We sincerely thank all the reviewers for their insightful suggestions.

1 Ablation Studies The issues raised by the reviewers on ablation studies are very sensible. Actually, we originally did comprehensive ablation studies, but they were omitted due to the space limit. We thought reporting more results on more tasks would be more important than reporting ablation studies. Apparently, we were wrong. We will add them back in the updated version, which will have 1 more page. Those ablation studies were systematically conducted on the LCQMC dataset (a large-scale Chinese question matching corpus).

1.1 Training Strategies In our proposed training strategy (BERT-glyph-joint), we first only fine-tune the BERT model using task-specific supervising signals. Next, we freeze BERT and then update parameters of the Glyph layer. Finally, we relax BERT and fine-tune the two models jointly. Baseline training strategies include (1) *Glyph-Joint*, in which BERT is not fine-tuned at the beginning, i.e., we first freeze BERT to train the glyph layer, and then jointly train both layers until convergence; and (2) the *joint* strategy, in which we directly train the two models together until convergence. Results are shown in Table 1. The proposed training strategy introduces a performance boost of F1 about +1.0 over the others.

1.2 image-classification training objective Table 2 explores the influence of the image-classification training objective. As can be seen, this auxiliary training objective introduces a +0.8 F1 performance boost.

1.3 Structures of the task-specific output layer We change transformers in the task-specific output layer to other structures such as BiLSTMs and CNNs to explore their effects. Results for different models on different tasks are shown in Table 3.

1.4 CNN structures Results for different CNN structures are shown in Table 4. As can be seen, the adoption of tianzige-CNN structure introduces a performance boost of F1 about +1.0.

Strategy	Precision	Recall	F1
BERT-glyph-joint	86.8	91.2	88.8
Glyph-Joint	82.5	94.0	87.9
joint	81.5	95.1	87.8

Table 1: Impact of different training strategies.

Strategy	Precision	Recall	F1
W image-cls	86.8	91.2	88.8
WO image-cls	83.9	93.6	88.4

Table 2: Impact of the image classification objective.

Strategy	Precision	Recall	F1
Transformers	86.8	91.2	88.8
BiLSMTs	81.8	94.9	87.9
CNNs	81.5	94.8	87.6
BiMPM	81.1	94.6	87.3

Table 3: Impact of different structures for the task-specific output layer.

	Precision	Recall	F1
Tianzige-CNN	86.8	91.2	88.8
Kim 2014	85.7	90.4	87.9
Vanilla-CNN	85.3	89.8	87.4
ResNet	84.5	90.8	87.5

Table 4: Impact of different CNN structures.

2.1 First Reviewer

2.1.1 Task details: We appreciate the helpful suggestions. To demonstrate the generalization power of the GLYCE model, we extensively tested our model on a wide range of NLP tasks. Experiments were conducted on 21 datasets across 7 different tasks. We will add all the details of each task in the appendix of the final version.

2.1.2 Training details: we are sorry for the missing training details. Please refer to Section 1.1. We will add these details in the final version.

2.1.3 More details about the glyph CNN itself: sorry for the confusion. The glyph-CNN is detailed in Section 2.2 in the original paper, but we will make it clearer in the updated version.

2.2 Second Reviewer

2.2.1 Appropriateness: Generally, we think that Glyce is a perfect fit for NeurIPS. NeurIPS/NIPS has a long-standing reputation for presenting fundamental deep learning technology or methodology that improved a wide range of NLP tasks, e.g., *Sutskever et al., Sequence to Sequence Learning with Neural Networks, NIPS2014*; *Mikolov et al., Distributed Representations of Words and Phrases and their Compositionality, NIPS2013*. GLYCE is actually along this line of research. It offers a universal methodology to deal with character graph of logographic languages, and achieves SOTA results on 21 datasets across 7 tasks.

2.2.2 Why visual features would help in certain cases: Sorry for the confusion. In logographic languages, the glyph of a character encodes semantic information. The meaning of a character can not only be inferred by its context (external), but also by its own glyph (internal). Glyph information is particularly helpful to model the meaning of rare characters, since there is not much context available to infer their meanings. For example, “鸣”(chirp) is composed of “口”(mouth) and “鸟”(bird), and “淼”(flood) is composed of three “水”(water). We can see that the glyph of a Chinese character is closely related to its meaning.

2.3 Third Reviewer

2.3.1 details of transformers: thank you for the advice. We will include those details in the updated version.

2.3.2 how many scripts in Table 1 are used: sorry for the confusion. We find that using all (i.e., 8) historical scripts is beneficial to all tasks, and thus we use all of them across all tasks.

2.3.3 whether the original BERT is fine-tuned: sorry for the confusion. Please refer to Section 1.1 on this issue. We will add it in the updated version.