A. Motivation

Modern visual representation models are built upon the attention mechanism inspired by biological vision systems. One drawback of it is the lack of a clear definition of the relationship between biological electrical signals and brain activity (energy). This drives us to break through the attention mechanism and attempt other physical laws. Heat conduction is a physical phenomenon in nature, characterized by the propagation of energy. The heat conduction process combines implicit attention computation with energy computation and has the potential to be a new mechanism for visual representation models.

B. HCO implementation using $\mathbf{DCT_{2D}}$ and $\mathbf{IDCT_{2D}}$

Assume a matrix denoted as $\mathbf{A}$ and the transformed matrix denoted as $\mathbf{B}$, the $\mathbf{DCT_{2D}}$ and the $\mathbf{IDCT_{2D}}$ can be performed by

$$\begin{aligned}
\mathbf{DCT_{2D}}: \mathbf{B}_{pq} &= \alpha_{\mathbf{p}}\alpha_{\mathbf{q}} \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \mathbf{A}_{mn}\mathbf{cos}\frac{(2m+1)p\pi}{2M}\mathbf{cos}\frac{(2n+1)q\pi}{2N}, \\
\mathbf{IDCT_{2D}}: \mathbf{A}_{mn} &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \alpha_{\mathbf{p}}\alpha_{\mathbf{q}}\mathbf{B}_{pq}\mathbf{cos}\frac{(2m+1)p\pi}{2M}\mathbf{cos}\frac{(2n+1)q\pi}{2N},
\end{aligned} \tag{10}$$

where $0{\leq}\{p,m\}{\leq}M-1$, $0{\leq}\{q,n\}{\leq}N-1$, $\alpha_{\mathbf{p}} = \begin{cases} \frac{1}{\sqrt{M}}, p=0 \\ \frac{2}{\sqrt{M}}, p>0 \end{cases}$, and $\alpha_{\mathbf{q}} = \begin{cases} \frac{1}{\sqrt{N}}, q=0 \\ \frac{2}{\sqrt{N}}, q>0 \end{cases}$. M and N respectively denote the row and column sizes of $\mathbf{A}$. Considering the matrix multiplication is GPU-friendly, we implement the $\mathbf{DCT_{2D}}$ and $\mathbf{IDCT_{2D}}$ in Eq. (10) by

$$\begin{aligned}
\mathbf{C} = (\mathbf{C}_{mp})_{M\times M} &= \left(\alpha_{\mathbf{p}}\mathbf{cos}\frac{(2m+1)p\pi}{2M}\right)_{M\times M}, \\
\mathbf{D} = (\mathbf{D}_{nq})_{N\times N} &= \left(\alpha_{\mathbf{q}}\mathbf{cos}\frac{(2n+1)q\pi}{2N}\right)_{N\times N}, \\
\mathbf{B} &= \mathbf{CAD^T}, \\
\mathbf{A} &= \mathbf{C^TBD}.
\end{aligned} \tag{11}$$

Suppose the number of total patches is N and the image is square, the shapes of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are all $\sqrt{N}\times\sqrt{N}$, which illustrates the computational complexity of (11) and HCO is $O(N^{1.5})$.

We compared our implementation of DCT/IDCT in vHeat with Torch-DCT, which is implemented based on $torch.fft$. Our implemented vHeat-B (661 img/s) is much faster than Torch-DCT (367 img/s), validating that our implemented GPU-friendly matrix multiplication is significantly efficient.

C. Experimental Settings

Model configurations. The configurations of vHeat-T/S/B models are shown in Table 8. The FLOPs and training parameters are reported after reparameterization in HCOs.

Image Classification. Following the standard evaluation protocol used in [38], all vHeat series are trained from scratch for 300 epochs and warmed up for the first 20 epochs. We utilize the AdamW optimizer [40] during the training process with betas set to $(0.9, 0.999)$, a momentum of 0.9, a cosine decay learning rate scheduler, an initial learning rate of 2×10^{-3}, a weight decay of 0.08, and a batch size of 2048. The drop path rates are set to $0.1/0.3/0.5$ for vHeat-T/S/B, respectively. Other techniques such as label smoothing (0.1) and exponential moving average (EMA) are also applied. No further training techniques are employed beyond these for a fair comparison. The training of vHeat-T/S/B takes 4.5/7/8.5 minutes per epoch on Tesla 16×V100 GPUs.

Object Detection. Following the settings in Swin [38] with the Mask-RCNN detector, we build the vHeat-based detector using the MMDetection library [7]. The AdamW optimizer [40] with a batch size of 16 is used to train the detector. The initial learning rate is set to 1×10^{-4} and is reduced by a factor of $10\times$ at the 9th and 11th epoch. The fine-tune process takes 12 ($1\times$) or 36 ($3\times$) epochs. We employ the multi-scale training and random flip technique, which aligns with the established practices for object detection evaluations.

Table 8. Configurations of vHeat. The contents in the tuples represent configurations for four stages.

Size	Tiny	Small	Base
Stem	3×3 conv with stride 2; Norm; GELU; 3×3 conv with stride 2; Norm		
Downsampling	3×3 conv with stride 2; Norm		
MLP ratio	4		
Classifier head	Global average pooling, Norm, MLP		
Layers	(2,2,6,2) (classification) (2,2,5,2) (others)	(2,2,18,2) (classification) (2,2,16,2) (others)	(2,2,18,2) (segmentation) (4,4,20,4) (others)
Channels	(96,192,384,768)	(96,192,384,768)	(128,256,512,1024) (segmentation) (96,192,384,768) (others)

Semantic Segmentation. Following the setting of Swin Transfomer [37], we construct a UperHead [72] on top of the pre-trained vHeat model to test its capability for semantic segmentation. The AdamW optimizer [40] is employed and the learning rate is set to 6×10^{-5} with a batch size of 16. The fine-tuning process takes a total of standard $160k$ iterations and the default input resolution is 512×512.

D. Additional Ablation Studies

Table 9. Evaluating different methods to align the shape of FVEs/k when loading ImageNet-1K pre-trained vHeat-B weights for detection and segmentation on COCO.

Method	AP^b	AP^m
Interpolating FVEs to predict k	47.4	42.9
Adding 0 to FVEs	47.4	42.7
Adding 0, then interpolating FVEs	**47.7**	**43.0**
Interpolating the predicted k	47.2	42.7

D.1. Interpolation of FVEs/k for downstream tasks

We have tried several approaches to align the shape for ablation. (1) Directly interpolate FVEs to the target shape of the input image. (2) Add 0 to the lower right region of FVEs to align the target shape. (3) Add 0 to the lower right region of FVEs to 512×512, and interpolate to the target shape. (4) Directly interpolate the predicted thermal diffusivity k to the target shape. The results are summarized in Table 9. Through the comparison, we select adding 0, then interpolating FVEs to the target shape for all downstream tasks.

D.2. Plain vHeat model

We've tested the performance of plain vHeat-B on ImageNet-1K classification. Keeping the same as DeiT-B, plain vHeat-B has 12 HCO layers, 768 embedding channels and the patch size is set to 16. Results are shown in Table 10. The superiority of plain vHeat-B over DeiT-B also validates the effectiveness of vHeat model.

Table 10. Plain vHeat-B vs. DeiT-B on ImageNet-1K with 300 epochs supervised training.

Model	#Param.	FLOPs	Acc
DeiT-B	86M	17.5G	81.8
Plain vHeat-B	88M	16.9G	**82.6**

D.3. Depth-wise convolution

We conduct experiments to validate the performance improvement from DWConv. We replace depth-wise convolution with layer normalization for vHeat-B. Results are summarized in Table 11, and vHeat-B achieves 83.8% Top-1 accuracy on ImageNet-1K classification, 0.2% lower than with DWConv, which validates the main gains come from the proposed HCO. Besides, when k is fixed as a large value, e.g. $k = 10.0$, replacing DWConv with layer normalization causes a significant performance drop (-0.7% top-1 accuracy). The comparison validates predicting k by FVEs can effectively improve the robustness of vHeat.

Table 11. Ablation experiments of depth-wise convolution (DWConv).

Model	DWConv	Acc
vHeat-B	✓	84.0
vHeat-B	✗	83.8 (-0.2)
vHeat-B (fix k=10.0)	✓	83.6
vHeat-B (fix k=10.0)	✗	82.9 (-0.7)

E. Receptive Field Visualization

The Effective Receptive Field (ERF) [42] of an output unit denotes the region of input that contains elements with a non-negligible influence on that unit. In Fig. 8, ResNet, ConNeXT, and Swin have local ERF. DeiT [61] and vHeat exhibit global ERFs. The difference lies in that DeiT has a $\mathcal{O}(N^2)$ complexity while vHeat enjoys $\mathcal{O}(N^{1.5})$ complexity.

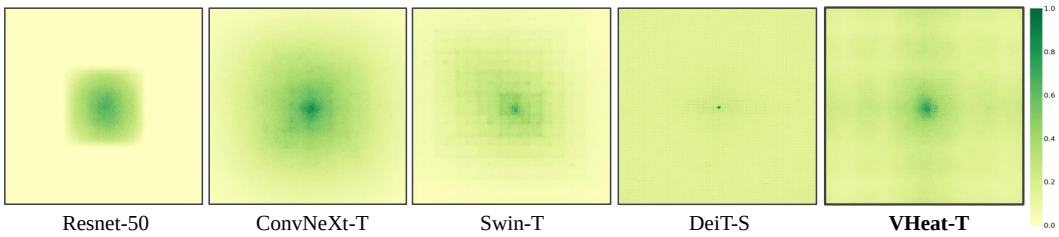

Figure 8. Visualization of the effective receptive fields (ERF) [42]. The visualization of baseline models are provided from VMamba [36]. Pixels of higher intensity indicate larger responses with the central pixel.

F. Heat Conduction Visualization

We visualize more instances of visual heat conduction, given a randomly selected patch as the heat source, Fig. 9, validating the self-adaptive visual heat conduction pattern through the prediction of k.

G. Analysis of k in each layer

We calculate average values of k in each layer of ImageNet-1K classification pre-trained vHeat-Tiny, Fig. 10. In stage 2 and stage 3, average values of k corresponding to deeper layers are larger, indicating that the visual heat conduction effect of deeper layers is stronger, leading to faster and farther overall content propagation.

H. Feature Map Visualization

We visualize the feature before/after HCO in a random layer in stage 2 with randomly selected images as input, Fig. 11. Before HCO, only a few regions of the foreground object are activated. After HCO, almost the entire foreground object is activated intensively.

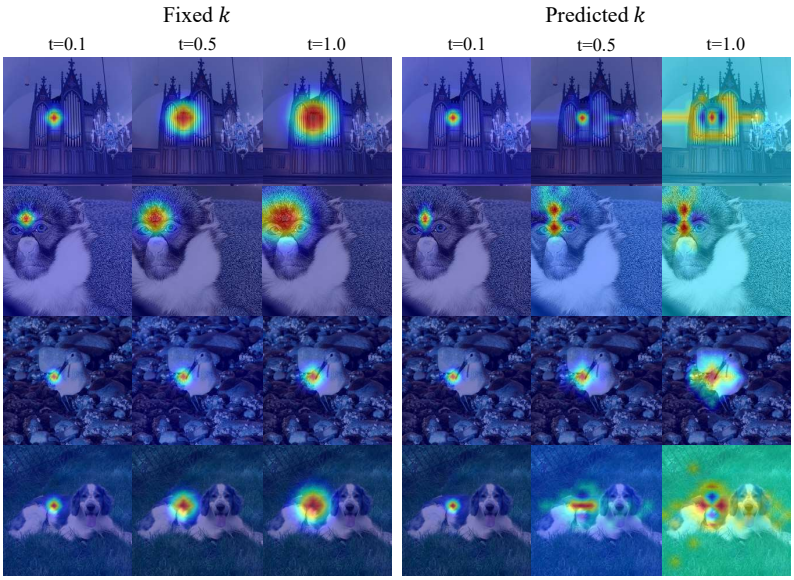

Figure 9. Temperature distribution (U^t) when using a randomly selected patch as the heat source. (Best viewed in color)

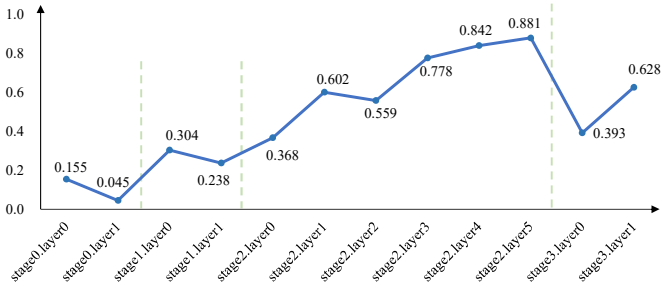

Figure 10. Average values of k in each layer.

I. Ablation of the Linear+SiLU branch

The Linear + SiLU branch is a gated MLP unit, which is inspired by the design used in LLaMA [62]. We conducted ablation experiments, and the results, as shown in Table 12, demonstrate that the primary performance improvement of vHeat comes from the HCO, rather than the gated MLP.

Model	#Param.	FLOPs	Top-1 acc. (%)
vHeat-B	68M	11.2G	84.0
Linear + ReLU	68M	11.2G	83.9
w/o (Linear + SiLU)	62M	10.3G	83.6
w/o HCO	49M	8.0G	76.7

Table 12. Ablation study of the Linear + SiLU branch.

J. Comparison with SOTAs

We compare vHeat-B with other base-level SOTA visual representation models (MetaFormer-v2-M48, CAFormer-B36 [75], iFormer-B [51], and BiFormer-B [81]) on Top-1 Accuracy on ImageNet-1K and test throughput (an A100 GPU with 128 batch size), Figure 12. Our proposed vHeat demonstrates comparable performance with substantially improved test throughput.

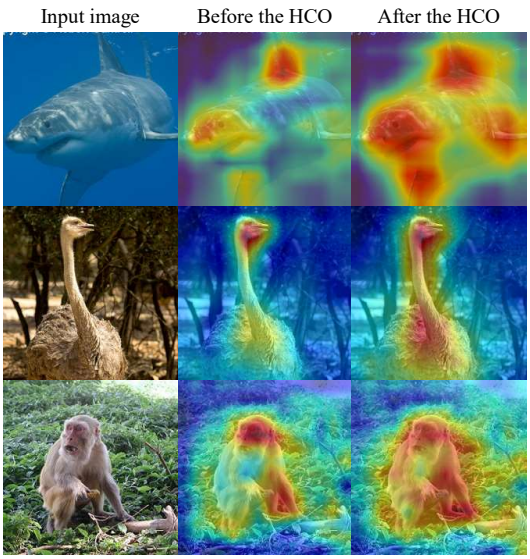

Figure 11. Visualization of the feature before/after HCO in a random layer in stage 2 with ImageNet-1K classification pre-trained vHeat-B. The images are randomly selected from ImageNet-1K.

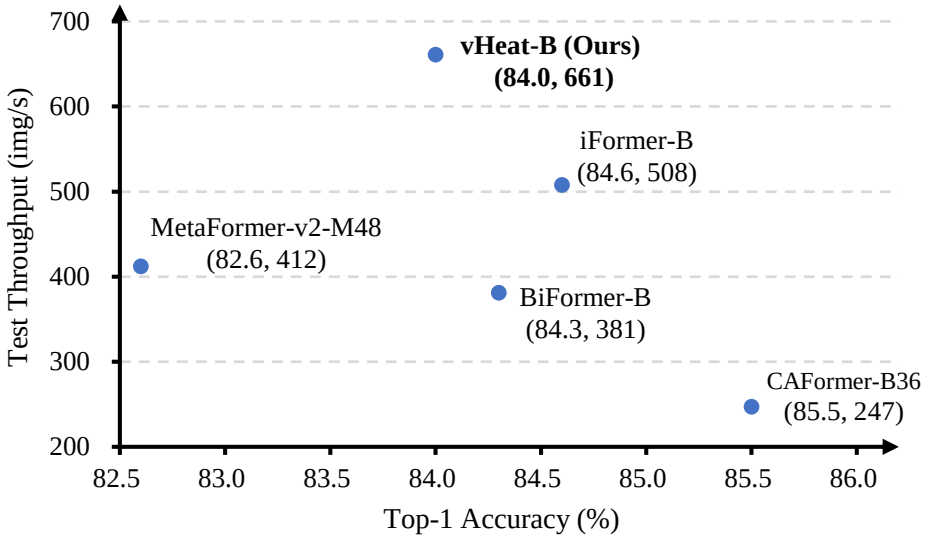

Figure 12. Comparison of vHeat-B and other base-level SOTA vision models.