Supplementary Materials for Multimodal Optimal Transport-based Co-Attention Transformer with Global Structure Consistency for Survival Prediction

A. Outline

The supplementary materials for this paper are organized as follows:

1. We demonstrate the qualitative visualization for the effect on the size of Micro-Batch Strategy.
2. We provide visualization cases of Co-Attention.

B. Effect on Size of Micro-Batch

In this section, we compare several variants of the proposed method to show the robustness to the size of Micro-batch strategy used in histology data, in which the results over the size of 128, 256 and 512 are presented in Fig. 1.

There are two variants of the proposed method mentioned in the quantitative analysis of Section 4.3 and MCAT to be compared for the qualitative analysis. As shown in Fig. 1, we visualize the co-attention values between the first 300 instances of a WSI and all genomic instances ($M_g = 6$) from the same patient.

We observe that 1) our method demonstrates the best consistency of activation among different sizes of Micro-Batch, while the variant (c) UMBOT $\rightarrow$ EMD shows considerably poor consistency, which validates that UMBOT-based co-attention used in our method is more robust to the size of Micro-Batch than the original OT (i.e. EMD). 2) When we replace UMBOT with the co-attention used in MCAT and train it over Micro-Batch, we found that the activation pattern of size 512 is significantly different from that of sizes 128 and 256, as shown in the variant (b) UMBOT $\rightarrow$ CoAttn. 3) Furthermore, results of all sizes in variant (b) are apparently distinguished from the co-attention of MCAT directly computed over all instances, further indicating the poor robustness of the original co-attention.

C. Visualization of Co-Attention

To show the interpretability, we visualize the co-attention values of all instances in each WSI for high and low cases, as shown in Fig. 2 of our method and Fig. 3 of MCAT [1]. In order to present a more obvious difference in co-attention values among various genomic instances, we consider the instance of Tumor Suppression (marked in red) as the reference and show the differences of other genomic instances from it, since there is only a slight difference in the values of co-attention.

By comparing MCAT with the proposed method, we found that the OT-based co-attention is concerned about different areas of the WSI for different genomic functional instances, while the dense co-attention used in MCAT focuses on the similar regions of histology for different functional instances of genes. As a result, the better performance of our method may benefit from these different concerns.

References

[1] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021. 1, 4

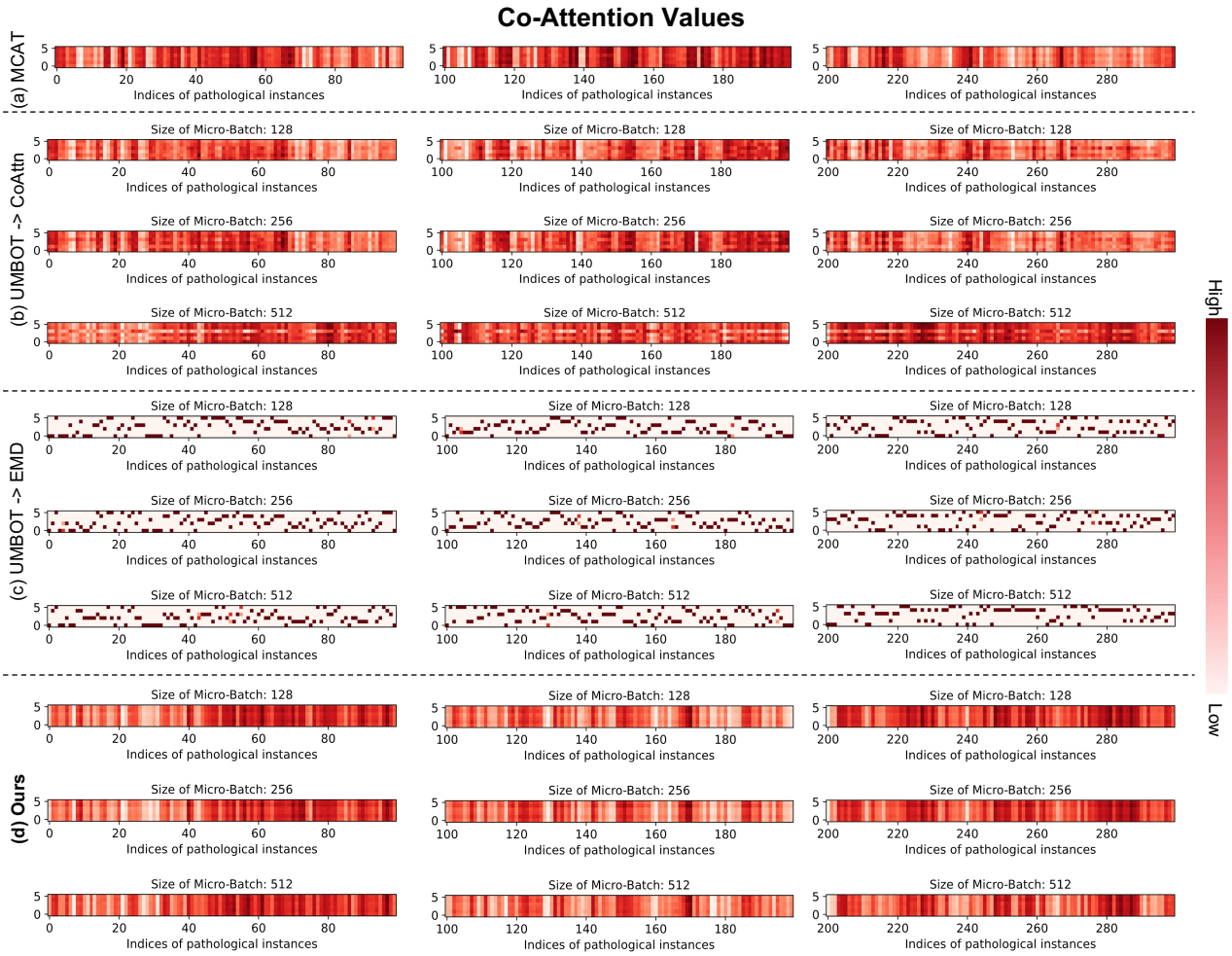

Figure 1. Co-Attention values between the first 300 instances of histology and all instances of genomics for the case *TCGA-06-0210* of GBMLGG: (a) MCAT, (b) UMBOT → CoAttn, (c) UMBOT → EMD and (d) our method, in which the size of Micro-Batch ranges from 128 to 512.

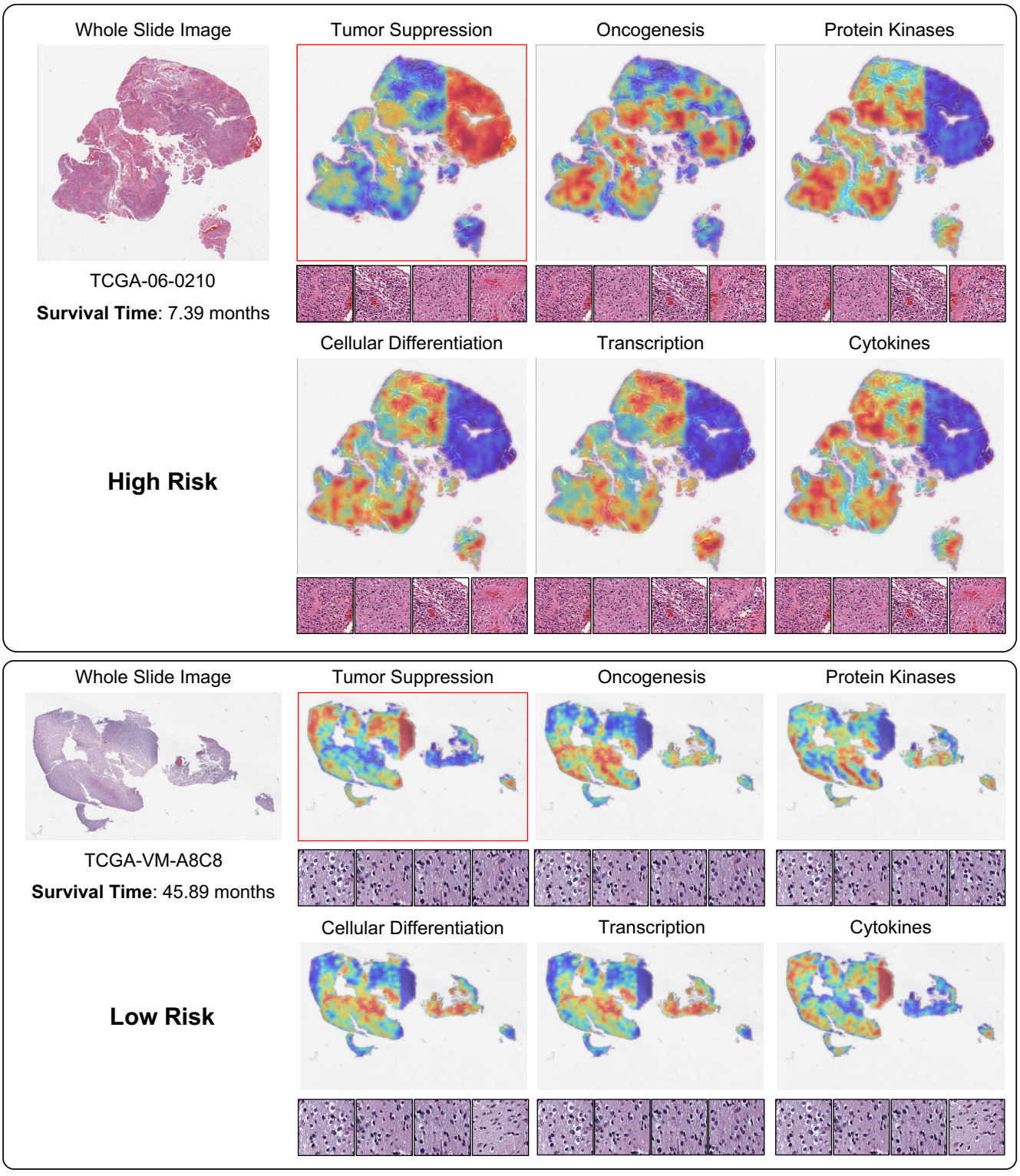

Figure 2. Co-Attention visualization of **our method** for high and low risk cases in GBMLGG, with corresponding top-4 highest attention patches for each genomic instance of unique functional category.

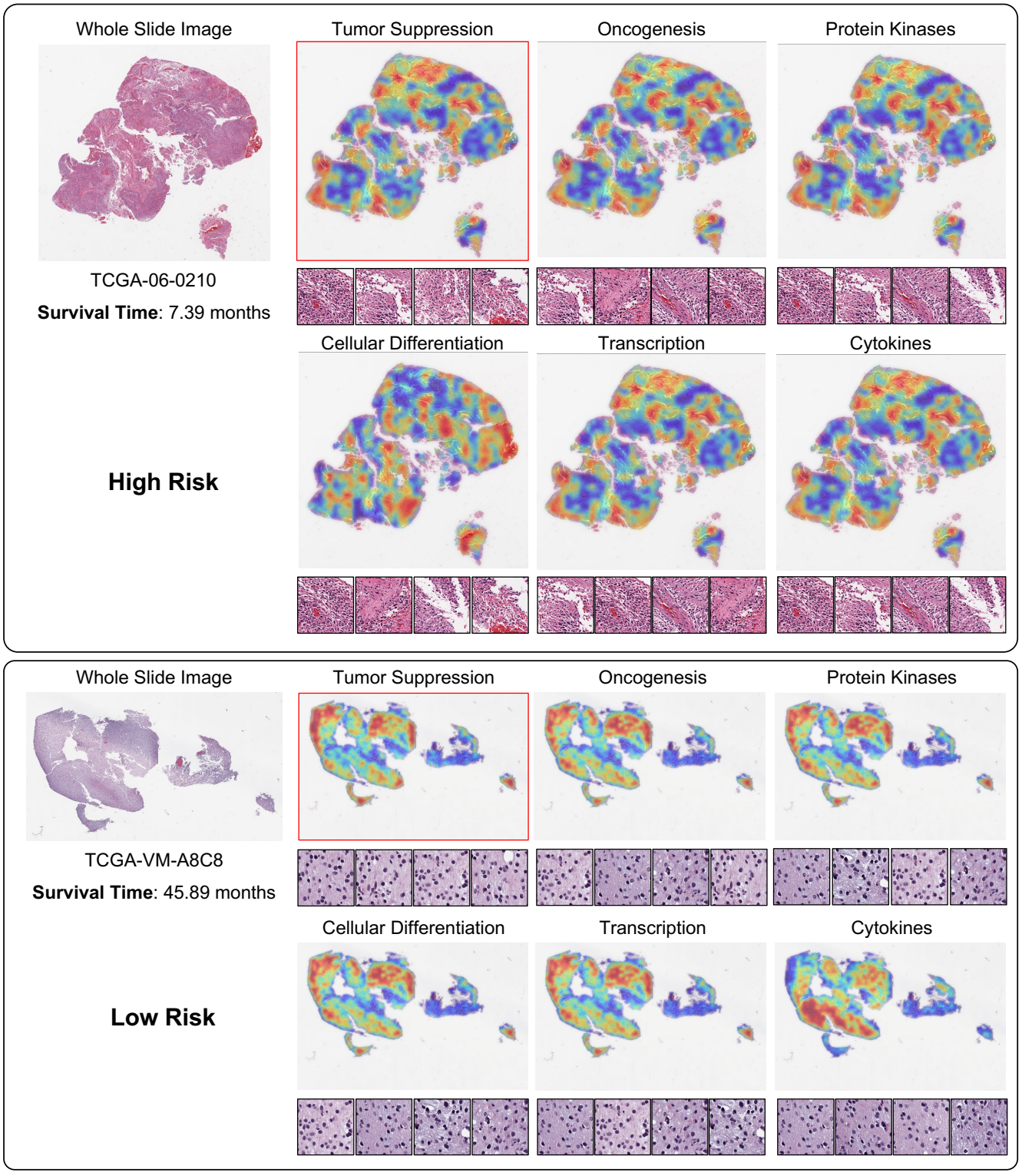

Figure 3. Co-Attention visualization of **MCAT** [1] for high and low risk cases in GBMLGG, with corresponding top-4 highest attention patches for each genomic instance of unique functional category.