphasizes scenes with strong global color and lighting contrast.
3	1	3	Attends to detailed textured surfaces with curved or swirling patterns over animals, urban structures.
3	1	4	Combines urban views, human figures and small natural elements within scenes.
3	1	5	Sensitive to fine linear and feather like textures, crystalline structures and ornate decorative.
3	1	6	Highlights warm interiors and rustic outdoor scenes, often with animals, aged stone.
3	1	7	Tracks human scale objects and figures in warm domestic and natural settings.
3	1	8	Captures overall ambience of busy streets, forests, coasts and cafes.
3	1	9	Encodes tonal style and illumination, favoring high contrast scenes with ancient stone, snow.
3	1	10	Prefers central symmetric or circular structures, quilted and floral textures.
3	1	11	Attends to detailed natural subjects, technology displays, aged stone and markets.
3	1	12	Focuses on manmade structures with intricate mechanisms, neon lighting, shattered glass and handwritten.
3	1	13	Encodes large geometric or object like forms under stylized tonal treatments.
3	1	14	Prefers graphic and typographic compositions with minimalist lines, mosaic.
3	1	15	Specializes in global color mood and night city atmospheres.
3	2	0	Prefers mid-scale subjects with pronounced shape and texture embedded in structured environments.
3	2	1	Sensitive to connected structures and interactions, combining human gestures, expressions.
3	2	2	Focuses on intimate social situations with a salient central figure or small group in expressive surroundings.
3	2	3	Emphasizes global lighting and simple geometry, capturing paths, silhouettes and warm tonal composition.
3	2	4	Texture oriented, highlighting material surfaces and minimalist layouts with a strong stylistic rendering.
3	2	5	Responds to strong perspective and framing cues that place a distinct subject in a dramatic spatial context.
3	2	6	Pattern and symbol selective, attending to text, graphic motifs and high frequency structured details.
3	2	7	Encodes clean daylight compositions where a clear subject is framed by architectural or natural structure.
3	2	8	Balances soft textures and broad environments, often with a gentle motion or vintage styled appearance.
3	2	9	Focuses on rigid man made forms and objects with stylized tonal treatment and reflective city ambience.
3	2	10	Selective to elongated or linear focal elements integrated into scenic views with a nostalgic mood.
3	2	11	Highlights rich textures of living beings and crafted objects in warm.
3	2	12	Mood driven, encoding scene wide illumination from dim interiors to bright natural and concert lighting.
3	2	13	Oriented toward journey and exploration themes, following paths, rugged terrain and adventurous scenes.
3	2	14	Social and narrative focused, combining expressive figures with symbolic or technical visual elements.
3	2	15	Strongly color selective, tracking lush natural scenes and festivities through global color palette and atmosphere.

Table 10. Representative head captions illustrating the semantic focus of different attention heads in DeiT-B.

Layer	Head	Head Caption
10	0	The attention head is selective towards outdoor landscapes and open environments.
10	1	The attention head is selective towards scenes where lighting and atmosphere dominate.
10	2	The attention head is selective towards human portraits and scenes characterized by distinct color palettes.
10	3	The attention head is selective towards wide open outdoor scenes that span both natural environments.
10	4	The attention head is selective towards human faces, small social groups and scenes with warm .
10	5	The attention head is selective towards emotionally expressive scenes that combine animals, human contact.
10	6	The attention head is selective towards reflective or curved structures, textured materials.
10	7	The attention head is selective towards man made objects together with relaxed human or animal subjects.
10	8	The attention head is selective towards strong graphic patterns.
10	9	The attention head is selective towards close range views with rich color.
10	10	The attention head is selective towards linear or striped textures together with dynamic subjects.
10	11	The attention head is selective towards elongated fine structures, expressive subjects.
11	0	The attention head is highly selective towards images containing animals and detailed natural textures.
11	1	The attention head is selective towards scenes with animal or human activity in rural and urban environments.
11	2	The attention head is highly selective towards human figures and facial expressions.
11	3	The attention head is selective towards warm color tones.
11	4	The attention head is highly selective towards salient living subjects and dynamic social scenes.
11	5	The attention head is selective towards combinations of animals, expansive natural landscapes and rustic.
11	6	The attention head is highly sensitive to rounded shapes and swirling structures.
11	7	The attention head is selective towards overall scene ambience.
11	8	The attention head is selective towards salient local objects with curved forms and fine detail.
11	9	The attention head is highly selective towards human made structures and technological objects.
11	10	The attention head is selectively driven by color palette and lighting.
11	11	The attention head is selective towards scenes that combine elongated objects and human or outdoor activity.
12	1	The attention head is highly selective towards small natural objects and fine textures.
12	2	The attention head is selective towards man made containers and functional objects.
12	3	The attention head is selective towards images containing animals and mid scale objects.
12	4	The attention head is selective towards human faces, everyday objects.
12	5	The attention head is strongly activated by elongated or slender shapes and intricate decorative elements.
12	6	The attention head is selective towards human figures and social activities.
12	7	The attention head is highly selective towards wildlife and natural textures.
12	8	The attention head is selective about broad environmental scenes and dynamic events.
12	9	The attention head is selective towards salient local structures with strong shape contrast.
12	10	The attention head is highly selective towards overall scene categories and spatial layouts.
12	11	The attention head is selective towards man made patterns and atmospheric scenes.