TransForming Visual Scene Graphs to Image Captions

将场景图转换为图像文本

Xu Yang[1] Jiawei Peng[1] Zihua Wang[1] Haiyang Xu[2*] Qinghao Ye[2]
Chenliang Li[2] Songfang Huang[2] Fei Huang[2] Zhangzikang Li[1] Yu Zhang[1*]

1 东南大学计算机科学与工程学院、软件学院、人工智能学院-模式学习与挖掘实验室（Palm）
2 阿里巴巴达摩院

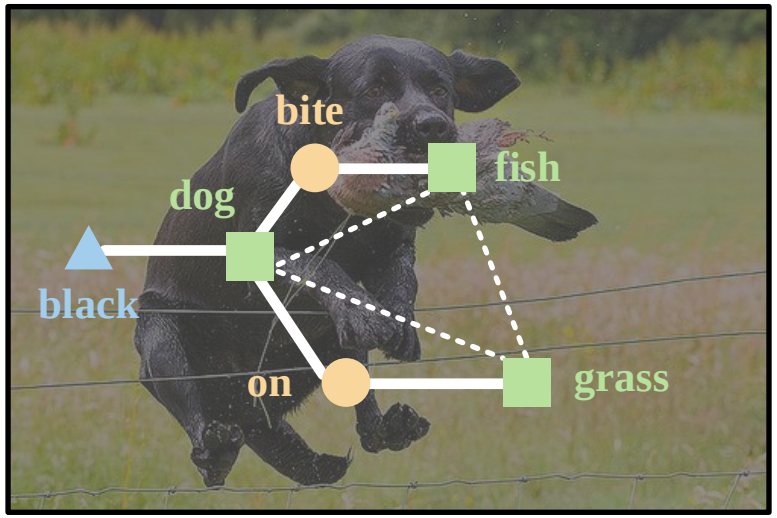
Image captioning

Generating a sentence to describe the image

A black dog biting a fish is running on the grass.

Image captioning with Scene Graphs

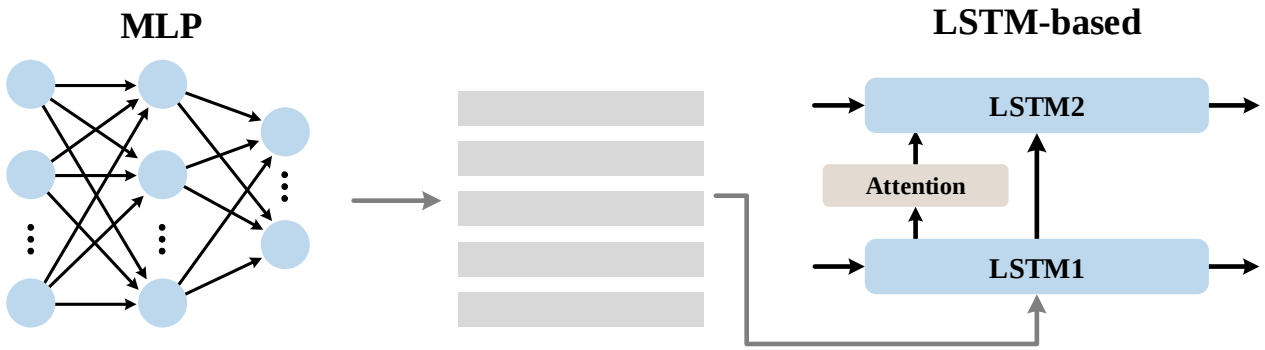
Structured information used to represent **objects**, **attributes**, and **relations** in an image



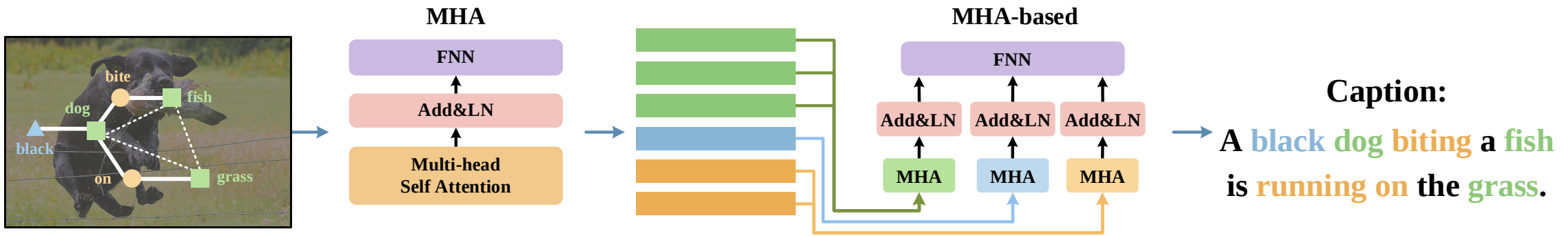
GNN-LSTM framework

Complex Training Strategies | **Heterogeneous** structure **Requires well-chosen training strategies.**

Less descriptive | Embeddings are indiscriminately selected Less descriptive captions

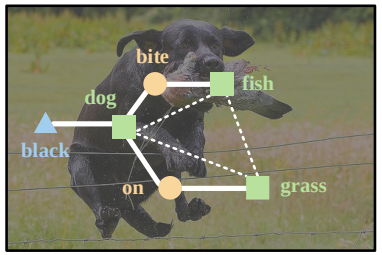


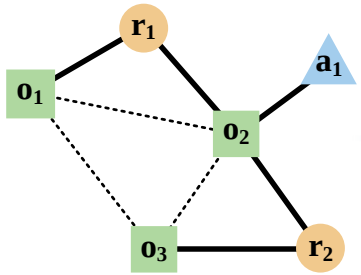
TransForming Scene Graphs to Captions (Ours)

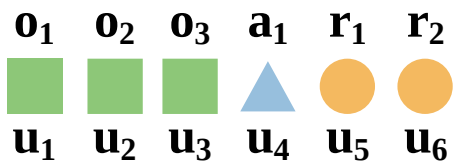


(a) Scene Graph (b) Encoder (c) Graph Embedding (d) Decoder

Input



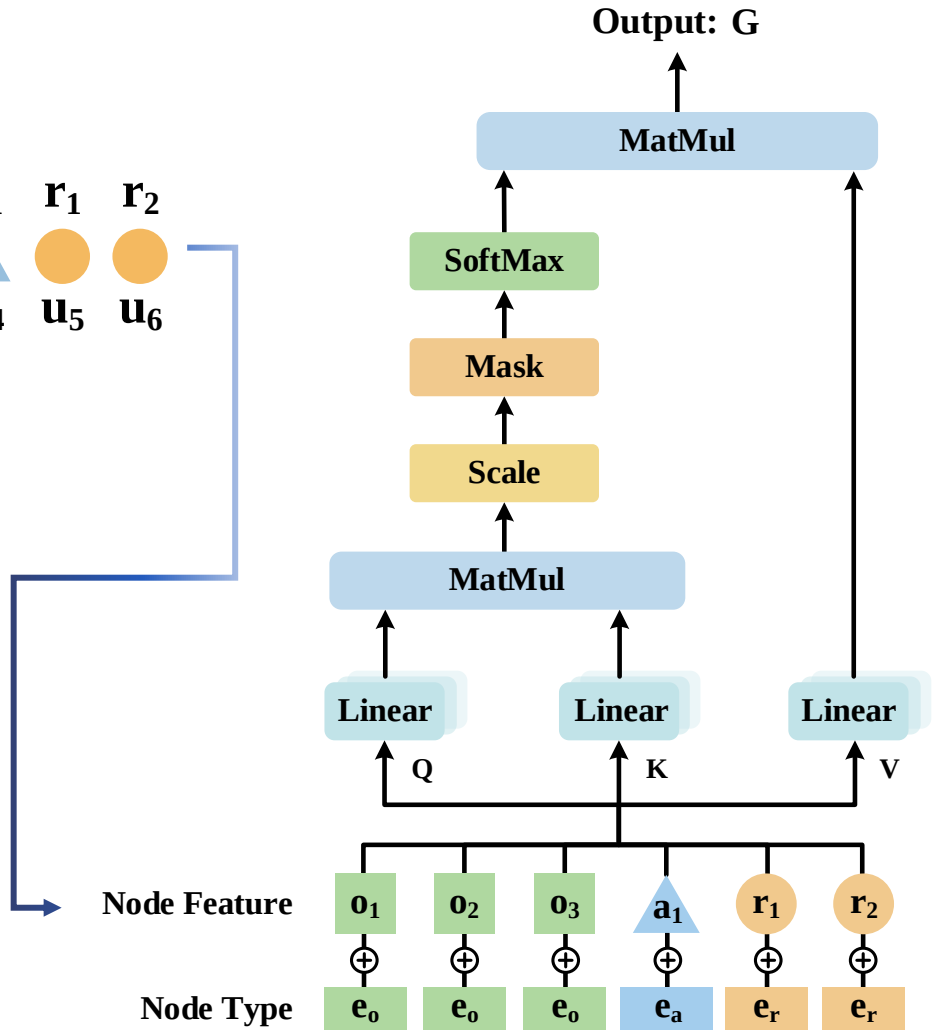




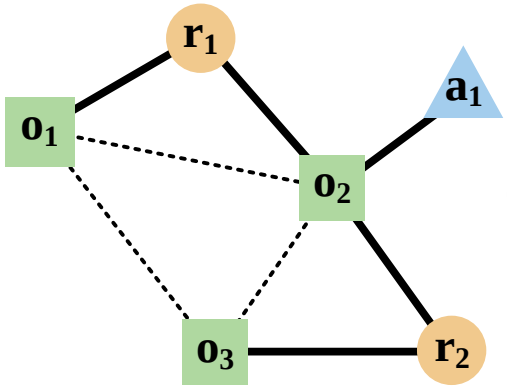
$$\textbf{Object:} \quad \boldsymbol{u}_i = \boldsymbol{o}_i + \boldsymbol{e}_o, \qquad 1 \le i \le N_o,$$

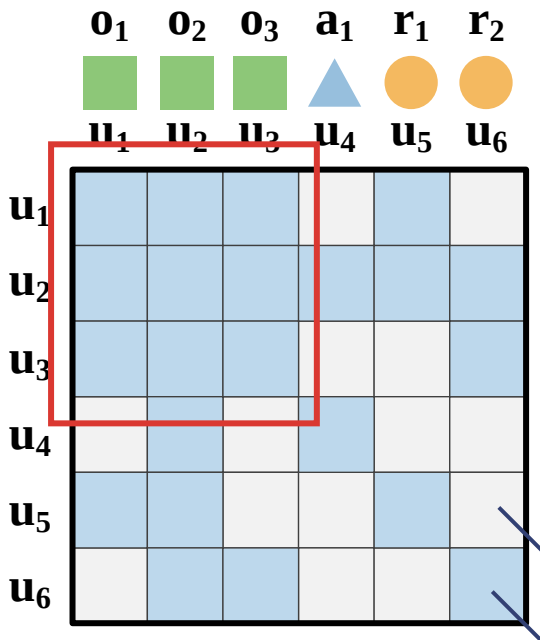
$$\textbf{Attribute:} \quad \boldsymbol{u}_{N_o+i} = \boldsymbol{a}_i + \boldsymbol{e}_a, \qquad 1 \le i \le N_a,$$

$$\textbf{Relation:} \quad \boldsymbol{u}_{N_o+N_a+i} = \boldsymbol{o}_i + \boldsymbol{e}_r, \qquad 1 \le i \le N_r,$$



Encoder

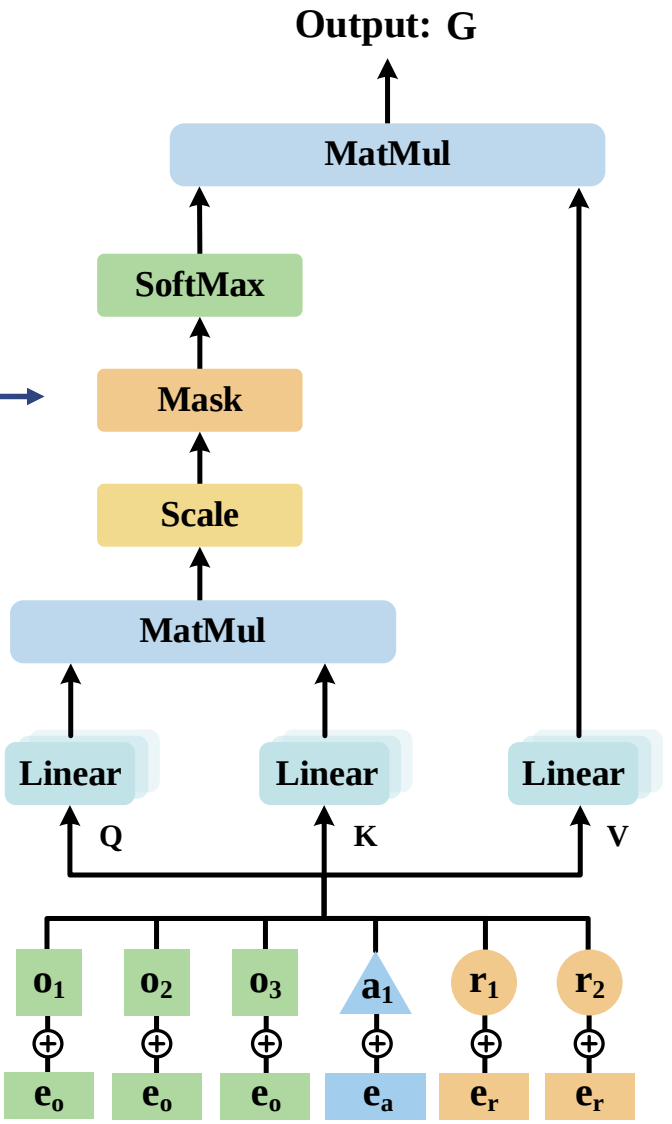




Symmetric Binary Mask Matrix

0

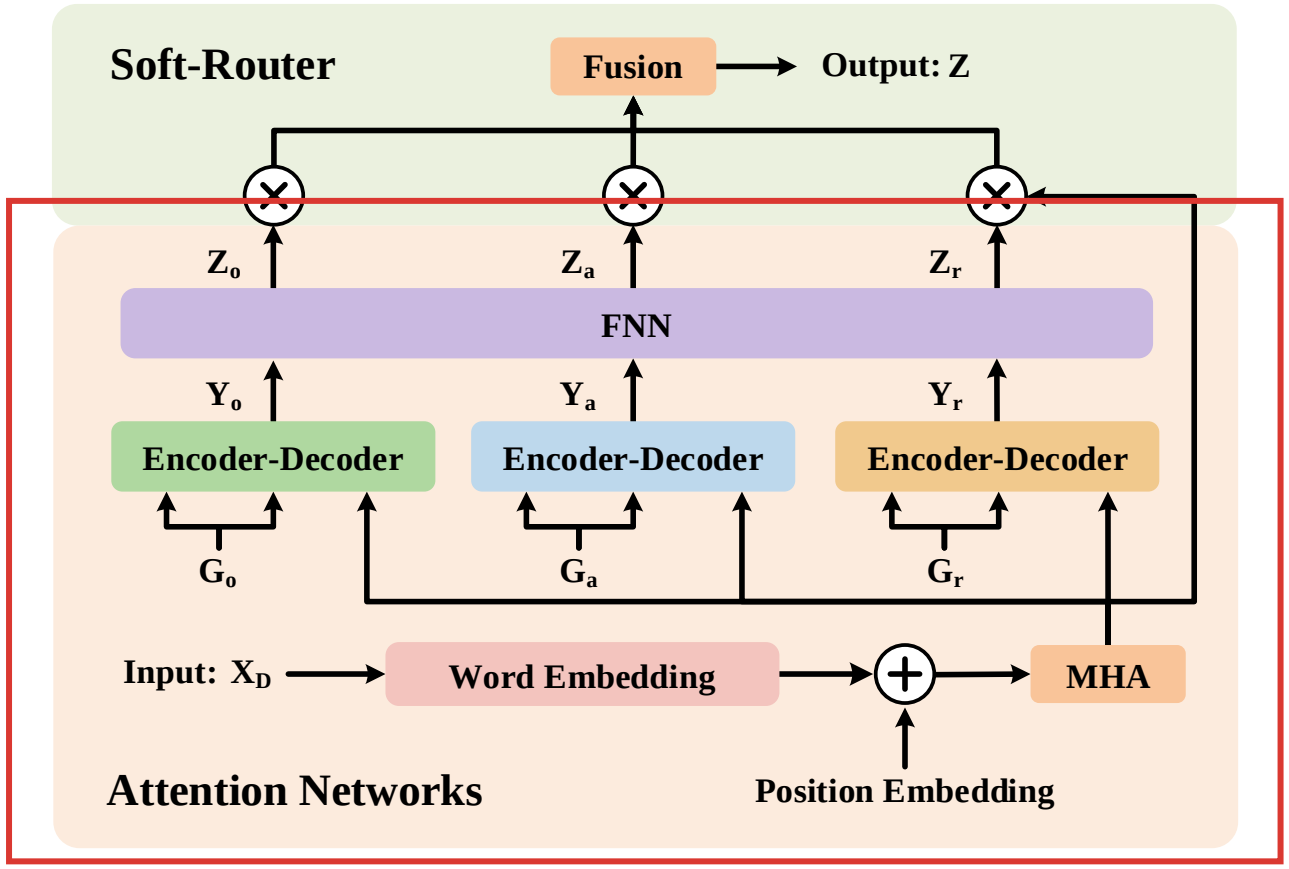
1



Masked:

- Control whether two nodes are connected or not
- Node can be updated by aggregating more neighbour embeddings

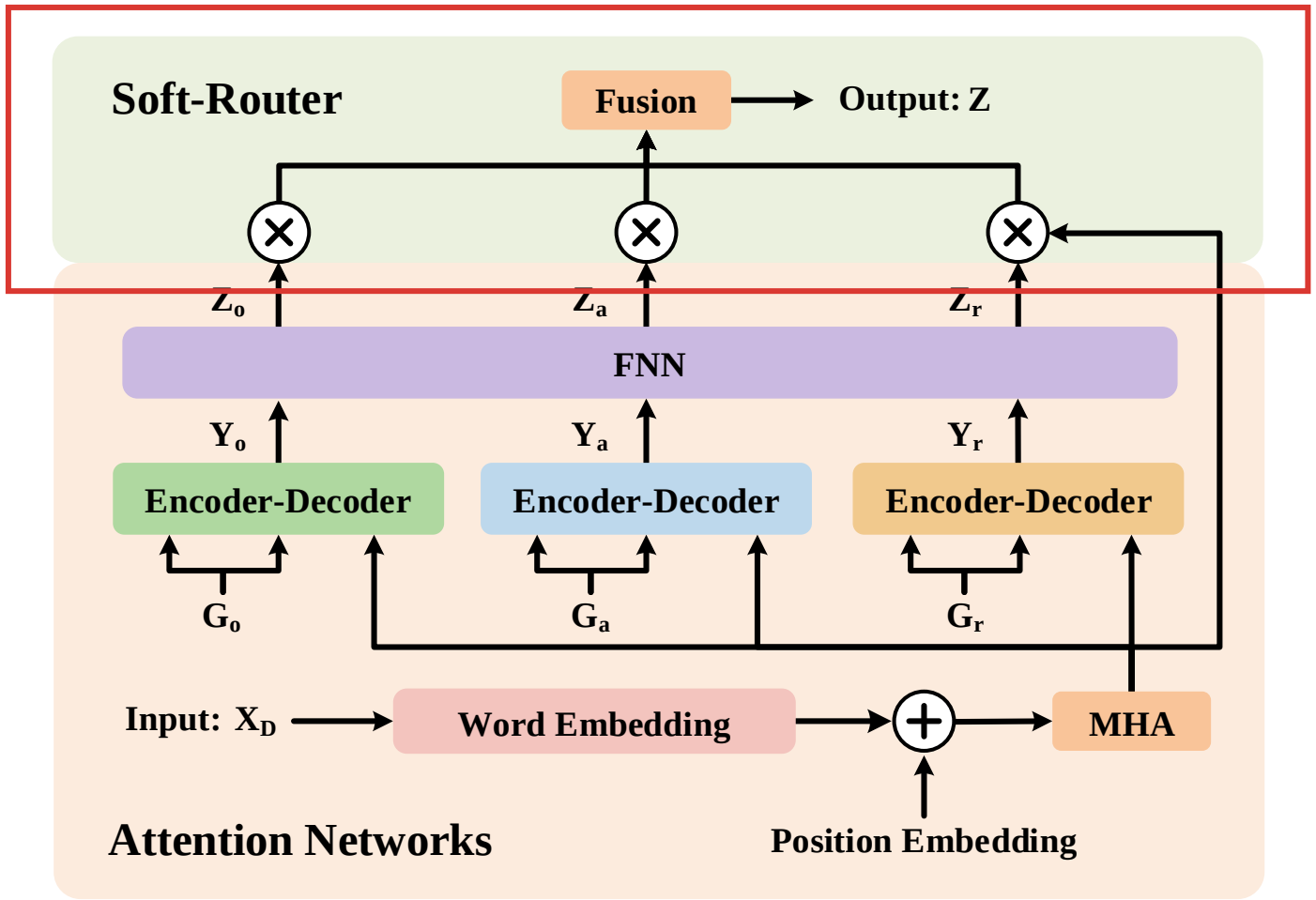
Decoder



$$
\begin{aligned}
\textbf{Input:} \quad & X_D, G_o, G_a, G_r \\
\textbf{SA:} \quad & X = \mathrm{MHA}(Q = K = V = X_D), \\
\textbf{EXP}_O\textbf{:} \quad & Y_o = \mathrm{MHA}(Q = X, K = V = G_o), \\
\textbf{EXP}_A\textbf{:} \quad & Y_a = \mathrm{MHA}(Q = X, K = V = G_a), \\
\textbf{EXP}_R\textbf{:} \quad & Y_r = \mathrm{MHA}(Q = X, K = V = G_r), \\
\textbf{FFN:} \quad & Z_o, Z_a, Z_r = \mathrm{FFN}(Y_o, Y_a, Y_r),
\end{aligned}
$$

The sketch of the MOE-decoder, we use different colors to denote different experts: green/blue/yellow correspond to object/attribute/relation experts.

Output



$$\begin{aligned}\textbf{Input:}\quad & x, z_o, z_a, z_r, \\ \textbf{ATT:}\quad & \alpha = \{\alpha_o, \alpha_a, \alpha_r\} \\ & \quad = \mathrm{Softmax}(\{x^T z_o, x^T z_a, x^T z_r\}) \\ \textbf{Output:}\quad & z = \alpha_o z_o + \alpha_a z_a + \alpha_r z_r,\end{aligned}$$

e.g.:

If the last word of generated caption is an **adjective** “black”, the next word is more like to be a **noun** and thus α_o should be a large value for using more **object embeddings** instead of the other embeddings.

TransForming Visual Scene Graphs to Image Captions

将场景图转换为图像文本

Thank you for listening!

Xu Yang[1] Jiawei Peng[1] Zihua Wang[1] Haiyang Xu[2*] Qinghao Ye[2]
Chenliang Li[2] Songfang Huang[2] Fei Huang[2] Zhangzikang Li[1] Yu Zhang[1*]