A Multimodal Network on Handwritten Chinese Character Error Correction

Haizhao Sun*
sunhaizhao@bupt.edu.cn
Yu Ning*
ningyuv@bupt.edu.cn
Xu Ji
jixv@bupt.edu.cn
Chuang Zhang
zhangchuang@bupt.edu.cn
Ming Wu
wuming@bupt.edu.cn

Beijing University of Posts and Telecommunications
China

Abstract

Handwritten Chinese characters possess complex internal structures and a vast array of categories, making errors highly diverse. Therefore, Handwritten Chinese Character Error Correction (HCCEC) cannot be simply framed as a classification problem, but should be expressed as an open vocabulary question without predefined categories. Beyond visual information, Chinese characters also carry semantic information such as structure and components, which can be represented as Ideographic Description Sequences (IDS). To harness multiple modalities effectively, we adopt a human-inspired approach to error identification, discerning differences in components and structures between erroneous and correct characters. Accordingly, we propose a multi-modal encoder-decoder network incorporating CLIP training methodology. Through pre-training similar to CLIP, the model aligns handwritten characters with their corresponding IDS. The multi-modal decoder deciphers features combining image and semantic information, outputting IDS. With the output IDS, identifying and correcting errors becomes straightforward. The experimental results indicate that our method, as an approach not reliant on pre-defined categories, achieves performance comparable to that of closed-set classification methods with pre-defined categories in the HCCEC task.

1 Introduction

Handwriting Chinese character error correction (HCCEC) is a new problem developed from Handwriting Chinese character recognition (HCCR)[12, 13, 15, 17]. It is often used in Chinese character learning scenarios. For beginners whose native language is not Chinese, and primary school students who are exposed to Chinese writing for the first time, handwritten Chinese character error correction will be difficult for them. Play a key role in learning.

*These authors contributed equally to this work.

Handwritten Chinese character correction refers to the process of evaluating and correcting incorrectly written Chinese characters. The evaluation phase aims to determine whether a given handwritten character contains a writing error, while the correction phase corrects it by predicting the target character the user wants to write.

Therefore, unlike Chinese character recognition tasks, handwritten Chinese character error correction requires processing erroneous characters that are not in the correct set of Chinese characters. Compared with Latin error correction, Chinese characters have many categories and complex internal structures. Many categories are highly similar and only have local differences. Therefore, the misrecognition of handwritten Chinese characters not only has the characteristics of small category differences and difficult recognition in fine-grained classification tasks[6], but also faces the difficulty of transferring attribute knowledge between categories in generalized zero-shot learning tasks, and the inability to predict the categories in advance in open vocabulary classification tasks. It is difficult to identify open sets. In addition, due to the lack of open-source datasets containing incorrectly written characters, training for error correction becomes an almost impossible task.

In order to solve the above problems, some existing HCCR methods propose radical-based and stroke-based methods[3]. These methods are very effective in traditional Chinese character recognition problems, but there are challenges in handwritten Chinese character error correction problems. Because traditional CCR methods usually identify a character as a specific category in the Chinese character vocabulary, even an error character will be recognized as a correct character. However, errors in handwritten Chinese characters may occur in various forms, and it is not possible to predefine every possible error category as a candidate.

While some research teams[8] have attempted to create error character datasets on their own and have categorized common handwritten Chinese character errors into radical-level errors, structure disorders, and stroke-level errors (as shown in Figure 5), these datasets are not yet open source and therefore difficult to be widely used. This current situation limits the development and application of handwritten Chinese character error correction technology.

People who learn Chinese, usually first pay attention to the overall structure of Chinese characters, and then decompose Chinese characters into different radicals based on this structure. This decomposition helps them move through known Chinese character structures and components to new characters they have never seen before. When evaluating the correctness of a Chinese character, people compare each radical to the corresponding radical in a similar-structured character at the same position, similar to the process we use when identifying animals: we would break the animal into its different body parts, and then divide these into Parts are compared with known animals. For example, if we see a lion with a snake's tail, we know it is an incorrect or unconventional creature. This method of comparison and decomposition is very effective in Chinese character learning and can help learners identify and correct errors.

Inspired by humans' recognition of fake Chinese characters, we design a new method that fully utilizes the structural and semantic multi-modal information of Chinese characters. This method starts from the overall structure and gradually analyzes the local details and components of the Chinese characters. By comparing each component of a handwritten Chinese character with the corresponding Chinese character part in the dictionary, we are able to diagnose whether the handwritten Chinese character image is a correct Chinese character or contains an incorrect Chinese character.

This method mimics human visual and cognitive processes, taking into account the relationship between the whole and its parts when analyzing Chinese characters, as well as the

Figure 1: Chimera[1] in mythology, similar to uncommon characters.

specific shapes of different radicals and strokes. Through this meticulous comparative analysis, writing errors can be more accurately identified, especially those subtle differences that may be ignored in conventional Chinese character recognition systems. Such a system is a valuable tool for students learning Chinese characters, helping them more effectively master the correct way of writing and improve the quality of their writing.

Ideographic Description Sequences (IDS), as defined by Unicode, is used to decompose Chinese characters. All training data are Chinese characters composed of known radicals and structures. We can consider that even error characters are composed of known radicals and structures, which makes even categories of error characters that do not appear in the training data can also be effectively handled. Through the transfer of part knowledge, we are able to solve the problems caused by unseen error character categories.

Specifically, our method utilizes a self-attention mechanism to learn the interaction of these visual and textual features by combining the visual features of handwritten Chinese characters with the corresponding IDS. This enables the model to focus on the overall structure and specific areas of the character image respectively based on different positions of the currently predicted IDS. Subsequently, the IDS is gradually generated through the multimodal decoder using next token prediction. This generation method is auto-regressive, addressing the challenge of target IDS belonging to an open vocabulary set. In other words, the predicted IDS is not constrained by any predefined category information and can be directly generated, allowing for novel combinations that were not seen during the training phase and offering limitless possibilities.

Our main contributions are as follows.

- A multimodal learning framework that combines visual and textual information is introduced to enhance the model's understanding of Chinese character structure and semantics and improve error correction accuracy.
- When data is scarce, this method can train a model that can recognize unseen Chinese characters through samples in the dataset, effectively dealing with the problem of data limitations.
- Generative technology is used to identify error characters in text images without predefining error categories, adapting to open vocabulary and diverse error types, enhancing the practicability and adaptability of the model.

2 Related Work

The need for handwritten Chinese character error correction has long existed, but it was only recently that it began to receive attention as an advanced problem in handwritten Chinese character recognition tasks. With the development of deep learning, CNN-based methods such as MCDNN[5] have achieved remarkable success in extracting robust features of Chinese characters, achieving close to human-level performance on the ICDAR 2013 competition handwritten Chinese character recognition task[14]. In addition, the Chinese Academy of Sciences constructed the HWDB[9] dataset for the competition, which was widely used by subsequent researchers.

Ideographic Description Sequence (IDS) is a Chinese character structure description grammar defined by the Unicode standard. It is composed of a description character and two or more specific characters (mainly Chinese characters) to represent the abstract structure of Chinese characters. HCL2020[7] is the first systematic dataset that combines handwritten Chinese characters with radicals and structures, but its main focus is on the style of handwritten Chinese characters.

Yu et al.[16] proposed a Chinese character recognition network that combines the CLIP structure[10], named CCR-CLIP, which is the first work to associate IDS with the CLIP structure. However, CCR-CLIP, as a CCR method, requires predefined categories and prompts during testing, and in the HCCEC problem, there are too many error categories to be predefined. In response to this problem, generateU[4] proposed a method that directly generates relevant categories without relying on any predefined category information, breaking the traditional open vocabulary detection paradigm and proposing the concept of generative detection[11].

3 Method

In this paper, we propose a handwritten Chinese character error correction framework illustrated in Figure 2 that includes multiple modalities. The framework consists of three parts, the contrastive learning pre-training part, the IDS generation part and the error correction part.

In the pre-training part, contrastive learning is employed using correctly handwritten Chinese character images and their corresponding IDS, enhancing the image encoder's ability to extract image features that align with the IDS. The image encoder is composed of 12

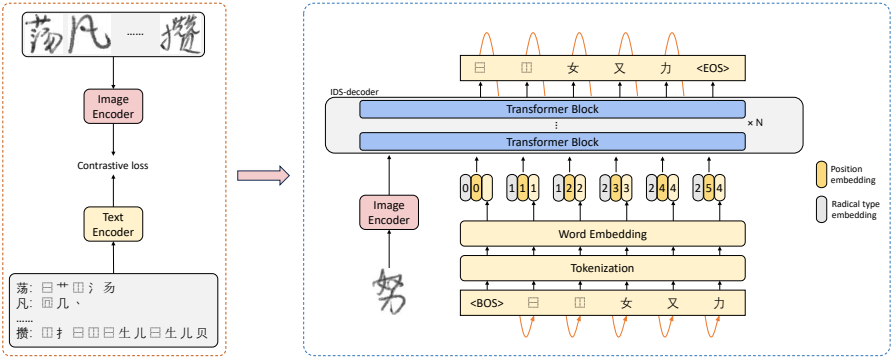


Figure 2: Network architecture of our model. First, an image encoder is obtained through CLIP-like image-text pretraining. Then, the image features are concatenated with IDS tokens for multimodal learning.

layers of ViT and is used to extract the visual features $\mathbf{F}^c \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ of the input handwritten character image. In order to represent features with one-dimensional vectors, we added EOS tokens at the end of the image patch sequence. Since the EOS token can pay attention to the embedding of all previous image patches, it can represent the global feature $\mathbf{f}^c \in \mathbb{R}^{1 \times C}$ of the image. Finally project $\mathbf{f}^c$ to embedded visual feature space $\mathbf{I} = \mathbf{f}^C \times \mathbf{W}^C$ Among them, $\mathbf{I}$ represents the embedded visual features of the input handwritten Chinese character image, $\mathbf{W}^c \in \mathbb{R}^{C \times C'}$ represents the projection matrix, C' is the alignment dimension.

The text encoder extracts radical features and structural features corresponding to IDS. The text encoder consists of 12 layers of Transformer encoders and 1 embedding layer. The IDS sequence is encoded through the text encoder into $\mathbf{f}^r \in \mathbb{R}^{1 \times D}$, similarly , the text encoder also projects features to $\mathbf{T} = \mathbf{f}^r \times \mathbf{W}^r$, $\mathbf{W}^r \in \mathbb{R}^{D \times C'}$ represents the projection matrix. The image features and text features are projected from feature spaces of dimensions C and D, respectively, into a common feature space of dimension C' using different projection matrices. Use contrastive loss $\mathcal{L}_{Contrastive}$ to align image features and IDS features. For a batch with training samples of N, the loss function is as follows:

$$\mathcal{L}_{Contrastive} = -\sum_{j=1}^{N} \log \frac{\exp(\mathbf{I}_j \cdot \mathbf{T}_j)}{\sum_{n=1}^{N} \exp(\mathbf{I}_j \cdot \mathbf{T}_n)} - \sum_{j=1}^{N} \log \frac{\exp(\mathbf{I}_j \cdot \mathbf{T}_j)}{\sum_{n=1}^{N} \exp(\mathbf{I}_n \cdot \mathbf{T}_j)}$$

$\mathbf{I}^j$, $\mathbf{T}^j$ respectively represent the embedded visual features and IDS features of the jth sample in the batch. In order to adapt to the problem that there are the same Chinese character samples in the same batch, align the labels with the first occurrence of the corresponding character during the calculation.

In the IDS generation part, the image encoder in the pre-training part is used. For each Chinese character image-IDS pair, $\mathbf{I}$ is used to represent the image, T_i represents IDS tokens, T_0 is the <BOS> token, T_{N+1} is the <EOS> token. Training the model using the loss of a multimodal language model:

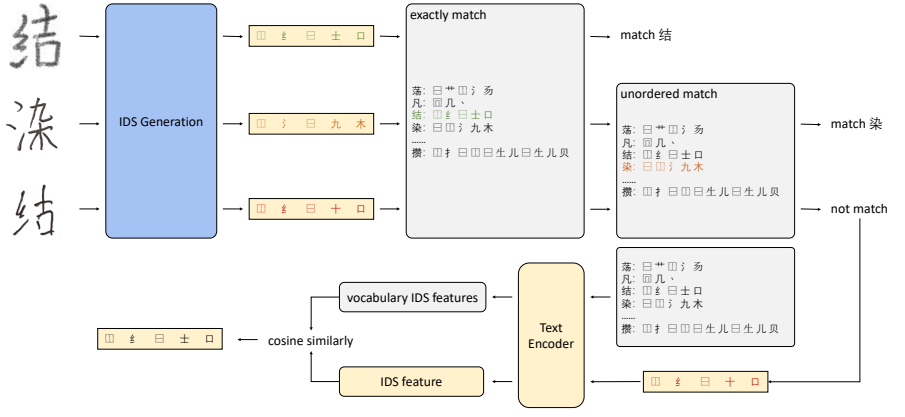


Figure 3: Correction workflow.

$$L_{MLM} = \frac{1}{N+1} \sum_{i=1}^{N+1} \mathrm{CE}(T_i, p(T_i|\mathbf{I}))$$

where CE is cross entropy loss.

In the error correction part, as shown in Figure 3, after obtaining the generated IDS, first accurately match the existing sequences in the dictionary. If the match fails, perform unordered matching. If the match still fails, use the pre-trained text encoder to extract IDS features. And calculate cosine similarity with the features of all sequences in the dictionary to get the ideal character corresponding to the most similar sequence.

4 Experiments

4.1 Dataset

Given that open-source error character datasets are extremely scarce, we analyzed the error patterns of incorrect Chinese characters and found that most of these errors stem from improper use of radicals and confusing structures. This phenomenon is very similar to the appearance of rare characters, which are rarely used in daily life and many people do not even recognize these characters. Considering that our target users are beginners of Chinese characters when beginners write an uncommon character that almost no one knows, it is most likely a writing error. Therefore, we decided to use the rare character handwriting dataset as an alternative to the error character handwriting dataset to test our system.

The HWDB[9] dataset was used in the Chinese handwriting recognition competition. HWDB1.0 contains 3,866 categories of Chinese characters, the vast majority of which (3,740 categories) are included in the commonly used 3,755 categories as specified by the Chinese national standard GB 18030-2022*, and HWDB1.1 contains 3,755 categories, all of which are in commonly used Chinese characters. HWDB1.2 contains 3,319 categories of Chinese characters, which do not intersect with commonly used Chinese characters.

*https://en.wikipedia.org/wiki/GB_18030

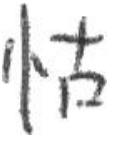
real character	fake character	uncommon character
		
⿰纟⿱士口	⿰纟⿱十口	⿰忄⿱十口

Figure 4: The left side is the correct word, the middle is the wrong word, and the right side is the uncommon word. It can be seen that the rare words and the wrong words have similar errors compared to the correct words.

The test set uses the test set of HWDB1.2, which removes letters, numbers and symbols, and only retains Chinese characters. Since Chinese character error correction tasks are usually aimed at beginners, and the written characters are basically common Chinese characters, when the error character dataset is difficult to obtain, rare characters that have not been seen in the training stage are used to replace the error character for testing.

4.2 Implementation Details

The method in this article is implemented using PyTorch, and all experiments are performed on an NVIDIA RTX 3090 GPU with 24 GB of memory. The Adam optimizer is used to train the model with an initial learning rate of 5e-5, and then follows linear decay, with the momentum $\beta 1$ and $\beta 2$ set to 0.9 and 0.999 respectively. The batch size was set to 64, and training was conducted for 1 epoch, after which the model reached saturation. The image encoder is initialized from a pre-trained contrastive model[16]. The multi-modal decoder consists of 6 randomly initialized Transformer blocks with 12 heads and the hidden dimension is set to 768.

4.3 Evaluation Metrics

Decomposition accuracy (DACC) is a more precise measure to evaluate subtasks. It evaluates the model's ability to decompose a given sample into corresponding IDS sequences. In this article, the correct set and error set are used to calculate DACC respectively. The definition of DACC is similar to line accuracy in line-level text recognition. The number of characters whose entire IDS perfectly matches the ground truth is denoted as $D_{correct}$, and the total number of characters is denoted as N.

$$DACC = \frac{D_{correct}}{N}$$

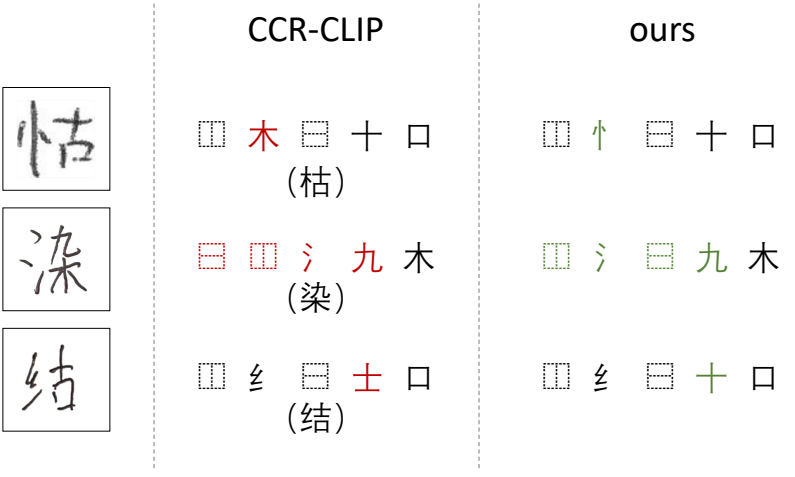


Figure 5: Comparison of the prediction results of CCR-CLIP without candidates with the prediction results of this method.

Method	error set	correct set
	DACC	DACC
HDE[2]	33.5%	92.1%
CCR-CLIP(-)	-	94.6%
Ours	31.5%	89.2%

Table 1: Results. The top row is the error character recognition results with the predefined category method, and the bottom row is the result without the predefined category method. (-) refers to removing the original predefined categories in the method.

5 Discussions

5.1 Results

As a HCCR method, CCR-CLIP cannot perform error character recognition without predefined categories shown in Table 1. However, the effect of our method is comparable to the HDE method with predefined categories on the error set. Our method is based on generation. The traditional correct set classification task does not lag behind the closed set classification method much. Compared to traditional character recognition methods, this approach allows for a more fundamental understanding of character structures and radicals. As a result, the model can accurately decompose unseen characters that have similar structures to the correct characters, even when encountering radicals with similar structures.

5.2 Visualization

In the comparison in Figure 5, we can clearly see that CCR-CLIP is more inclined to reproduce characters that have appeared in the training data during the generation process, thus producing an illusory effect to some extent. Our method, on the other hand, focuses more on the structure of characters and the accurate generation of radicals, demonstrating a high degree of mastery of details and improved adaptability to new characters. This shows that our technology can better maintain structural integrity and recognition accuracy when dealing with unseen characters, thus showing greater adaptability and reliability in practical applications. This is especially important when doing Chinese character recognition and generation, because the complexity of Chinese characters requires the model to not only remember the characters that have been learned, but also creatively recombine radicals and strokes to form new characters.

6 Conclusion

In this paper, we propose a multimodal handwritten Chinese character error correction network, which is inspired by the way humans recognize incorrect Chinese characters and focuses errors in Chinese characters on Chinese character structure and radicals. The model obtains the ability to decompose Chinese characters into corresponding IDS through a CLIP model and then uses an image encoder with this ability to combine three embeddings of characters, structures, and radicals to generate accurate error character IDS through a multimodal decoder. And use the generated results to predict the ideal characters corresponding to the error character.

This method overcomes the problem of sparse Chinese character error character data and the absence of open-source handwritten Chinese character error character datasets. Resolved the difficulty that error character categories cannot be predefined.

References

[1] Tommy Adapto. Chimera. Online. URL https://www.deviantart.com/draethius/art/Chimera-346202862. Available from: [Website Name] (Accessed: Month Day, Year).

[2] Zhong Cao, Jiang Lu, Sen Cui, and Changshui Zhang. Zero-shot handwritten chinese character recognition with hierarchical decomposition embedding. *Pattern Recognition*, 107:107488, 2020. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107488. URL https://www.sciencedirect.com/science/article/pii/S0031320320302910.

[3] Jingye Chen, Bin Li, and Xiangyang Xue. Zero-shot chinese character recognition with stroke-level decomposition, 2021. URL https://arxiv.org/abs/2106.11613.

[4] Lin Chuang, Jiang Yi, Qu Lizhen, Yuan Zehuan, and Cai Jianfei. Generative region-language pretraining for open-ended object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] Dan Cirean and Ueli Meier. Multi-column deep neural networks for offline handwritten chinese character classification. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2015. doi: 10.1109/IJCNN.2015.7280516.

[6] Jingcai Guo, Zhijie Rao, Zhi Chen, Jingren Zhou, and Dacheng Tao. Fine-grained zero-shot learning: Advances, challenges, and prospects. *arXiv preprint arXiv:2401.17766*, 2024. URL https://arxiv.org/abs/2401.17766.

[7] Peiyi Hu, Mengqiu Xu, Ming Wu, Guang Chen, and Chuang Zhang. Handwritten style recognition for chinese characters on hcl2020 dataset. In *Pattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16–18, 2020, Proceedings, Part II 3*, pages 138–150. Springer, 2020.

[8] Yunqing Li, Jun Du, Jianshu Zhang, and Changjie Wu. A tree-structure analysis network on handwritten chinese character error correction. *IEEE Transactions on Multimedia*, 25:3615–3627, 2023. doi: 10.1109/TMM.2022.3163517.

[9] Cheng-Lin Liu, Fei Yin, Qiu-Feng Wang, and Da-Han Wang. Icdar 2011 chinese handwriting recognition competition. In *2011 International Conference on Document Analysis and Recognition*, pages 1464–1469, 2011. doi: 10.1109/ICDAR.2011.291.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[11] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *T-PAMI*, 2024.

[12] Canyu Xie, Songxuan Lai, Qianying Liao, and Lianwen Jin. High performance offline handwritten chinese text recognition with a new data preprocessing and augmentation pipeline. In Xiang Bai, Dimosthenis Karatzas, and Daniel Lopresti, editors, *Document Analysis Systems*, pages 45–59, Cham, 2020. Springer International Publishing. ISBN 978-3-030-57058-3.

[13] Chen Yang, Qing Wang, Jun Du, Jianshu Zhang, Changjie Wu, and Jiaming Wang. A transformer-based radical analysis network for chinese character recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3714–3719, 2021. doi: 10.1109/ICPR48806.2021.9412439.

[14] Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Icdar 2013 chinese handwriting recognition competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1464–1470, 2013. doi: 10.1109/ICDAR.2013.218.

[15] Haiyang Yu, Jingye Chen, Bin Li, and Xiangyang Xue. Chinese character recognition with radical-structured stroke trees, 2022. URL https://arxiv.org/abs/2211.13518.

[16] Haiyang Yu, Xiaocong Wang, Bin Li, and Xiangyang Xue. Chinese text recognition with a pre-trained clip-like model through image-ids aligning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11943–11952, October 2023.

[17] Xinyan Zu, Haiyang Yu, Bin Li, and Xiangyang Xue. Chinese character recognition with augmented character profile matching. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 60946102, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3547827. URL https://doi.org/10.1145/3503161.3547827.