

# Supplementary Materials

## 1 MORE CORRECTION EXAMPLES

As illustrated in Table 1, Table 2, Table 3, and Table 4, we present examples of homophone confusion correction in Chinese and closely pronounced words in English within the context of VSDial-question, the Chinese and English version of VSDial-caption, and Flickr8K datasets. These examples demonstrate our approach’s effectiveness in addressing auditory ambiguities inherent in spoken language. The **blue** words represent recognition confusion with the Whisper model, while the **red** indicates the correction results with our CIEASR model.

Table 1: More correction examples in VSDial-question.




Scene Image	Whisper	CIEASR
	起手是男性还是女性 Translation: Is the starter male or female?	骑手是男性还是女性 Translation: Is the rider male or female?
	有数吗 Translation: Are there numbers?	有树吗 Translation: Are there trees?
	他睡着了吗 Translation: Is he asleep?	她睡着了吗 Translation: Is she asleep?

Table 2: More correction examples in the Chinese version of VSDial-caption.

Scene Image	Whisper	CIEASR
	三个骑马的人拿着棍子追足球 Translation: Three people on horseback holding sticks chase a soccer ball.	三个骑马的人拿着棍子追逐球 Translation: Three people on horseback holding sticks chase a ball.
	一只羊站在岩石上站着叫 Translation: A sheep stands on a rocky mountain and stands calling.	一只羊站在岩石上长着角 Translation: A sheep with horns stands on a rocky mountain.
	一名男子将相机指向汽车的测试镜 Translation: A man points the camera at the car's test mirror.	一名男子将相机指向汽车的侧视镜 Translation: A man points the camera at the car's side mirror.

Table 3: More correction examples in the English version of VSDial-caption.




Scene Image	Whisper	CIEASR
	A family has a <b>sanity</b> betrothed to themselves while they fly a kite.	A family has a <b>sandy beach</b> all to themselves while they fly a kite.
	A blue plate with a <b>bottle</b> of bananas in the center.	A blue plate with a <b>bushel</b> of bananas in the center.
	A <b>valerie</b> plant hides small yellow blooms not yet bloomed.	A <b>velvety</b> plant hides small yellow blooms not yet bloomed.

Table 4: More correction examples in Flickr8K.

Scene Image	Whisper	CIEASR
	Four boys posing while one boy sets his <b>dream</b> down.	Four boys posing while one boy sets his <b>drink</b> down.
	Two dogs wearing sweaters <b>playing</b> in the grass.	Two dogs wearing sweaters <b>play</b> in the grass.
	A man with a texas flag stands in the snow by <b>chance</b> .	A man with a texas flag stands in the snow by <b>tents</b> .

## 2 TRAINING DETAILS

Our CIEASR model is trained on the VSDial-question and VSDial-caption dataset for 5 epochs with a batch size of 64, each epoch containing 120k samples. For the Flickr8K dataset, FunASR is employed to filter out erroneous samples with a WER higher than 60%, indicating a significant discrepancy between the speech content

and the labels. This includes misreadings, omissions, and interferences that occurred during the recording of the speech. Then we train our model for 20 epochs with a batch size of 32, each epoch containing 8k samples, and evaluate our CIEASR model and the Whisper model on the filtered Flickr8K dataset.

We use the AdamW optimizer with 0.05 weight decay. We use a linear warmup cosine learning schedule with 2000 warmup steps which begins with a linear warm-up phase at 1e-6, increasing to 1e-4, and then follows a cosine decay pattern towards 1e-5.

We employ mixed precision training, where the frozen visual encoder and Whisper model operate using half-precision (float16) for enhanced computational efficiency. In parallel, the Q-Former and projection layers utilize full precision (float32) to maintain computational stability, thus balancing computational efficiency and model performance fidelity.

We train all models with the devices of  $2 \times \text{A800s}$ . Training on VSDial-question and VSDial-caption is completed within 6 hours. Training on Flickr8K is completed within 2 hours.

Table 5: Hyper-parameters of the CIEASR model.

Hyper-parameters	VSDial-question	VSDial-caption	Flickr8K
epochs	5	5	20
train_batch_size	64	64	32
eval_batch_size	64	64	32
accum_grad_iters	1	1	1
init_lr	1e-4	1e-4	1e-4
min_lr	1e-5	1e-5	1e-5
warmup_lr	1e-6	1e-6	1e-6
warmup_steps	2000	2000	2000
weight_decay	0.05	0.05	0.05
decoder_max_len	50	50	50
num_query_token	32	32	32

Table 6: More correction examples of homophone confusion

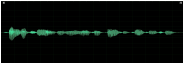

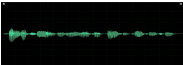

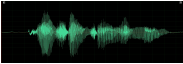



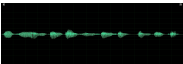

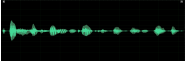





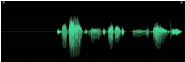



Dataset	Origin Speech	Pure Speech Recognition	Scene Image Cue	CIEASR
VSDial-question		"is this a pen"		"is this a pen"
VSDial-question		"is this a pen"		"is this a pen"
VSDial-question		"is this a pen"		"is this a pen"
VSDial-caption-cn		"is this a pen"		"is this a pen"
VSDial-caption-cn		"is this a pen"		"is this a pen"

Table 7: More correction examples of homophone confusion

Dataset	Origin Speech	Pure Speech Recognition	Scene Image Cue	CIEASR
VSDial-caption-en		"is this a pen"		"is this a pen"
VSDial-caption-en		"is this a pen"		"is this a pen"
Flickr8K		"is this a pen"		"is this a pen"
Flickr8K		"is this a pen"		"is this a pen"
Flickr8K		"is this a pen"		"is this a pen"