
Register and [CLS] tokens induce a decoupling of local and global features in large ViTs

Alexander Lappe^{1,2} Martin A. Giese¹

¹Hertie Institute, University Clinics Tübingen ²IMPRS-IS
alexander.lappe@uni-tuebingen.de

Abstract

Recent work has shown that the attention maps of the widely popular DINOv2 model exhibit artifacts, which hurt both model interpretability and performance on dense image tasks. These artifacts emerge due to the model repurposing patch tokens with redundant local information for the storage of global image information. To address this problem, additional register tokens have been incorporated in which the model can store such information instead. We carefully examine the influence of these register tokens on the relationship between global and local image features, showing that while register tokens yield cleaner attention maps, these maps do not accurately reflect the integration of local image information in large models. Instead, global information is dominated by information extracted from register tokens, leading to a disconnect between local and global features. Inspired by these findings, we show that the [CLS] token itself leads to a very similar phenomenon in models without explicit register tokens. Our work shows that care must be taken when interpreting attention maps of large ViTs. Further, by clearly attributing the faulty behavior to register and [CLS] tokens, we show a path towards more interpretable vision models.

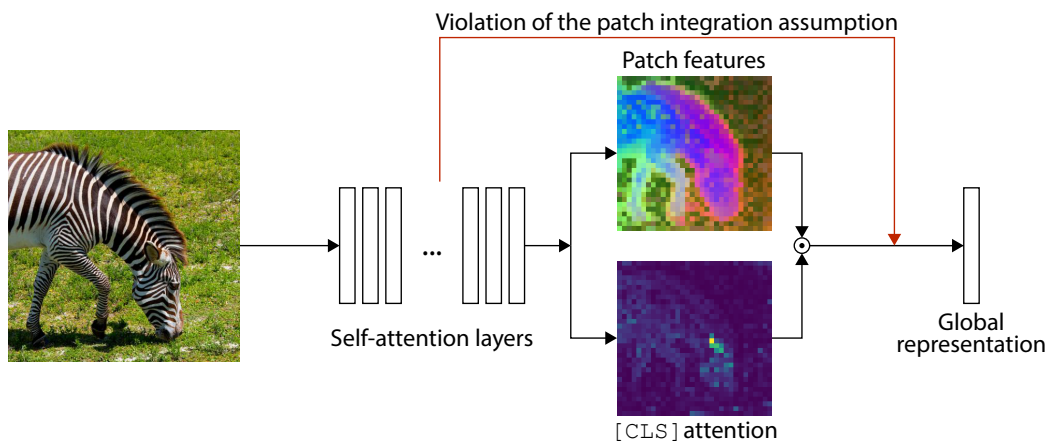


Figure 1: The global image representation computed by a Vision Transformer is usually understood as a weighted average of the local patch features, where weights are given by the [CLS] token attention scores. This notion, which we refer to as the *patch integration assumption*, underlies attribution methods like plotting the attention maps to identify patches that strongly contribute to the global output. In this work, we show that both register tokens and the [CLS] token lead to violations of this assumption in large ViT models.

1 Introduction

In recent years, the application of the Transformer architecture [1] to computer vision [2] has given rise to powerful, highly general feature extractors which can be used for numerous downstream vision tasks. These models are pretrained on large datasets in either supervised [3, 4, 5, 6] or self-supervised [7, 8, 9, 10] fashion, and then fine-tuned or used directly as off-the-shelf feature extractors. A useful property of the Vision Transformer architecture is that it simultaneously extracts global and dense local features, making it suitable for both classification, as well as more fine-grained tasks such as object detection [11], segmentation [12] or depth estimation [10, 13]. Further, the fact that the global image representation is computed as a convex combination of the dense features enables direct comparisons between patch and global representations [14, 15, 16]. The weights used for this convex combination are the attention outputs of the global [CLS] token to the patch tokens. Caron et al. [7] popularized the investigation of the attention scores as a way of attributing model outputs to image regions, as one can seemingly clearly identify the patches from which the global information is extracted. After the publication of the now widely used DINOv2 model [10], Darcet et al. [17] showed that the model’s attention maps exhibit artifacts with very large weights. Upon further investigation, they noted that these artifacts also appear in other models trained with different training strategies, and that the patch tokens corresponding to the artifacts contain global image information rather than information about the image patch as originally assumed. Since this phenomenon made the attention maps less interpretable and degraded the model’s performance on dense tasks, the authors proposed the introduction of additional register tokens in which global information can be stored, to prevent it from being encoded in the patch tokens. The inclusion of such register tokens achieved the desired effect of removing artifacts from the attention maps. Further, while there has been some debate on how informative attention maps are in general [18, 19], the resulting attention maps have also been argued to be more interpretable [17, 20]. Subsequently, further work proposed alternative ways of including register tokens in the model without repeating the expensive pretraining stage [20, 21].

However, the encoding of global information in the patch tokens points towards a deeper underlying problem. Even if this information is stored in register tokens instead, it is unclear whether the denoised attention maps over the patch tokens actually reflect the global model output faithfully. Since evidently some global information is integrated before the last layer and can flow from the register tokens to the [CLS] token, it must be studied to which degree the patch features still contribute to the final output. In this work, we show that in large models, the global image representation relies primarily on information extracted from the register tokens, which needs to be considered when studying the denoised attention maps. This effect does not occur in the smaller model variants, which display a tight correspondence between local and global features. Inspired by this finding, we examine whether a disregard of the local patch features can also occur in overparameterized models without explicit register tokens. We observe that the [CLS] token itself may be interpreted as a register and show that the implicit backward-attention mechanism introduced by the residual connection leads to the same effect.

Contributions. Our results show that the intuitive correspondence between global and local image representations holds in smaller variants of the ViT trained with DINOv2, but breaks down for larger models. Importantly, the last-layer attention maps do not faithfully represent the mechanism through which global image representations are formed in large DINOv2 models with register tokens or residual connections in the [CLS] token. These effects need to be taken into account when studying attention maps to attribute model outputs to image regions, or ground global feature representations in local patch features. Our findings suggest that future iterations of generalist vision models should be built without register tokens and residual connections in the [CLS] token to ensure model interpretability.

2 Preliminaries

Vision Transformers. Given an image $\mathbf{x} \in \mathbb{R}^{h \times w \times 3}$, a vanilla Vision Transformer divides it into small patches and feeds these patches through a network of self-attention layers. This process yields a feature map $\mathbf{z} \in \mathbb{R}^{p_1 \times p_2 \times d}$, where p_1 and p_2 denote the spatial dimensions and d the model’s feature dimension. To learn a global representation of the image, an additional [CLS] token is introduced, which is treated exactly like the image patch tokens. The output of the [CLS] token at the last layer corresponds to the global image embedding. Mathematically, the self-attention mechanism for a

single attention head works as follows: The [CLS] token outputs a query vector $\mathbf{q} \in \mathbb{R}^m$ which is compared to the keys output by each token denoted by $\mathbf{k}_1, \dots, \mathbf{k}_{p_1 p_2 + 1} \in \mathbb{R}^m$ to yield the attention vector

$$\mathbf{a} := \text{softmax}(\langle \mathbf{q}, \mathbf{k}_1 \rangle, \dots, \langle \mathbf{q}, \mathbf{k}_{p_1 p_2 + 1} \rangle). \quad (1)$$

The output of the attention mechanism for the [CLS] token is then given by the convex combination

$$\mathbf{o}_{cls} := \sum_i a_i \mathbf{v}_i, \quad (2)$$

where \mathbf{v}_i denotes the value vector of the i -th token. The final processing after the attention mechanism differs slightly between models. In the DINOv2 model we consider here, the attention is followed by a residual connection, layer normalization and a shallow multi-layer perceptron to yield the final output. As the residual connection is applied after self-attention and does not affect the features extracted via the attention map, we exclude it from the computations when comparing contributions of register tokens and patch tokens in Section 3. We treat the influence of the residual connection separately in Section 4.

The patch integration assumption. Vision Transformers are therefore usually understood as patch feature extractors, where the [CLS] token learns to select the final features from the most relevant patches for global image understanding. We refer to this notion as the *patch integration assumption*. Based on the fact that the [CLS] token undergoes the same transformations as the patch tokens and can be written as a weighted average thereof, the patch integration assumption has been used to attribute global model behavior to specific image regions [14, 15, 16]. In particular, the attention vector \mathbf{a} of the [CLS] token is a popular object of study to determine the patches that the model relies on for its global image representation [2, 7, 22, 23, 24]. A body of work on interpretability of ViTs has suggested alternative methods for attributing model behavior, going beyond studying the [CLS] attention scores [25, 26, 27, 28]. However, only Darcet et al. [17] challenged the *patch integration assumption* itself by observing the emergence of high-norm tokens in large models.

Transformers with register tokens. As mentioned previously, Darcet et al. [17] found that very large models trained with DINOv2 store global image information in patches that otherwise contain redundant information, degrading the quality of the dense patch features. As a remedy, they introduced additional register tokens that are appended to the patch tokens just as the [CLS] token. These are supposed to be used for storing global information to avoid the use of the patch tokens themselves for this purpose. In this case, the output of the [CLS] token becomes

$$\mathbf{o}_{[CLS]} = \underbrace{\sum_{i \in \{\text{patches}\}} a_i \mathbf{v}_i}_{\text{patch contribution}} + \underbrace{\sum_{j \in \{\text{registers}\}} a_j \mathbf{v}_j}_{\text{register contribution}}. \quad (3)$$

Decoupling of patch and global features. Darcet et al. [17] show that the resulting attention maps of DINOv2 models including registers are less noisy, and the resulting patch features show better performance on dense prediction tasks. However, it is possible that the inclusion of registers decouples the global image embedding from the local patches, as the [CLS] token is no longer a convex combination of only the local features. In case the [CLS] token attends primarily to the register tokens, the attention maps on the patches, while shown to be less noisy, might be uninformative due to dominance of non-local features. Further, if global information is obtained from the registers instead of integrating information from the patches, feature spaces between local and global representations might not be aligned, hurting the ability to precisely specify the local grounding of global information [14, 15, 29].

3 The influence of register tokens on the patch integration assumption

3.1 Larger models attend more to register tokens and less to patch tokens

High-norm patch tokens have so far only been observed in the larger variants of the ViT architecture. This raises the question of whether only larger models rely heavily on additional register tokens. To answer this question, we test how much attention the last-layer [CLS] token places on the register

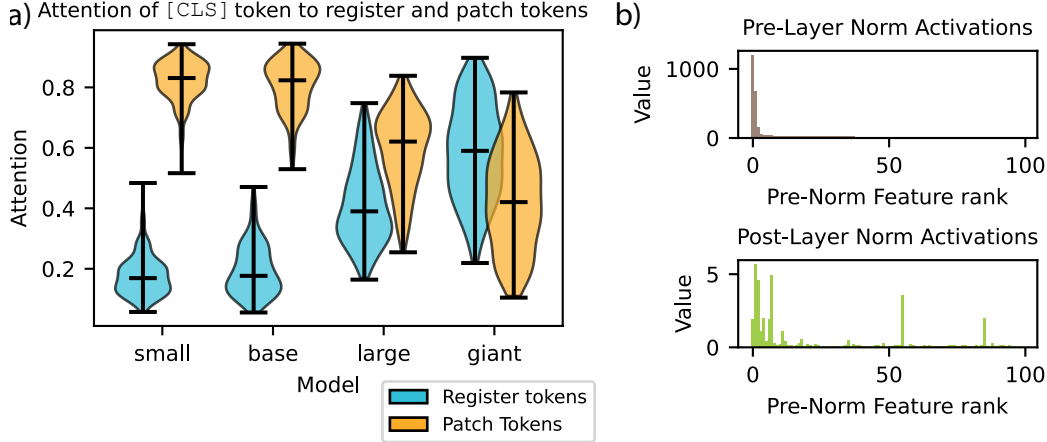


Figure 2: **a)** The amount of attention placed by the [CLS] token of the last layer onto the patch and register tokens, respectively. Smaller models attend primarily to the patch tokens, whereas bigger models attend more strongly to the register tokens. **b)** Mean activations of the highest-norm register token in the last layer of the 'giant' model. The 100 dimensions with highest activations before the layer norm are shown. Register tokens show large activations in a small subspace, making them seemingly image-independent as measured by pairwise cosine similarity (top panel). However, the layer norm downscales these dimensions *before* the self-attention mechanism (bottom panel).

and patch tokens, respectively. We study the DINOv2 models with register tokens as published by the authors on huggingface.com and probe them using the validation images of the MS COCO dataset [30]. We study the 'small', 'base', 'large' and 'giant' models, which differ in number of self-attention layers, hidden dimension and number of attention heads. They have 21M, 86M, 300M, and 1,100M total parameters, respectively. We extract the post-softmax attention vectors of the [CLS] token as presented in Eq. (3), average them across attention heads and then sum the resulting scalars for the register tokens and the patch tokens respectively, to examine how attention is partitioned between the two types of tokens. For each image, the sum of register attention and patch attention is one. Results are shown in Fig. 2.

We observe clearly that smaller models attend primarily to patch tokens, and the 'large' and 'giant' variants place more attention on the register tokens. Strikingly, the variance of preference for either register or patch tokens in the largest model is considerable, with some images attending mainly to patch tokens, and some almost ignoring them entirely.

3.2 Information in register tokens is only seemingly image independent

Image dependency of high-norm tokens. The role of information encoded by high-norm tokens in the original DINOv2 model is not entirely clear. Darcet et al. [17] trained linear probes on the high-norm tokens, concluding that they contain more global information and less positional information than the regular patch tokens. On the other hand, Wang et al. [31] claimed that high-norm tokens are image independent and can be predicted by the first singular vector of a linear approximation of the attention layer itself. They support these findings by noting that the pairwise cosine similarity between high-norm tokens of different images is extremely high. This conclusion raises the question of why the [CLS] token of the original DINOv2 model would be subject to image-independent information. By studying representations *within* the self-attention layer, we first reveal that these previous findings also apply to the register tokens, corroborating the assumption that they indeed take over the role of high-norm patch tokens. Further, we show that the previous results are not at all at odds, and that high-norm register tokens do carry image-dependent information.

Influence of layer norm. To understand in how far register tokens encode image-dependent information, we study their hidden states at the penultimate layer of the 'giant' model. The penultimate layer is the most relevant, since the final [CLS] output is formed by attending to these representations. First, we note that the hidden states of the register tokens do have abnormally high norm, and that the

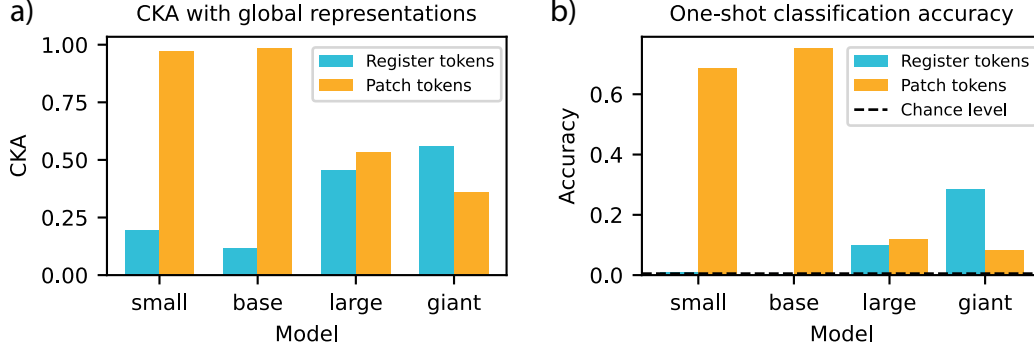


Figure 3: **a)** Centered kernel alignment between the global $[\text{CLS}]$ token output, and $[\text{CLS}]$ token output computed while only attending to either register tokens or patch tokens. Patch tokens yield a faithful representation of the global output for smaller models, but the connection between local and global features breaks down with increasing model size. **b)** One-shot classification accuracy on the 1000 Imagenet classes. The classifier is trained on the global $[\text{CLS}]$ output and then tested on output based on patch and register tokens, respectively. Attending only to the patch tokens yields poor performance in the larger models, corroborating the finding that global representations are not formed by attending to the patch features.

same token has the highest norm among all tokens for any input image by an order of magnitude. Second, we observe that the finding from Wang et al. [31], that activations are very similar across images, hold for the register tokens as well. The average pairwise cosine similarity between images in the highest-norm token is 0.9989, seemingly confirming that these tokens are image independent. However, studying the feature activations more closely shows that this need not be the case. Fig. 2 shows the top-100 feature dimensions of the highest-norm token (features with highest activation), averaged across all images. The plot clearly shows that the vector is dominated by extreme outliers which in turn dominate the cosine similarity. This leads us to two important points: First, note that the hidden state itself does not appear in the self-attention mechanism via a dot-product comparison. Instead, keys and values, which *are* compared via the dot product, are computed as linear maps of the hidden state, allowing the model to ignore these dimensions and focus on potentially image dependent dimensions instead. This mirrors exactly the methodological differences between [17] and [31], demonstrating why the authors came to opposite conclusions. Further, before entering the self-attention mechanism, the ViT applies a layer norm to all tokens, thereby reweighting the features and downscaling high-norm tokens. Fig. 2 shows the same data *after* applying the layer norm, showing that outlier dimensions are scaled down before computing self-attention.

3.3 Registers tokens dominate global image representations in large models

So far, we have shown that larger models attend strongly to register tokens, and we have discussed how register tokens can encode image-specific information. It remains to study how strongly the register tokens actually influence the output of the self-attention mechanism. After all, the information extracted by register tokens might be redundant w.r.t. to the patch tokens, or their value vectors could have low norm, leading to weak contribution despite large attention.

To examine the influence of patch tokens and register tokens on the $[\text{CLS}]$ output, we constrain the final self-attention layer to attend to only one of the two token types. Specifically, we set all attentions to one of the two token types to zero, yielding a $[\text{CLS}]$ output that is based only on the other. We also compute the unaltered model output based on all tokens and investigate the similarity of the resulting global representations using linear centered kernel alignment (CKA) [32]. Given two (mean-centered) matrices of neural network activations $X \in \mathbb{R}^{n \times d_1}$, $Y \in \mathbb{R}^{n \times d_2}$ computed on n test samples, linear CKA is defined as

$$CKA(X, Y) = \frac{\text{tr}(XX^\top YY^\top)}{\sqrt{\text{tr}(XX^\top XX^\top) \text{tr}(YY^\top YY^\top)}} \quad (4)$$

and can be interpreted as measuring the alignment of pairwise similarities between the two matrices. Fig. 3 a) shows the CKA between the unaltered output, and the patch-based and register-based output,

respectively. CKA was again computed using the samples from the MS COCO test set. We observe that for the two smaller models, the patch-based representation shows close-to-perfect alignment with the $[\text{CLS}]$ output. This indicates that the notion of the global output being an aggregate of last-layer local patch features is accurate. However, with increasing model size, the global output and the patch-based output become highly disconnected, and outputs based on register tokens represent the global output more faithfully in the giant model. The same analysis for the Imagenet test set [33] is presented in the appendix, mirroring the qualitative results. Fig. 3 b) displays one-shot top-5 classification accuracy when training the classifier on the global $[\text{CLS}]$ output and evaluating on either the patch-based or register-based output. Results were generated by randomly sampling one training image and one test image per class from the 1000 classes in the Imagenet test set. In accordance with the results from panel a), we observe that classifiers relying only on patch features perform poorly for the two larger models. This demonstrates that class-relevant global information is *not* extracted from the patch features.

3.4 Attention maps with registers are clean but do not reflect global image representations

The original purpose of including register tokens in the DINOv2 model was to remove high-norm tokens from the image patches to obtain better dense representations and cleaner attention maps. However, Fig. 3 casts doubt on how informative the cleaned-up attention maps are if the patch features extracted via the attention maps do not accurately represent the global model outputs. We show an example in Fig. 4. As expected, we observe that the attention map output by the giant model with registers is visually cleaner than the one output by the vanilla model. Next, we examine how faithfully the features extracted according to this attention map represent the total layer output by computing the cosine similarity between the extracted patch features, and the total layer output including register tokens. The cosine similarity is -0.0092, meaning that the features extracted according to the attention map are completely orthogonal to the features extracted by the $[\text{CLS}]$ token when including the register tokens. While not all images elicit disconnects of this magnitude, the example demonstrates the violation of the *patch integration assumption* in large models with registers, showing that its attention maps should not be relied on.

At the global level, let us consider the connection between the attention maps over patches and the CKA results shown in Fig. 3. The purpose of the $[\text{CLS}]$ token attention is to extract global image features that facilitate downstream tasks. Relevant for task performance is how images are embedded *relative* to each other, which, as is done in CKA, can be measured by the Gram matrix $\mathbf{X}\mathbf{X}^\top$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ contains the feature representations of a set of test samples. This Gram matrix encapsulates the similarity structure of learned image representations. Writing $\mathbf{X} = \mathbf{X}_p + \mathbf{X}_r$ as the sum of patch contributions and register contributions Eq. (3), we can decompose the Gram matrix as

$$\mathbf{X}\mathbf{X}^\top = \mathbf{X}_p\mathbf{X}_p^\top + \mathbf{X}_r\mathbf{X}_r^\top + \mathbf{X}_p\mathbf{X}_r^\top + \mathbf{X}_r\mathbf{X}_p^\top. \quad (5)$$

The attention maps only give insight into which patches are relevant for computing the patch-based representation and thus embedding an image into the representational geometry determined by $\mathbf{X}_p\mathbf{X}_p^\top$.

However, Fig. 3 shows that the geometries of $\mathbf{X}_p\mathbf{X}_p^\top$ and $\mathbf{X}\mathbf{X}^\top$ are misaligned, indicating that the attention map does not yield sufficient insight into how an image is embedded in the models’s global representational geometry.

3.5 Connections to overparameterization and neural collapse

The analysis of Darcet et al. [17] showed that high-norm patch tokens only appear in the larger variants of the ViT architecture. Our results on register tokens are similar, showing that both $[\text{CLS}]$ token attention to registers, as well as their influence on the global image representations increase substantially with model size. These findings relate to literature on how representations are formed in overparameterized models. Work on neural collapse [34] demonstrates that models with large capacity tend to learn simple, very low-dimensional representations at the last layer and that such representations yield better performance [35, 36, 37] and generalization abilities [34]. Further, it has been shown for overparameterized classification models, that low-intrinsic dimensionality as well as linear separability of classes emerge already before the final layer [38, 39]. Integrating our findings with this body of work, we hypothesize that larger ViT variants are overparameterized to the point at which DINOv2 training yields simplistic representations at the last layer, relying on a small number of tokens in which global information has already been integrated. The introduction of

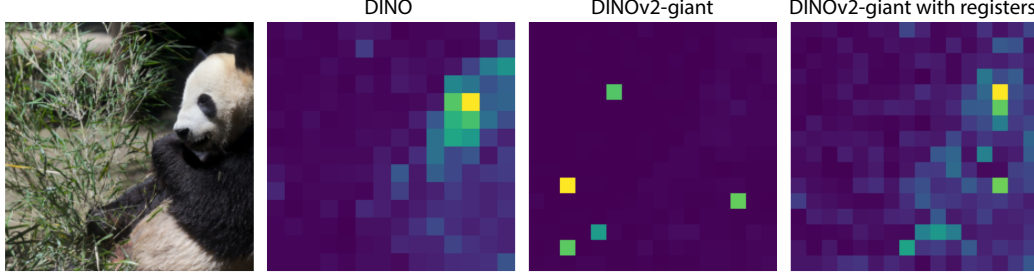


Figure 4: Attention maps of the final $[\text{CLS}]$ token. A convex combination of the corresponding patch features yields the patch-based contribution to the global image representation. As noted by Darcet et al. [17], the attention map of the DINOv2 model exhibits large artifacts. These are removed by including register tokens in the model, seemingly leading to a more interpretable attention map. However, when computing the $[\text{CLS}]$ output based on the convex combination of patch features in the model with registers, its cosine similarity to the total output of the final layer is -0.0092. In other words, attending to the patch tokens yields a representation completely orthogonal to the one including the register tokens, showing that the attention map fails to attribute global information to image patches.

register tokens merely shifts this mechanism outside of the patch features, preserving simple last-layer representations in the register tokens and thus still allowing the model to violate the *patch integration assumption*.

4 The influence of the $[\text{CLS}]$ token on the patch integration assumption

4.1 Very large models attend primarily to features from the skip connection

So far, we have discussed how register tokens lead to disconnects between local and global image representations and thereby the violation of the patch integration assumption. We have hypothesized that this behavior is due to overparameterization, allowing the model to solve its image-level training objective before the final layer, and storing the information in the register tokens. This line of reasoning motivated an inquiry into the mechanisms through which similar behavior could arise in overparameterized models without explicit register tokens. We have already discussed how the original DINOv2 model adapts patches containing little local information to instead store global information. In this section, we explore another mechanism that allows the model to ignore patch information at the last layer, namely residual connections within the $[\text{CLS}]$ token.

Attending to the previous layer. Previous work has shown that overparameterized networks may form more or less final representations of the input relatively early in the model hierarchy [38, 39]. Since the $[\text{CLS}]$ token in the ViT is appended to the patch tokens before the first layer, the standard ViT already contains one register token, which can be used to store emerging global information. Importantly, recent implementations of ViTs include skip connections within the attention block, resulting in the $[\text{CLS}]$ token output

$$\mathbf{o}_{[\text{CLS}]} = \sum_i a_i \mathbf{v}_i + \mathbf{o}_{[\text{CLS}]}^{\text{prev}} = \underbrace{\sum_{i \in \{\text{patches}\}} a_i \mathbf{v}_i}_{\text{patch contribution}} + \underbrace{a_{[\text{CLS}]} \mathbf{v}_{[\text{CLS}]} + \mathbf{o}_{[\text{CLS}]}^{\text{prev}}}_{\substack{\text{skip contribution} \\ \text{non-patch contribution}}}, \quad (6)$$

where $\mathbf{o}_{[\text{CLS}]}^{\text{prev}}$ denotes the output of the $[\text{CLS}]$ token at the previous layer, and $a_{[\text{CLS}]}$ and $\mathbf{v}_{[\text{CLS}]}$ denote the attention value to the $[\text{CLS}]$ token and its value vector, respectively. This allows the output to attend to the $[\text{CLS}]$ token via two different mechanisms: The first is part of the standard self-attention as given in equations (1) and (2). The second is given implicitly through the relative scales of the contributions of the attention mechanism and the skip connection to the output $\mathbf{o}_{[\text{CLS}]}$ as given in Eq. (6). These relative scales depend on the model parameters; therefore the model learns during training how strongly to attend to the skip connection.

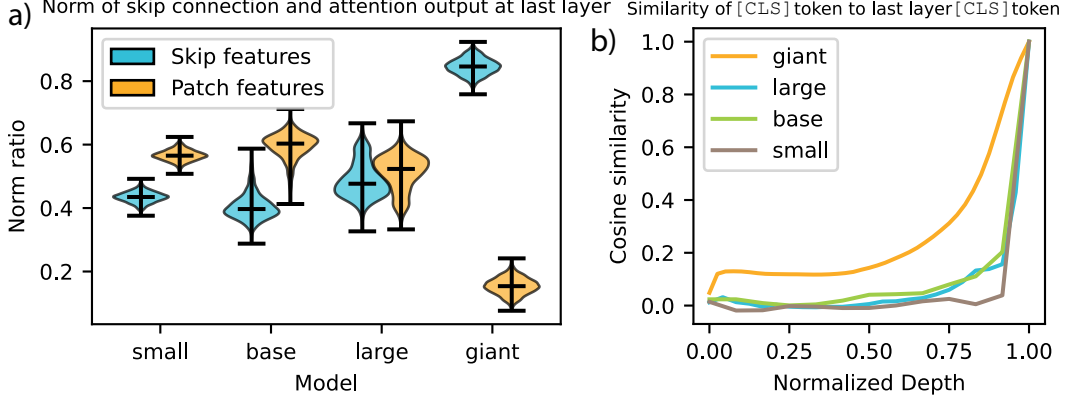


Figure 5: a) After the self-attention mechanism, a skip connection sums the attention output and the hidden states from the previous layer, providing an alternative way for the $[\text{CLS}]$ token to attend to itself. Since the attention weights are given only implicitly, we plot the L_2 -norm of the contributions of the skip connections and the patch features to the $[\text{CLS}]$ token as a proxy. We observe that in the ‘giant’ model, the global output is primarily determined by previously computed features, rather than the patch features. b) We show the cosine similarity of the $[\text{CLS}]$ token at all model layers to the $[\text{CLS}]$ token at the last layer. The three smaller models exhibit a large jump at the very last layer, indicating that the $[\text{CLS}]$ token at the last layer is strongly influenced by the patch tokens. Conversely, the $[\text{CLS}]$ token of the ‘giant’ model converges to the final output more smoothly, explaining its low reliance on the patch features at the last layer.

Measuring attention to patch features and skip features.

To examine how strongly models without register tokens attend to the skip connection, we study the original implementation of DINOv2 without registers, as well as DeiT III [6] for a supervised model, CLIP [3] for language supervision, and iBOT [9]. For readability we present findings for DINOv2 in the main text and show the generality of the findings using the other models in the Appendix. Since attention weights for the skip connection are not computed explicitly, we compute as a proxy the L_2 -norms of the patch contribution and the skip contribution from Eq. (6). The results of this analysis on the MS COCO validation set are shown in Fig. 5. We observe that the three smaller models attend to both patch and skip features, but the ‘giant’ model output is strongly dominated by the skip features. Panel b) sheds further light on this phenomenon: The similarity of the $[\text{CLS}]$ token of the last layer and all previous layers is low for the smaller models, indicating that the final $[\text{CLS}]$ output is determined by integrating the final patch features. This is in line with the patch integration assumption. However, the structure of the output of the ‘giant’ model emerges earlier, and the penultimate output is already very similar to the final one. Therefore, the last layer receives the quasi-final output through the skip connection and only needs to attend weakly to patch features to compute small corrections. We show that other ViT models trained with different recipes mimic this behavior in the Appendix (Fig. 7).

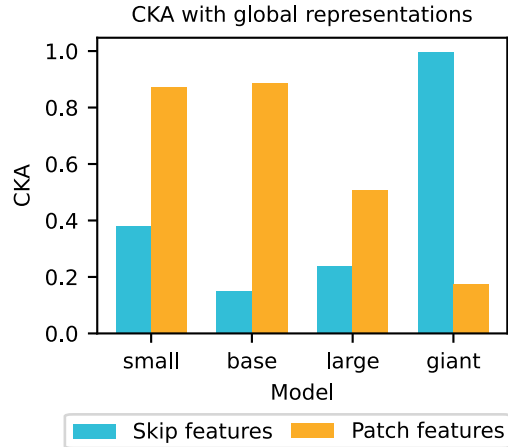


Figure 6: Centered kernel alignment between the standard $[\text{CLS}]$ token output, and $[\text{CLS}]$ token output computed using either only patch token features or the skip connection features from DINOv2. We observe that the findings from Fig. 5, that the giant model attends primarily to the skip connection, does indeed lead to disconnected representations between global features and patch features. This shows that the skip connection for the $[\text{CLS}]$ token can by itself lead to a violation of the patch integration assumption.

Earlier, we found for models with register tokens that large attention to registers leads to a disconnect between local and global image features. Does the same hold for large attention to skip features? To answer this question, we compute the centered kernel alignment between the global model output, and model output based only on patch or skip features, respectively. The results are shown in Fig. 6 (results for other models are presented in the Appendix, see Fig. 8). Again, we find that for the larger models, the features extracted from the patches do not yield a faithful representation of the global model output. We conclude that the skip connection of the `[CLS]` token indeed disconnects the global output from the patch tokens in the larger models.

5 Discussion

A tight correspondence between local and global image features computed by Vision Transformers is desirable for both interpretability and tasks combining local and global image information like object detection. For smaller DINOv2 models, including the widely used 'base' model, we found no evidence for disconnects between local and global features which is in line with previous findings regarding high-norm tokens [17]. On the other hand, we have observed that the global output of 'giant' model variants does not correspond to the final patch features extracted via the self-attention mechanism in models with registers, making their attention maps unreliable. Therefore, we hypothesize that these disconnects arise as a consequence of overparameterization, which gives models the flexibility to integrate global information already in intermediate layers. Combining results from Darcet et al. [17] with the work at hand, we identify three mechanisms through which the *patch integration assumption* may break down in overparameterized models. First, patch tokens can be repurposed to store prematurely emerging global information. Second, an approach of alleviating the first issue based on introducing register tokens seems to only move the problem outside of the patch tokens, still allowing the model to disregard patch features. Third, the skip connection in the self-attention layer enables the `[CLS]` token to extract image representations gradually, resulting in the fact that the last attention layer does not extract features that accurately represent the final output.

Future design choices. We have shown that while register tokens remove artifacts from the patch representations, they do not fully solve the problem of degenerate attention maps as they dominate the final attention-layer's output. This effect is particularly strong for the `[CLS]` token itself, which can be interpreted as a register token with an especially strong bias to attend to itself via the residual connection. We therefore argue that models that satisfy the *patch integration assumption* should be built without register tokens and hidden-layer `[CLS]` tokens. Another approach might be explicit regularization, although experiments on DeiT III show that drop-path regularization on its own does not prevent the violation of the *patch integration assumption*. In any case, it remains to ensure that global information is not stored in repurposed patch tokens. Here, approaches that impose direct regularization on the patch tokens appear to be the most promising. Along those lines, Wang et al. [31] proposed enforcing patch tokens to have similar representations as their neighbours, effectively removing high-norm patch tokens without the introduction of an additional mechanism for storing global information. The authors test their approach by finetuning on a smaller dataset and show that the resulting model performs well on dense tasks.

Limitations and future work. Our analysis focuses on diverging global encoding geometries between register tokens and patch tokens. Since the variance over samples with respect to the registers is high (see Fig. 2), future work could study the properties that lead an image to be processed primarily by patch/register tokens. As we have seen that centered kernel alignment between register tokens and patch tokens is low, a systematic analysis of the respective directions of variance in the representations could shed further light on the effects we have discussed. Since both attention to the skip connection and attention to explicit registers emerge in larger models, a more focused inquiry into the interdependence of these effects would be interesting. Finally, it remains to study the impact of violations of the *patch integration assumption* on performance on global image tasks. As mentioned in Section 3.5, previous work has shown a relationship between simplistic last-layer embeddings and model performance. A comparison of complexity-matched models that are forced to satisfy the *patch integration assumption* and models with registers is necessary for determining whether there is a tradeoff between image-level task performance and correspondence of local and global features.

Broader impacts. We do not foresee malicious use of our work, or tangible potential negative impact as a whole. On the contrary, we hope that our contribution will aid in better understanding of foundation vision models.

6 Acknowledgements

AL and MG are supported by ERC-SyG 856495. MG is supported by HFSP RGP0036/2016, BMBF FKZ 01GQ1704. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Alexander Lappe. Finally, we are grateful to Vojta Smekal, Marta Poyo Solanas, Prerana Kumar, and Lucas Martini for valuable discussions and support.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.
- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021.
- [4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR, July 2021.
- [5] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10347–10357. PMLR, July 2021.
- [6] Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13684, pages 516–533. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-20052-6 978-3-031-20053-3. doi: 10.1007/978-3-031-20053-3_30.
- [7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, October 2021. doi: 10.1109/ICCV48922.2021.00951.
- [8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, June 2022. doi: 10.1109/CVPR52688.2022.01553.
- [9] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer, January 2022.
- [10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, July 2023. ISSN 2835-8856.
- [11] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon W Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing Objects with Self-supervised Transformers and no Labels. In *Proceedings of the British Machine Vision Conference 2021*, page 365, Online, 2021. British Machine Vision Association. ISBN 978-1-901725-68-1. doi: 10.5244/C.35.365.

- [12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, April 2023.
- [13] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision Transformer Adapter for Dense Predictions. In *The Eleventh International Conference on Learning Representations*, September 2022.
- [14] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay Attention to Your Neighbours: Training-Free Open-Vocabulary Semantic Segmentation, September 2024.
- [15] Chong Zhou, Chen Change Loy, and Bo Dai. Extract Free Dense Labels from CLIP. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13688, pages 696–712. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-19814-4 978-3-031-19815-1. doi: 10.1007/978-3-031-19815-1_40.
- [16] Andrew F Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M Henderson, Leila Wehbe, and Michael J Tarr. BRAIN MAPPING WITH DENSE FEATURES: GROUND-ING CORTICAL SEMANTIC SELECTIVITY IN NATURAL IMAGES WITH VISION TRANSFORMERS. In *ICLR*, 2025.
- [17] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers, April 2024.
- [18] Sarthak Jain and Byron C. Wallace. Attention is not Explanation, May 2019.
- [19] Sarah Wiegrefe and Yuval Pinter. Attention is not not Explanation, September 2019.
- [20] Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision Transformers Don’t Need Trained Registers, June 2025.
- [21] Yinjie Chen, Zipeng Yan, Chong Zhou, Bo Dai, and Andrew F. Luo. Vision Transformers with Self-Distilled Registers, May 2025.
- [22] Lennart Brocki, Jakub Binda, and Neo Christopher Chung. Class-Discriminative Attention Maps for Vision Transformers. *Transactions on Machine Learning Research*, August 2024. ISSN 2835-8856.
- [23] Yaodong Yu, Tianzhe Chu, Shengbang Tong, Ziyang Wu, Druv Pai, Sam Buchanan, and Yi Ma. Emergence of Segmentation with Minimalistic White-Box Transformers, August 2023.
- [24] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching Matters: Investigating the Role of Supervision in Vision Transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7486–7496, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.00723.
- [25] Bill Psomas, Ioannis Kakogeorgiou, Konstantinos Karantzas, and Yannis Avrithis. Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, September 2023.
- [26] Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00084.
- [27] Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L Zhang. ViT-CX: Causal Explanation of Vision Transformers.
- [28] Alexandre Englebert, Sédrick Stassin, Géraldine Nanack, Sidi Ahmed Mahmoudi, Xavier Siebert, Olivier Cornu, and Christophe De Vleeschouwer. Explaining through Transformer Input Sampling. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 806–815, Paris, France, October 2023. IEEE. ISBN 9798350307443. doi: 10.1109/ICCVW60793.2023.00088.
- [29] Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking Self-Attention for Dense Vision-Language Inference. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15079, pages 315–332. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-72663-7 978-3-031-72664-4. doi: 10.1007/978-3-031-72664-4_18.
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

- [31] Haoqi Wang, Tong Zhang, and Mathieu Salzmann. SINDER: Repairing the Singular Defects of DINOv2, July 2024.
- [32] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR, May 2019.
- [33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.
- [34] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, October 2020. doi: 10.1073/pnas.2015509117.
- [35] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [36] Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How Far Pre-trained Models Are from Neural Collapse on the Target Dataset Informs their Transferability. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5526–5535, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.00511.
- [37] Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse Engineering Self-Supervised Learning.
- [38] Wojciech Masarczyk, Mateusz Ostaszewski, Ehsan Imani, Razvan Pascanu, and Piotr Miłoś. The Tunnel Effect: Building Data Representations in Deep Neural Networks.
- [39] Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A. Poggio. Feature learning in deep classifiers through Intermediate Neural Collapse. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28729–28745. PMLR, July 2023.

A Supplementary Results

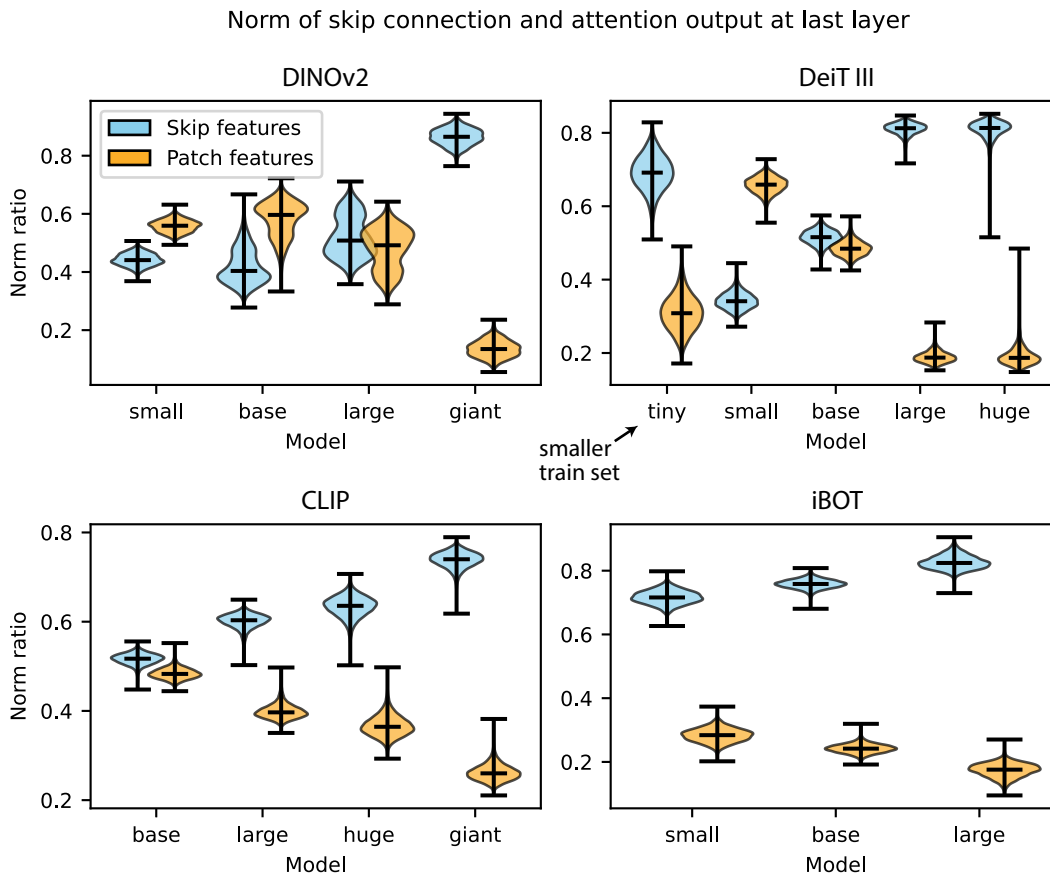


Figure 7: Attention to the $[CLS]$ token as discussed in Section 4, for Vision Transformers trained with different recipes. Note that attention to the skip connection consistently grows with model size. For DeiT III, we also include the tiny model, which was trained on a smaller data set. Interestingly, the tiny model mimics the behavior of the huge one trained on more data, demonstrating the importance of the the relationship between model capacity and data set size as discussed in the main text.

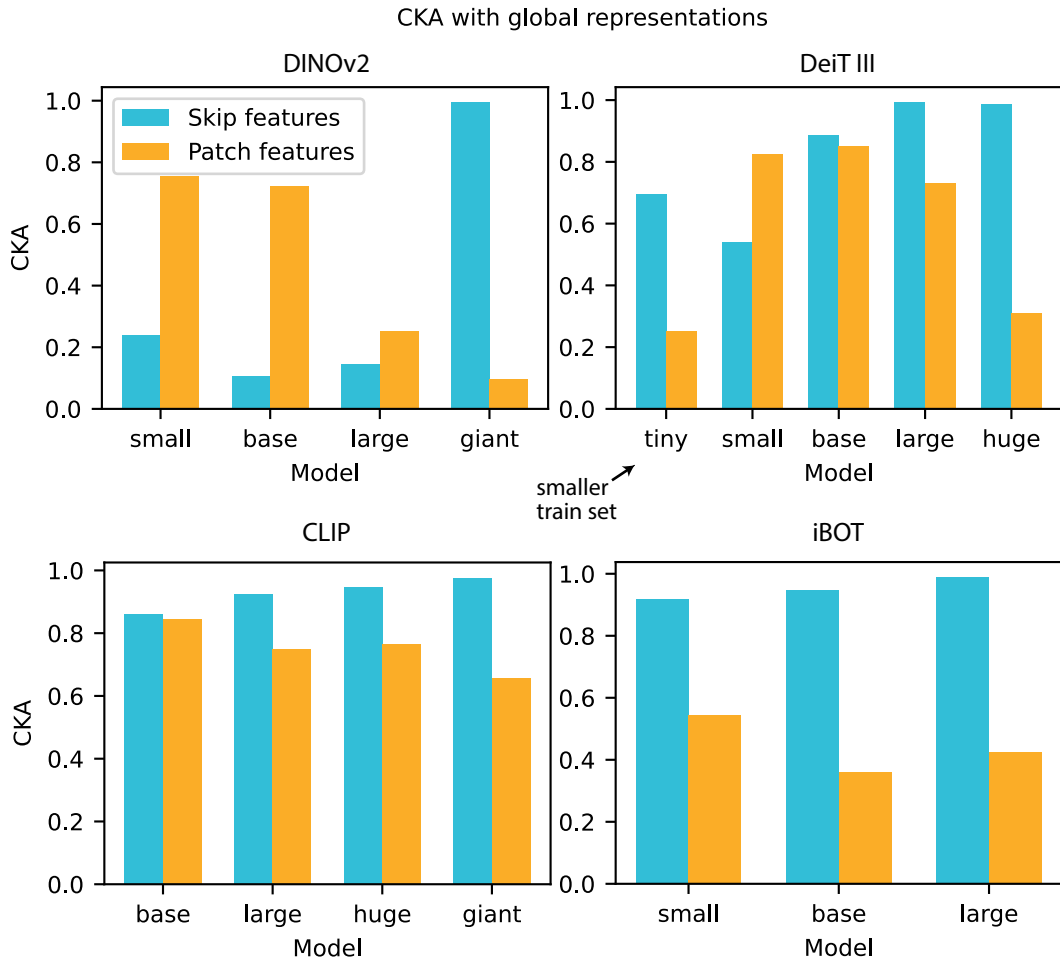


Figure 8: CKA results for the skip connection as discussed in Section 4 for different pretraining recipes. A disconnect between the local and global image features as discussed in the main text occurs across all pretraining recipes with growing model size.

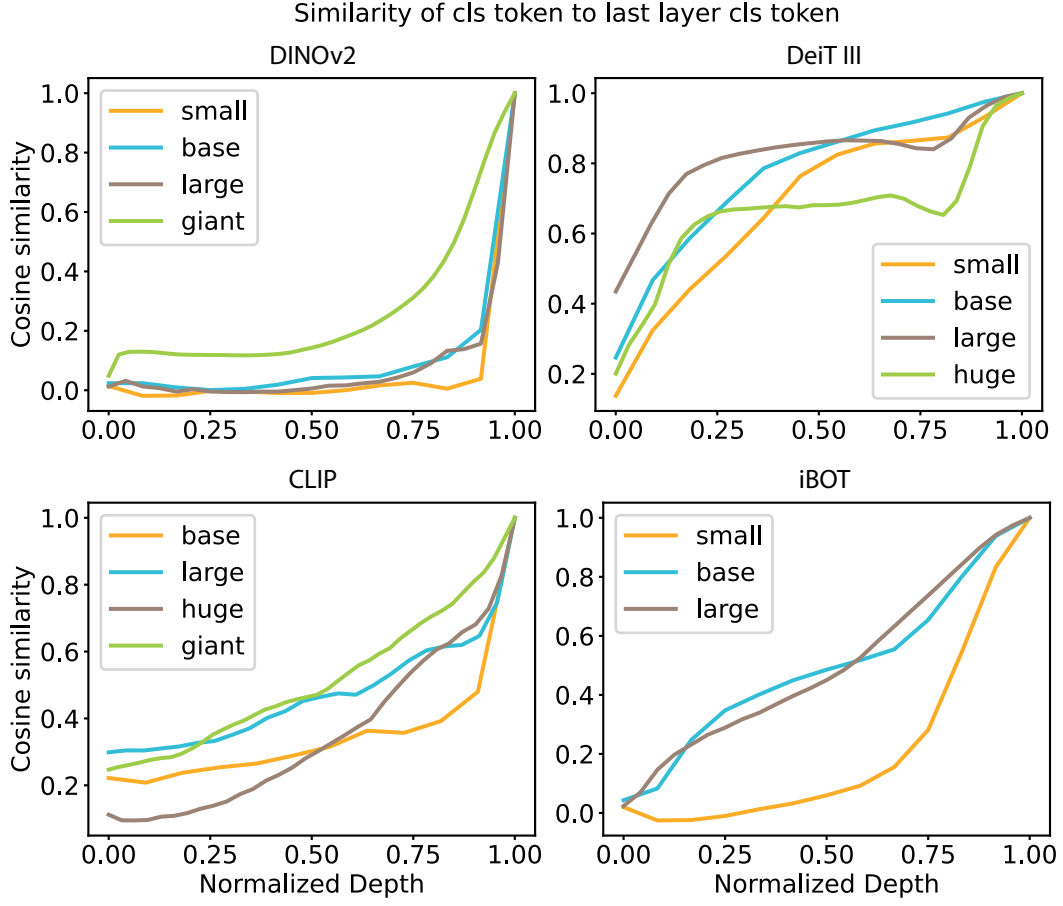


Figure 9: Layerwise similarity between the $[\text{CLS}]$ token and the last-layer $[\text{CLS}]$ token as discussed in Section 4. CLIP and iBOT mirror the DINOv2 behavior, whereas DeiT III does not exhibit a clear relationship between model size and early $[\text{CLS}]$ integration. Note that DeiT III is the only model trained with drop-path regularization, which may explain why all model sizes integrate final output information early in the hierarchy.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Claims in the abstract and introduction are that including register tokens in ViTs detaches the [CLS] output from the patch features, which is shown in the experiments.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We include a dedicated section on limitations of our work indicating deeper inquiries which are not part of the paper, as well as the necessity for further, computationally expensive, experiments on models without register tokens.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theoretical assumptions are necessary to support the claims made in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All models used in the analysis are publicly accessible, and we describe where to download them in the main text. We believe that experimental details are sufficient to reproduce the results. Additionally, we provide code that reproduces the paper's results and is inexpensive to run.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide scripts to automatically download the models used and run all analysis on an image dataset of choice.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Datasets and models are described in the main text, training details are not applicable to our work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: When presenting strength of attention to register tokens, we report the entire distribution so that the reader can appreciate the high variance. We do not report statistical significance or error bars for our experiments as there is no sampling aspect, except for the one-shot-classification experiment.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: We don't explicitly state compute times/resources as we believe it is obvious to the field that the experiments shown are inexpensive to run.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We are not aware of any ethics violations regarding this work.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We include a short section dedicated to this question.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not foresee any malicious applications of our work.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We clearly indicate the authors of models and datasets used for our experiments.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[NA\]](#)

Justification: We do not provide new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: There are no experiments including human subjects in this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: There are no experiments including human subjects in this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: There is no non-standard contribution of any LLM to this work.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.