Eurographics Symposium on Geometry Processing 2023
P. Memari and J. Solomon
(Guest Editors)

Volume 42 (2023), Number 5

SUPPLEMENTARY MATERIALS: Attention And Positional Encoding Are (Almost) All You Need For Shape Matching

Alessandro Raganato[1] and Gabriella Pasi[1] and Simone Melzi[1]

[1]Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicoccca, Italy.

In this document, we report all the additional details that do not find a place in the main manuscript due to lack of space.

1. Details on the datasets

In this Section, we report some additional details on the datasets involved in our evaluation.

1.1. Human shapes datasets

F$_{\sim 7K}$ [BRLB14]: 10 subjects, in the same 10 articulated poses, all represented with the same triangular mesh with 6890 vertices.

F$_{1K}$: The same shapes of F$_{\sim 7K}$, remeshed to the same triangular mesh with $1K$ vertices. The same mesh is applied to all the shapes, which thus share a 1:1 correspondence.

F$_{1K}$ N: The same shapes of F$_{1K}$ with Gaussian noise applied to the 3D coordinates with standard deviation equal to 0.01. This noise should simulate some error in the acquisition process.

F$_{1K}$ O: The same shape from F$_{\sim 7K}$ with a different sampling of $1K$ points. Some of them have been randomly moved far from the surface, generating a sparse set of outliers.

S19 [MMR*19]: 44 shapes from different repositories, whit various triangulation, numbers of points (from $\sim 5K$ to more than $200K$), poses, and subjects. A list of 430 pairs is provided to evaluate the resiliency of the method to different connectivities and densities.

1.2. Animal shapes datasets

SMAL 100 random pairs selected among 300 shapes generated with the SMAL parametric model that have never been seen during the training. The shapes belong to all the classes generated by SMAL (cat, dog, cow, horse, hippo).

HIPPO 100 random pairs composed only of hippos (with different sizes, proportions, and poses) generated with the SMAL model that have never been seen during the training.

TOSCA [BBK08] is a benchmark of synthetic triangular meshes belonging to different classes with various poses (cat, dog, horse, wolf, 3 human subjects, among others). We remesh them to $\sim 10K$ vertices. Shapes in each class are isometric, and with the vertices ordered in 1:1 correspondence. We only consider pairs composed of shapes from the same class. For the test on the animal shapes, we consider only the classes of dog, horse, and wolf. We exclude the cat class involved in the training of a specific model we test in our experiments.

2. Evaluation

2.1. Surface attention in Ours$_{SA}$

Following [TCM*21], for each point $x_p \in \mathcal{X}$, we estimate its area contribution as the inverse of the local point density: $\mathcal{A}(\mathcal{X})_p = (|\{x_j \in \mathcal{X} \text{ s.t. } \|x_j - x_p\|_2 < r\}|)^{-1}$, where $|\cdot|$ denotes the cardinality of a set, and r is a local radius set to 0.05. In our case, the attention energy ξ defined in equation 4 in the main paper, is changed as:

$$\xi = \frac{\xi \mathcal{A}_C(\mathcal{X}, SEP, \mathcal{Y})}{\sum \xi \mathcal{A}_C(\mathcal{X}, SEP, \mathcal{Y})} \qquad (1)$$

where $\mathcal{A}_C(\mathcal{X}, SEP, \mathcal{Y})$ is the concatenation of the $\mathcal{A}(\mathcal{X})$, $\mathcal{A}(\mathcal{Y})$, and a dummy separator SEP set to 1.

2.2. Evaluation metrics

Given the ground truth correspondence $\Pi^{GT}_{\mathcal{X},\mathcal{Y}}$ between X and Y (*i.e.*, $y = \Pi^{GT}_{\mathcal{X},\mathcal{Y}}(x) \in \mathcal{Y}$ is the correct match $\forall x \in \mathcal{X}$), we compute the geodesic error $\mathcal{E}^{\Pi}_{\mathcal{X},\mathcal{Y}}(x)$ of the estimated correspondence Π as:

$$\mathcal{E}^{\Pi}_{\mathcal{X},\mathcal{Y}}(x) = \mathcal{G}_{\mathcal{Y}}(\Pi^{GT}_{\mathcal{X},\mathcal{Y}}(x), \Pi(x)), \qquad (2)$$

where $\mathcal{G}_{\mathcal{Y}}$ is the geodesic distance on the surface Y. The average geodesic error, namely AGE, is the average value of this error on all the points that discretize $\mathcal{X}$. The average geodesic error, namely AGE, which is the mean value of this error on all the points that discretize $\mathcal{X}$:

$$AGE^{\Pi}_{\mathcal{X},\mathcal{Y}} = \frac{1}{n_{\mathcal{X}}} \sum_{x \in X} \mathcal{E}^{\Pi}_{\mathcal{X},\mathcal{Y}}(x), \qquad (3)$$

is the main metric adopted to globally assesses the estimated correspondence. In the same way, we can compute these errors in the opposite direction if the $\Pi^{GT}_{\mathcal{Y},\mathcal{X}}$ is available. In the tables, for each dataset, we report the average value of the *AGE* on a collection of pairs. For visualization, we encode this error in a colormap, points with large errors have dark colors while 0 errors are white.

A common metric we can adopt when the ground truth correspondence is not available is the Chamfer distance, which for the pair X,Y is defined as:

$$\sum_{x\in X}\min_{y\in Y}\|x-y\|_2^2+\sum_{y\in Y}\min_{x\in X}\|y-x\|_2^2. \tag{4}$$

This measure is more specific for shape registration, but it can be useful also for shape matching in absence of Π^{GT}.

2.3. Evaluation with augmentation

As described in the main test, we apply two types of augmentation at training time: random rotations and random permutation. The loss we adopt requires the 1:1 correspondence between $\mathcal{X}$ and $\mathcal{Y}$ that is not available after the transformations we apply in the augmentation step. For this reason, for every pair, we apply the inverse of these transformations to the output of our model to make the loss evaluation meaningful.

3. Details about our choices

3.1. Matching and not registration

As we state in the main paper, with our work, we aim to target the shape matching problem instead of the registration one. Even if potentially more precise, we believe that focusing on the analysis of the components and design of the transformer, it is better to target the more general shape matching task. Moreover, the metric (*AGE*) we adopt to evaluate and compare the performances is properly defined for the shape matching and not for the registration problem. To be as general as possible in our pipeline, we do not require any template or particular properties for one of the two shapes in each pair which is proper for the registration setting. Given that, we recognize that, following the method proposed in [TCM*21], the procedure we apply to compute the estimated correspondence belongs to the registration setting. The loss we exploit provides strong constraints about the desired solution and makes the shape matching problem (combinatorial in its nature) easier to optimize. Furthermore, by implementing the same loss, our model is directly comparable to SRTT, which is the main objective of our evaluation. Our main limitation is the fixed number of points that we can input into our model. We consider this a strong limitation for the registration, even if in shape matching, this does not hurt the performance too much, as we can appreciate from the quantitative evaluation, in which we outperform the competitors most of the time. In the future, we aim to target this limitation and, once do that, extend our model and experiment with its potential in the registration task.

3.2. Implementation details

Our implementation follows the Transformer encoder *base* setting [VSP*17]. Specifically, we use 6 layers, 512 as a model dimension, 2048 as a hidden dimension, and 8 attention heads of 64 dimension each. As an initial dimensionality augmentation, as shown in Figure 2 in the main paper, we use 6 linear layers of size (16,32,64,128,256,512) interleaved by the Tanh activation function. Similarly, as a final dimensionality reduction, we use 6 linear layers of size (256,128,64,32,16,3) with the Tanh activation function between all of them. We train our model for 5000 epochs on an NVIDIA A100 using Adam optimizer [KB15] with a constant learning rate of 1e-4, and batch size of 8. We release our complete implementation at: https://github.com/raganato/SGP23_AttPos4ShapeMatching.

We note that, to allow the model to learn sparse attention weights, we could replace the softmax operator in the attention heads with α-entmax functions [MA16, PNM19, CNM19]. However, this did not lead to any change in performance in the ablation setting (see Table 1). For simplicity, we keep the softmax operator.

Moreover, we note that we could also train our model to minimize the Chamfer distance respectively between $\mathcal{X}$ and $\widehat{\mathcal{Y}}$ or between $\mathcal{Y}$ and $\widehat{\mathcal{X}}$, without using the ground truth correspondence (see Equation 5 in the main paper). However, our main interest is in the analysis of the proposed architecture in relation to the matching problem, thus we do not explore this possibility and leave it as a future direction.

Table 1: *Ablation study*

Method	♯Params	F_{1K}	F_{1K}N	DEV
Ours	19.2M	0.0880	0.0826	0.0489
Ours-entmax	19.2M	0.0880	0.0826	0.0489

3.3. Details on Figure 8

Figure 8 in the main paper shows a 3D Principal Component Analysis (PCA) visualization using the embedding projection online tool https://projector.tensorflow.org/.

4. Additional Results

In this Section, we collect some additional figures and quantitative results.

4.1. Additional visualizations

In Figure 1, we depict the comparison between **Ours** and SRTT on a second pair from SMAL dataset. As can be seen, the poses of these animals are quite extreme. Also, in this case, our method outperforms the competitor. In particular, while SRTT is wrongly mapping two legs of the cow on the same leg of the hippo (visualized through the dark yellow color of the points in these regions), our result is correct. Figures 2 and 3 contain other results for the texture transfer application on pairs from the challenging S19 [MMR*19] dataset. In Figure 2, we visualize the texture transfer results produced by our method. Finally, In Figure 3, we report another comparison with SRTT. Once again, we highlight that even if numerically worse than SRTT on S19, we are accurate enough to target this application.

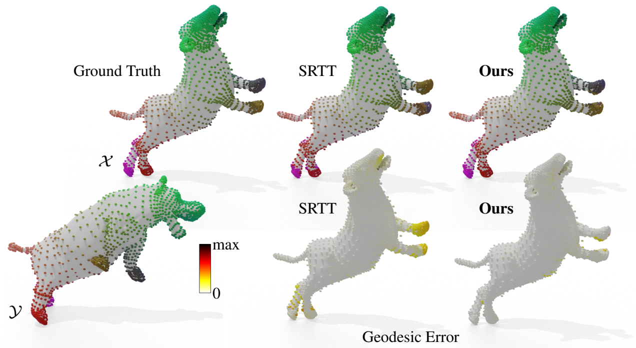


Figure 1: *Comparison between **Ours** and SRTT [TCM*21] on a pair from the SMAL dataset.*

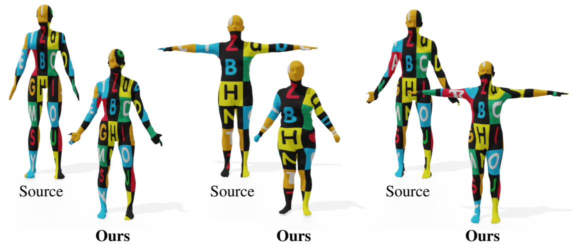


Figure 2: *Three texture transfer results (source on the top and transfer on the bottom of each pair) obtained from the correspondences estimated by our method on the challenging SHREC 19 [MMR*19]. Even if numerically worse than the scores of SRTT on the same dataset, we are accurate enough to target this application.*

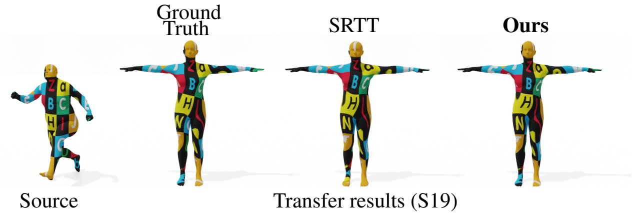


Figure 3: *Texture transfer results for one pair from S19. From left to right, we visualize the source shape with the texture, the ground truth transfer, the output of SRTT, and our result.*

4.2. Test on shapes with different numbers of vertices

In Table 2, we collect the quantitative evaluation of shapes with different densities. In the first and second columns, we have the results for the F_{1K} and $F_{\sim 7K}$ that are also in Table 1 of the main manuscript. In the last column we report the results for the same pairs but with $\mathcal{X}$ from F_{1K} and $\mathcal{Y}$ from $F_{\sim 7K}$. The last one is a more complicated setting due to the different densities of the two points distributions. The errors do not change too much in any case. As can be appreciated, even if a bit less stable, our method outperforms SRTT in every case. Furthermore, as highlighted in the main document, **Ours**$_\star$ is more stable than **Ours**, proving that the additional training with a different sampling of the points can help. We aim to explore this direction further in the future. Finally, neither **Ours**$_{SA}$ and **Ours**$_{SA\star}$ outperform **Ours**$_\star$ confirming that the surface attention by itself does not improve our solution even in the case of shapes with different numbers of vertices.

Table 2: *Comparison to SRTT [TCM*21] on shapes with different numbers of vertices.*

Method	F_{1K}	$F_{\sim 7K}$	F_{1to7K}
SRTT	0.042	0.051	0.050
SRTT$_\star$	0.036	0.044	0.043
Ours	0.014	0.024	0.029
Ours$_\star$	0.013	0.020	0.022
Ours$_{SA}$	0.015	0.023	0.027
Ours$_{SA\star}$	0.017	0.022	0.024

5. Analysis of the attention

In this section, we report all the additional analyses that were not included in the main paper due to the limited space.

We highlight here the main insights we inherit from these visualizations.

- **Ours** shows clear and sharp self and cross-attention patterns. Each head specializes in a specific pattern across the heads.
- **No Bid** does not present any cross-attention patterns, and the subdivision between the entries related to the two shapes is not sharp as in **Ours**.
- **No RA** does not specialize the heads. As an example, Head 1 switches from self to cross-attention in Layers 4 and 5.
- **No Pos** does not generate any meaningful pattern, as we understand from the bad quantitative results of this model.

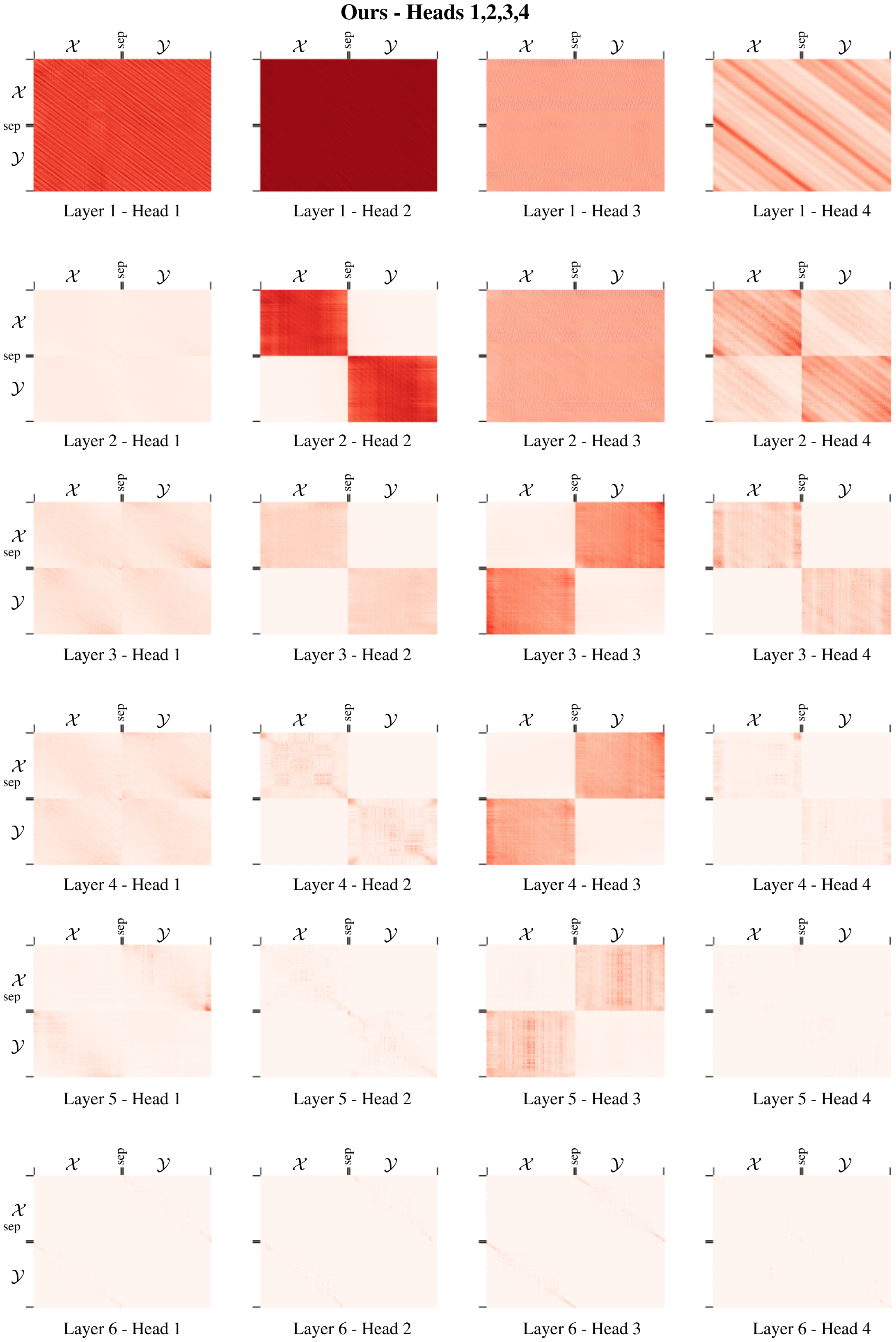
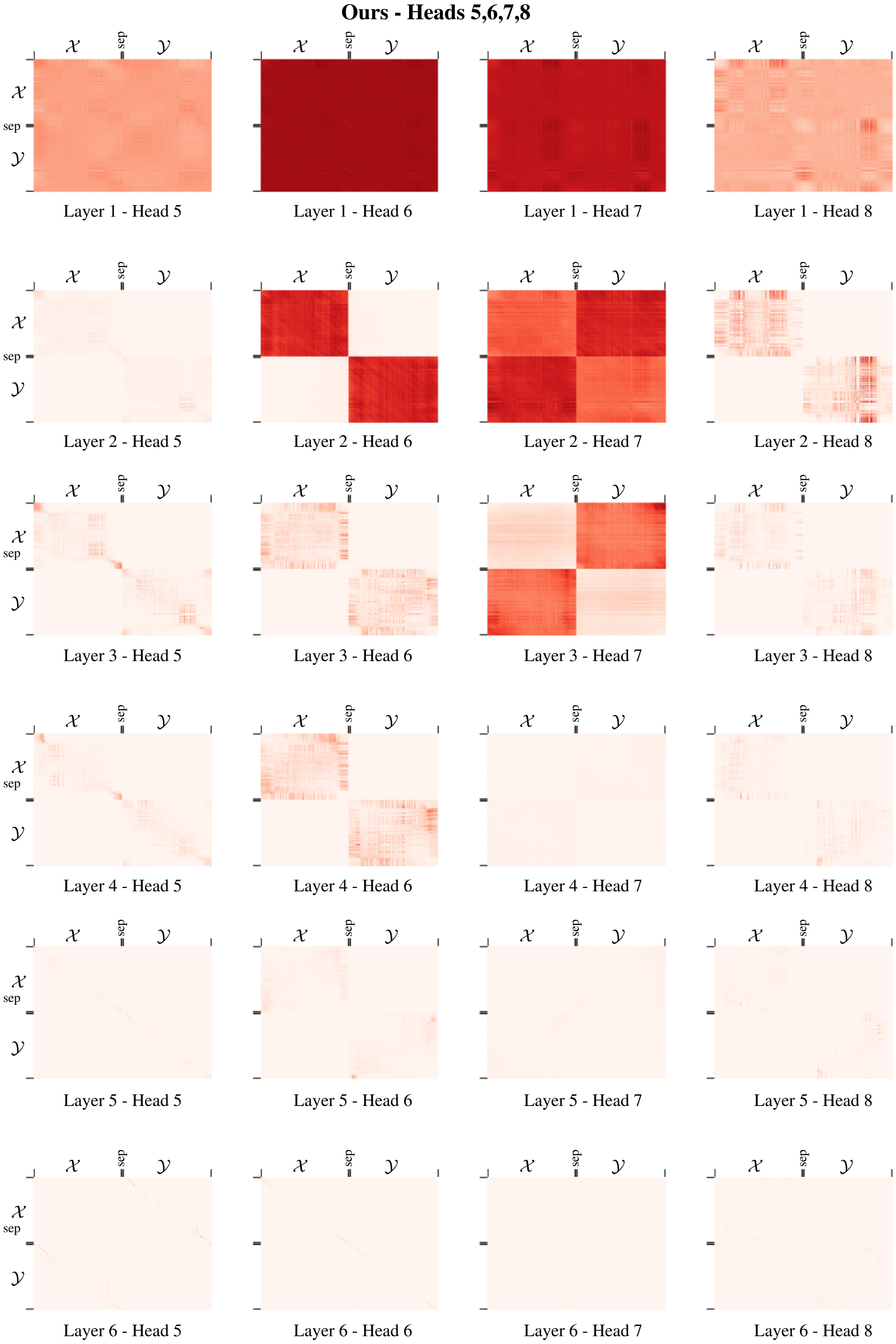
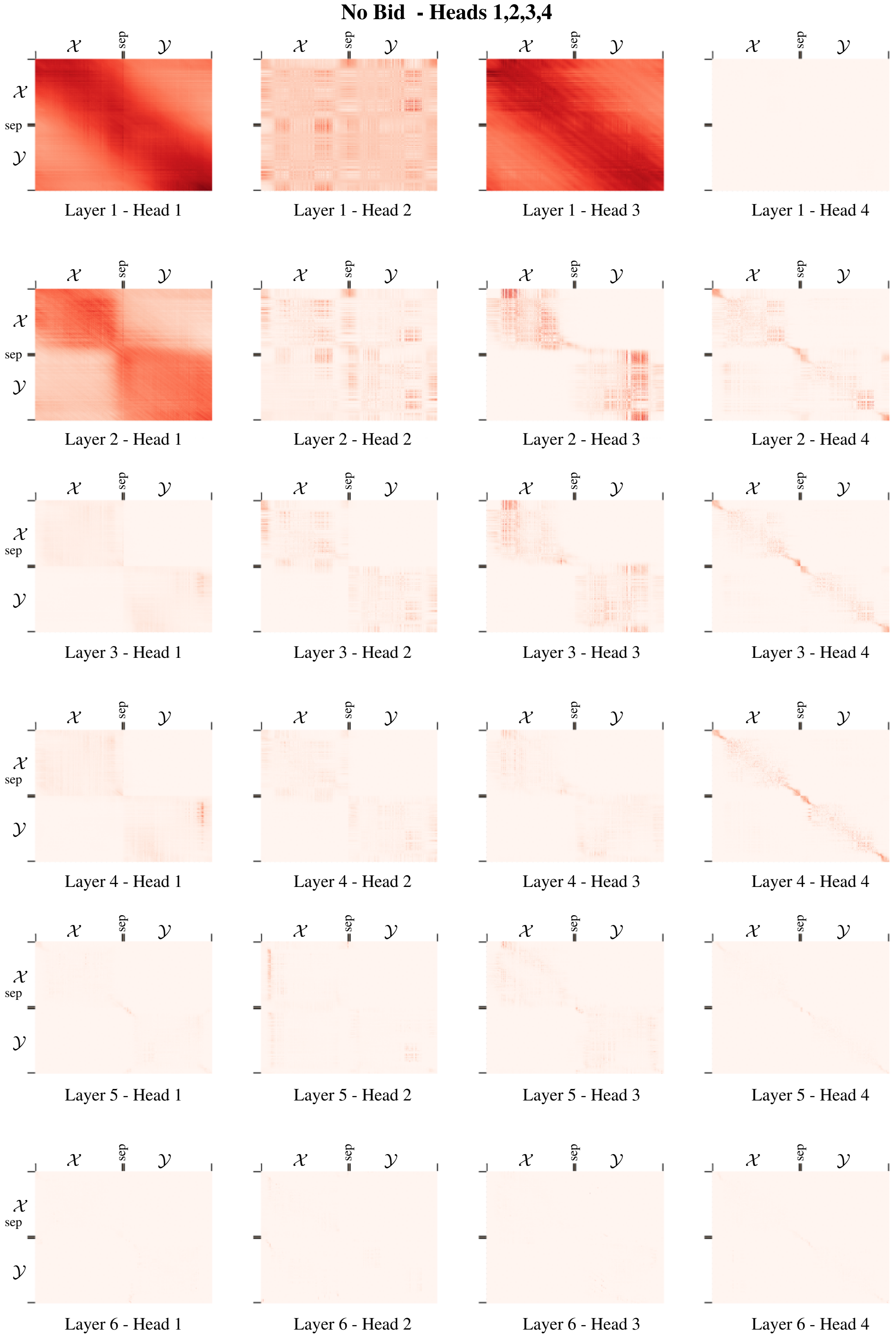


Figure 4: *Visualization of the first four attention heads of **Ours** model across the 6 layers of the architecture. The stronger the red, the higher the attention score.*



Layer 1 - Head 5
Layer 1 - Head 6
Layer 1 - Head 7
Layer 1 - Head 8
Layer 2 - Head 5
Layer 2 - Head 6
Layer 2 - Head 7
Layer 2 - Head 8
Layer 3 - Head 5
Layer 3 - Head 6
Layer 3 - Head 7
Layer 3 - Head 8
Layer 4 - Head 5
Layer 4 - Head 6
Layer 4 - Head 7
Layer 4 - Head 8
Layer 5 - Head 5
Layer 5 - Head 6
Layer 5 - Head 7
Layer 5 - Head 8
Layer 6 - Head 5
Layer 6 - Head 6
Layer 6 - Head 7
Layer 6 - Head 8

Figure 5: *Visualization of the last four attention heads of* ***Ours*** *model across the 6 layers of the architecture. The stronger the red, the higher the attention score.*


No Bid - Heads 1,2,3,4

Layer 1 - Head 1 Layer 1 - Head 2 Layer 1 - Head 3 Layer 1 - Head 4

Layer 2 - Head 1 Layer 2 - Head 2 Layer 2 - Head 3 Layer 2 - Head 4

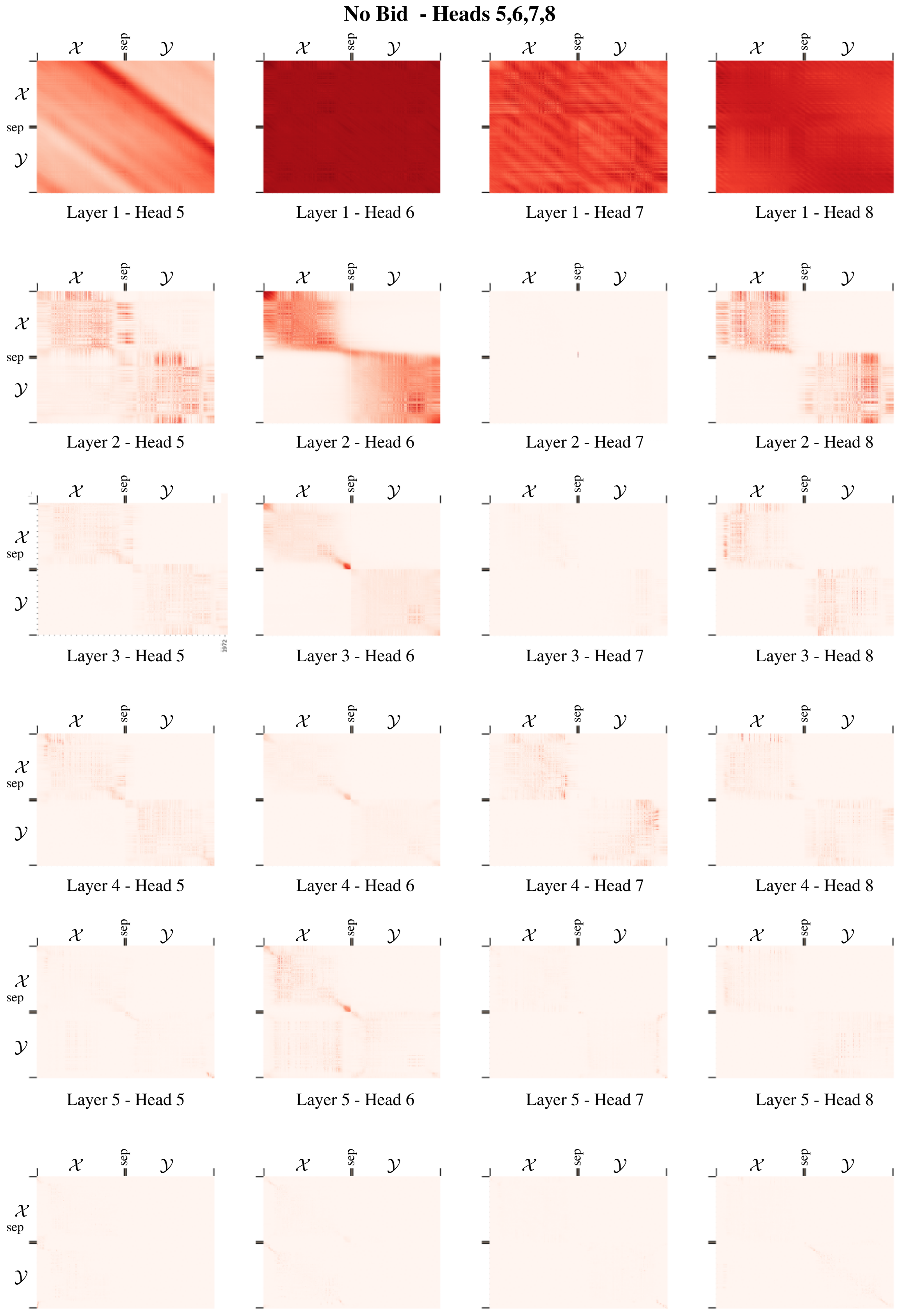
Layer 3 - Head 1 Layer 3 - Head 2 Layer 3 - Head 3 Layer 3 - Head 4

Layer 4 - Head 1 Layer 4 - Head 2 Layer 4 - Head 3 Layer 4 - Head 4

Layer 5 - Head 1 Layer 5 - Head 2 Layer 5 - Head 3 Layer 5 - Head 4

Layer 6 - Head 1 Layer 6 - Head 2 Layer 6 - Head 3 Layer 6 - Head 4

Figure 6: *Visualization of the first four attention heads of* ***No Bid*** *model across the 6 layers of the architecture. The stronger the red, the higher the attention score.*



Layer 1 - Head 5 Layer 1 - Head 6 Layer 1 - Head 7 Layer 1 - Head 8

Layer 2 - Head 5 Layer 2 - Head 6 Layer 2 - Head 7 Layer 2 - Head 8

Layer 3 - Head 5 Layer 3 - Head 6 Layer 3 - Head 7 Layer 3 - Head 8

Layer 4 - Head 5 Layer 4 - Head 6 Layer 4 - Head 7 Layer 4 - Head 8

Layer 5 - Head 5 Layer 5 - Head 6 Layer 5 - Head 7 Layer 5 - Head 8

Layer 6 - Head 5 Layer 6 - Head 6 Layer 6 - Head 7 Layer 6 - Head 8

Figure 7: *Visualization of the last four attention heads of* ***No Bid*** *model across the 6 layers of the architecture. The stronger the red, the higher the attention score.*

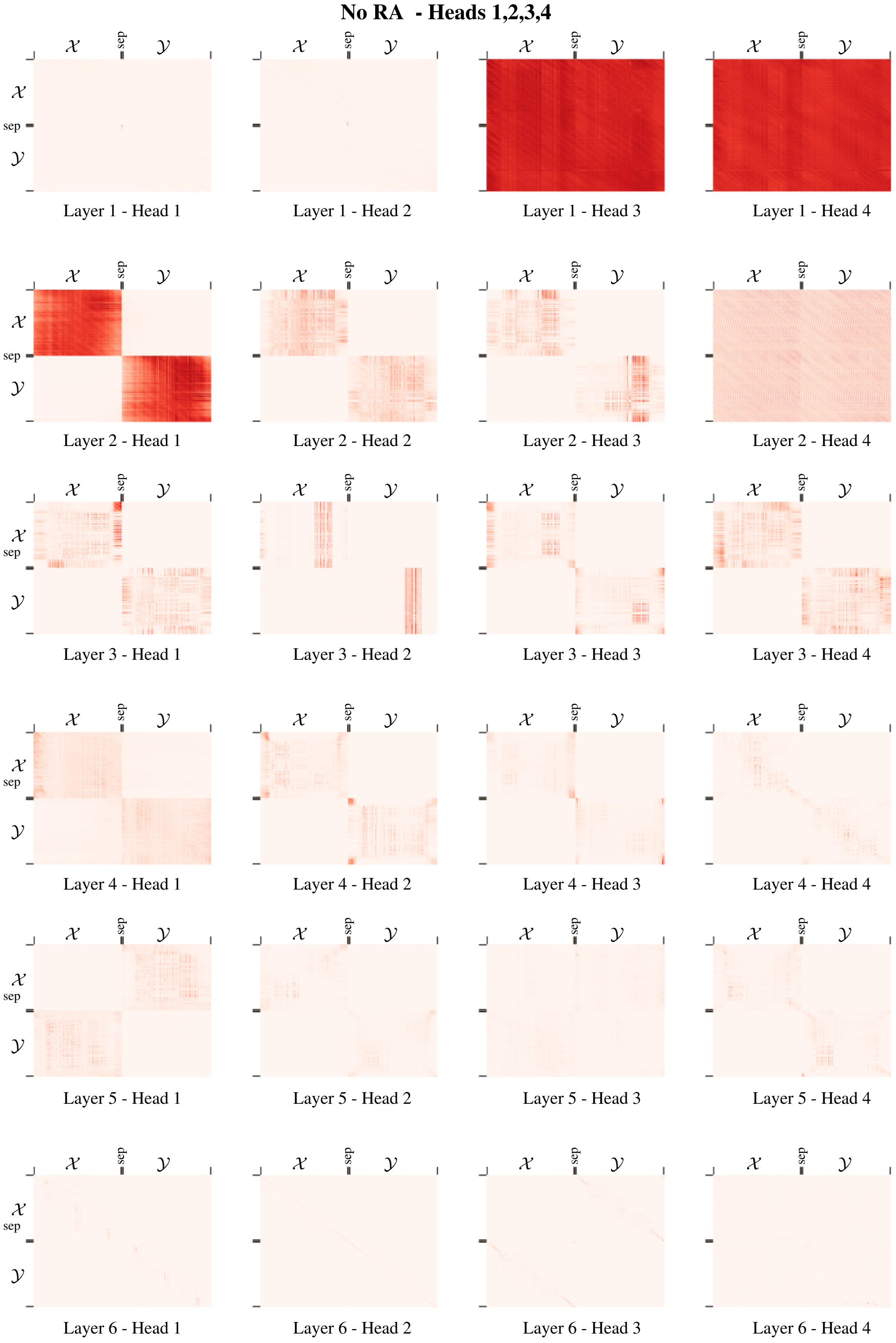
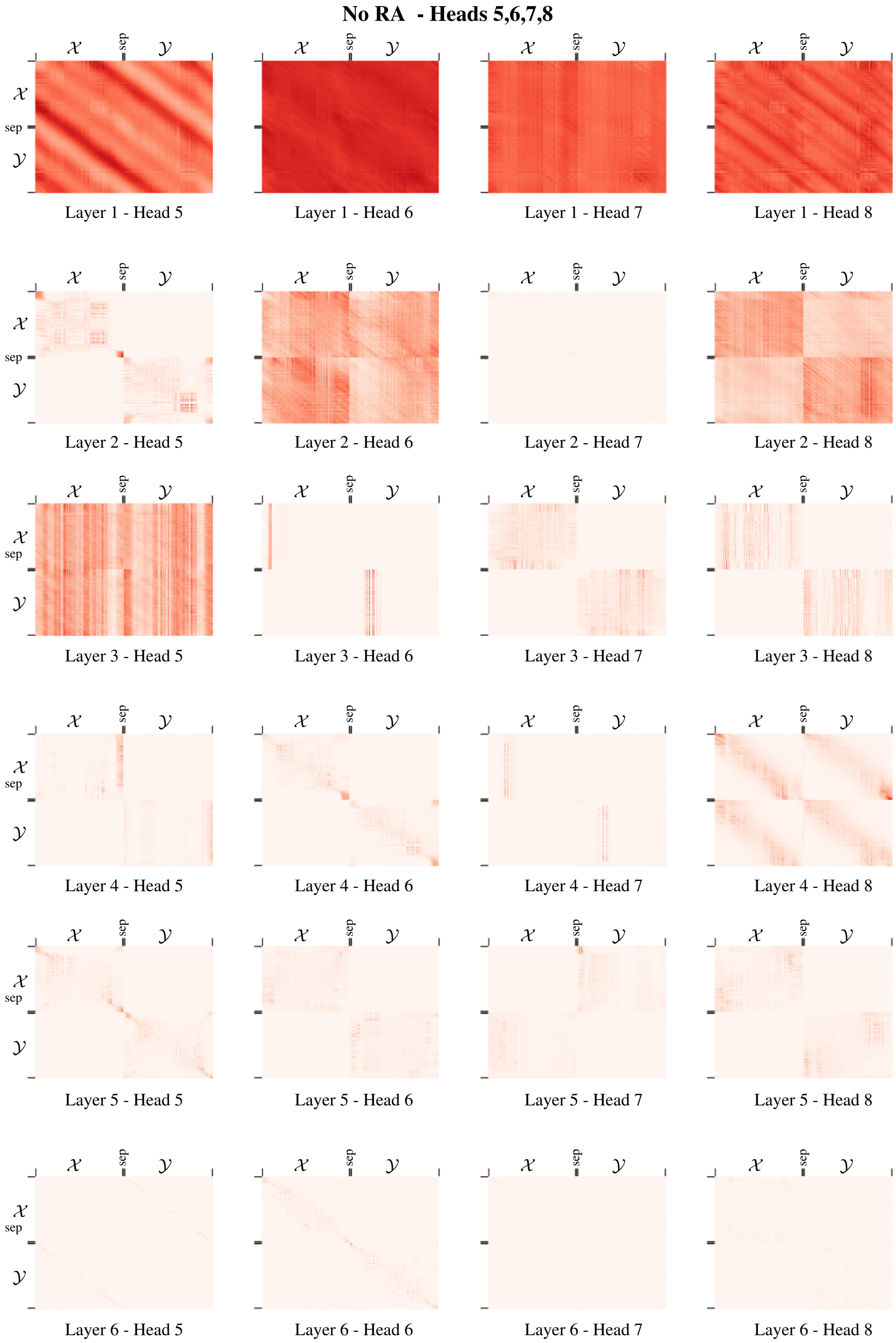


Figure 8: *Visualization of the first four attention heads of* ***No RA*** *model across the 6 layers of the architecture. The stronger the red, the higher the attention score.*



Layer 1 - Head 5 Layer 1 - Head 6 Layer 1 - Head 7 Layer 1 - Head 8

Layer 2 - Head 5 Layer 2 - Head 6 Layer 2 - Head 7 Layer 2 - Head 8

Layer 3 - Head 5 Layer 3 - Head 6 Layer 3 - Head 7 Layer 3 - Head 8

Layer 4 - Head 5 Layer 4 - Head 6 Layer 4 - Head 7 Layer 4 - Head 8

Layer 5 - Head 5 Layer 5 - Head 6 Layer 5 - Head 7 Layer 5 - Head 8

Layer 6 - Head 5 Layer 6 - Head 6 Layer 6 - Head 7 Layer 6 - Head 8

Figure 9: *Visualization of the last four attention heads of* ***No RA*** *model across the 6 layers of the architecture. The stronger the red, the higher the attention score.*

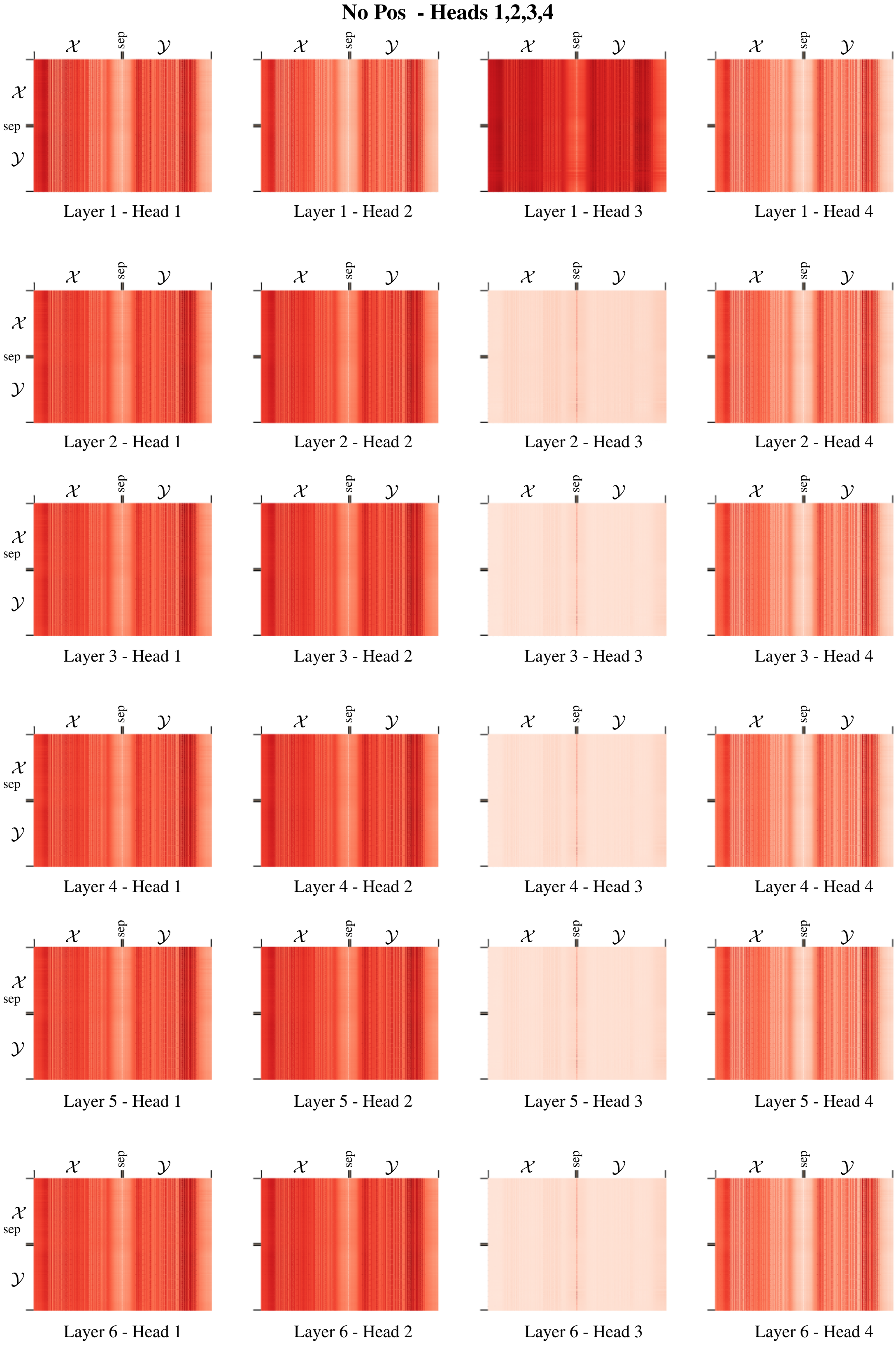


Figure 10: *Visualization of the first four attention heads of* ***No Pos*** *model across the 6 layers of the architecture. The stronger the red, the higher the attention score.*

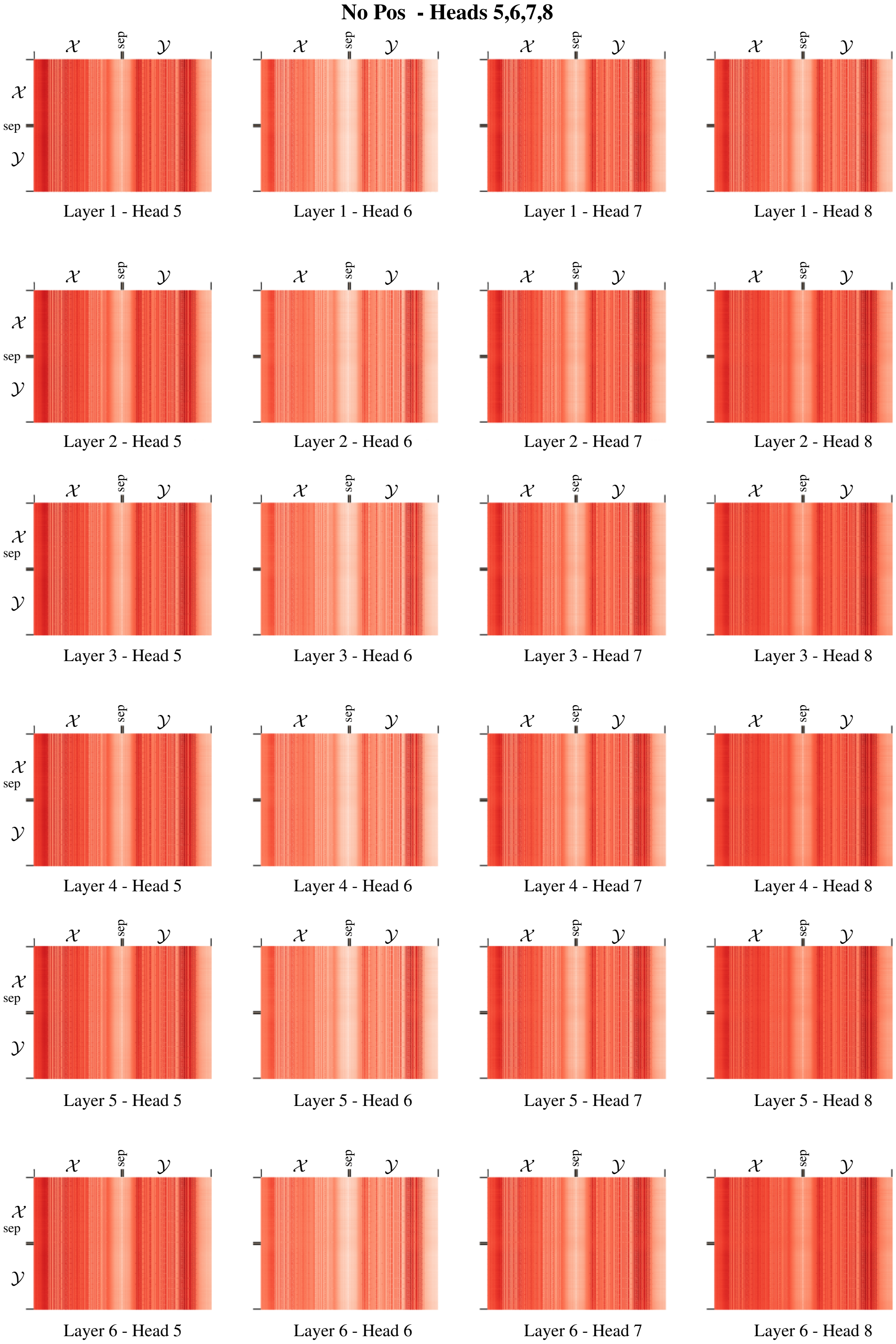


Figure 11: *Visualization of the last four attention heads of* ***No Pos*** *model across the 6 layers of the architecture. The stronger the red, the higher the attention score.*

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the RTX A5000 GPUs granted through the Academic Hardware Grant Program to the the University of Milano-Bicocca for the project "Learned representations for implicit binary operations on real-world 2D-3D data", and the CSC – IT Center for Science, Finland, for computational resources.

References

[BBK08] Bronstein A., Bronstein M., Kimmel R.: *Numerical Geometry of Non-Rigid Shapes*. Springer, New York, NY, 2008. 1

[BRLB14] Bogo F., Romero J., Loper M., Black M. J.: FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (Piscataway, NJ, USA, June 2014), IEEE. 1

[CNM19] Correia G. M., Niculae V., Martins A. F. T.: Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 2174–2184. 2

[KB15] Kingma D. P., Ba J.: Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015), Bengio Y., LeCun Y., (Eds.). 2

[MA16] Martins A., Astudillo R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning* (2016), PMLR, pp. 1614–1623. 2

[MMR*19] Melzi S., Marin R., Rodolà E., Castellani U., Ren J., Poulenard A., Wonka P., Ovsjanikov M.: Matching Humans with Different Connectivity. In *Eurographics Workshop on 3D Object Retrieval* (2019), Biasotti S., Lavoué G., Veltkamp R., (Eds.), The Eurographics Association. 1, 2, 3

[PNM19] Peters B., Niculae V., Martins A. F. T.: Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 1504–1519. 2

[TCM*21] Trappolini G., Cosmo L., Moschella L., Marin R., Melzi S., Rodolà E.: Shape registration in the time of transformers. In *NeurIPS* (2021), Ranzato M., Beygelzimer A., Dauphin Y., Liang P., Vaughan J. W., (Eds.), pp. 5731–5744. 1, 2, 3

[VSP*17] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I.: Attention is all you need. In *Proc. of NeurIPS* (2017). 2