Table 3: Non-iterative results on two monolingual datasets compared with MMEA methods where $X\%$ represents the percentage of reference entity alignments used for training.

	Models	FB15K-DB15K			FB15K-YAGO15K		
		H@1	H@10	MRR	H@1	H@10	MRR
20%	MMEA	.265	.541	.357	.234	.480	.317
	EVA∗	.199	.448	.283	.153	.361	.224
	MSNEA∗	.114	.296	.175	.103	.249	.153
	MCLEA∗	.295	.582	.393	.254	.484	.332
	MEAformer	**.417**	**.715**	**.518**	**.327**	**.595**	**.417**
	w/ MERP	**.434**	**.728**	**.534**	**.325**	**.598**	**.416**
50%	MMEA	.417	.703	.512	.403	.645	.486
	EVA∗	.334	.589	.422	.311	.534	.388
	MSNEA∗	.288	.590	.388	.320	.589	.413
	MCLEA∗	.555	.784	.637	.501	.705	.574
	MEAformer	**.619**	**.843**	**.698**	**.560**	**.778**	**.639**
	w/ MERP	**.625**	**.847**	**.704**	**.560**	**.780**	**.640**
80%	MMEA	.590	.869	.685	.598	.839	.682
	EVA∗	.484	.696	.563	.491	.692	.565
	MSNEA∗	.518	.779	.613	.531	.778	.620
	MCLEA∗	.735	.890	.790	.667	.824	.722
	MEAformer	**.765**	**.916**	**.820**	**.703**	**.873**	**.766**
	w/ MERP	**.773**	**.918**	**.825**	**.705**	**.874**	**.768**

Table 4: Iterative results on two monolingual datasets.

	Models	FB15K-DB15K			FB15K-YAGO15K		
		H@1	H@10	MRR	H@1	H@10	MRR
20%	EVA∗	.231	.488	.318	.188	.403	.260
	MSNEA∗	.149	.392	.232	.138	.346	.21
	MCLEA∗	.395	.656	.487	.322	.546	.400
	MEAformer	**.578**	**.812**	**.661**	**.444**	**.692**	**.529**
50%	EVA∗	.364	.606	.449	.325	.560	.404
	MSNEA∗	.358	.656	.459	.376	.646	.472
	MCLEA∗	.620	.832	.696	.563	.751	.631
	MEAformer	**.690**	**.871**	**.755**	**.612**	**.808**	**.682**
80%	EVA∗	.491	.711	.573	.493	.695	.572
	MSNEA∗	.565	.810	.651	.593	.806	.668
	MCLEA∗	.741	.900	.802	.681	.837	.737
	MEAformer	**.784**	**.921**	**.834**	**.724**	**.880**	**.783**

4.1.5 ***Implementation Details.*** Due to the varying experimental datasets and environments across the aforementioned works, we reproduce them with the following settings to ensure fairness and consistency with our model: ***(i)*** All networks have a hidden layer dimension of 300. The total epochs are set to 500 with an optional iterative training strategy applied for another 500 epochs, following [20]. Training strategies, including cosine warm-up schedule (15% steps for LR warm up), early stopping, and gradient accumulation, are adopted. The AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is used with a fixed batch size of 3500. ***(ii)*** To demonstrate the model's stability, following [4, 20], the vision encoders Enc_v are set to ResNet-152 [16] on DBP15K following EVA/MCLEA, where the vision feature dimension d_v is 2048, and set to VGG-16 [26] on FBDB15K/FBYG15K with $d_v = 4096$. ***(iii)*** Following [44], the Bag-of-Words (BoW) is selected for encoding relations (x^r) and attributes (x^a) as fixed-length (e.g., $d_r = 1000$) vectors. We use the pre-trained 300-d GloVe vectors together with the character bigrams for surface representation after applying machine translations for entity names, as described in [24]. ***(iv)*** An alignment editing method is employed to reduce the error accumulation [30].

Specifically, for MEAformer, the intermediate dimension d_{in} in FFN is set to 400. τ is set to 0.1, and the head number N_h in MHCA is set to 1. We eliminate the attribute values for input consistency, and extend MSNEA with iterative training capability.

We claim that the surface information will only be involved in bilingual datasets (DBP15K) following Lin et al., and we exclude those extra entity descriptions referred in minority EA works [33]. Experiments are conducted on RTX 3090Ti GPUs without parallel.

4.2 Overall Results

The results of the bilingual datasets are presented in Table 1 (non-iterative) and Table 2 (iterative), while the results for monolingual datasets are shown in Table 3 (non-iterative) and Table 4 (iterative). It is evident that our model outperforms the baselines across all datasets under all metrics. Particularly, MEAformer surpasses those methods in typical **supervised** scenario by a large margin on Hits@1 of DBP15K (5 ∼ 6%) and FBDB15K/FBYG15K (3 ∼ 11%). Moreover, it consistently gains improvements on other settings compared to the SOTA methods, while maintaining a similar number of learnable parameters (approximately 14M).

It is reasonable that all models have their performance improved when incorporating the **surface** modality, as textual information inherently serves as a strong supervisory signal in all alignment fields [25, 42]. Nevertheless, our model still exceeds those high-performing baselines and increases the current SOTA Hits@1 scores from (.929/.964/.990) to (.948/.977/.991) on *ZH-EN*/*JA-EN*/*FR-EN* datasets within the DBP15K, respectively.

As an important real world application, **unsupervised** MMEA effectively eliminates the need for manual labeling and annotation costs by generating pseudo-labeled seed alignments based on the similarity of features extracted by pre-trained models. Concretely, we follow Lin et al. [20] to take entities' surface (S) or visual (V) features as the unsupervised references. The results are illustrated in Table 1 and 2 with "Unsup." as the indication, where our MEAformer consistently achieves the SOTA results.

We note that the **MERP** policy could further increase model's performance across all datasets but it is not involved in our standard MEAformer, since we aim to keep it separate from the model and establish it as a generic technique for (MM)EA methods. Furthermore, we abandon MERP when training model with iterative or unsupervised strategies, since the generated pseudo seed alignments may not be entirely accurate and could mislead the model, exacerbating the error cascade issue throughout the replay stage.

4.2.1 ***Efficiency Analysis.*** To gain further insights into our model, we examine the efficiency behaviors of four MMEA algorithms on two datasets with identical 3000 epochs and an early stopping strategy. As shown in Figure 3, MEAformer consistently outperforms others throughout the training progress (left), and it also excels in the trade-off between convergence time and performance (right). We attribute MSNEA's subpar behavior to the constraints imposed by translation-based models' geometric models, which limit their ability to capture complex structural relationships among entities for EA. Please see the appendix for details.