Multi-Modal Multi-Label Classification under Missing Textual/Visual Modalities via Cross-Modal Feature Simulation*

Hassan Ismkhan[0000−0002−8351−3770] and Hamid Bouchachia[0000−0002−1980−5517]

the Department of Computing and Informatics, Faculty of Science and Technology at Bournemouth University
{hismkhan,hbouchachia}@bournemouth.ac.uk

Abstract. Multi-modal movie genre classification relies heavily on textual modalities, such as plot summaries, which provide rich semantic cues essential for accurate multi-label prediction. However, textual modalities are often incomplete, noisy, or entirely missing in real-world scenarios—a critical challenge for text-oriented applications where textual input plays a central role. We propose a cross-modal feature simulation framework designed to enhance robustness against missing textual modalities (while symmetrically handling missing visual modalities). Lightweight bidirectional simulators are pre-trained to reconstruct textual embeddings from available visual features, and visual embeddings from textual ones—using reconstruction losses on paired samples. During both training and inference, absent textual embeddings are dynamically imputed from visual cues (and vice versa), ensuring complete modality representations. The resulting embeddings—real or imputed—are fused via multi-head self-attention and fed to a multi-label classification head. Evaluated on the MM-IMDb benchmark across various missing-modality configurations, including severe textual absence, our approach demonstrates exceptional resilience, effectively transferring visual information into semantically consistent textual representations at the feature level without requiring raw text generation. This textual-focused, encoder-agnostic strategy achieves near-zero performance degradation even when textual modalities are unavailable, making it particularly suitable for text-scarce multi-modal environments. The source-code is available here: https://github.com/h-ismkhan/Multi-Modal-Multi-Label-Classification-under-Missing-Textual-Visual-Modalities.

Keywords: Textual Data · Missing Modality · Multi-Label Classification.

1 Introduction

Movie genre classification is a fundamental task in multimedia information retrieval and recommendation systems, enabling effective organization, search,

* Supported by EU, PRESERVE Project.

and personalization of vast film collections. With the proliferation of online platforms hosting user-generated content and official movie databases, automatic genre prediction from multi-modal data—combining visual elements such as posters and textual elements such as plot summaries—has attracted considerable research interest [1][?]. Traditional approaches relied on unimodal features, often focusing on textual metadata like plots or reviews for classification [6, ?]. However, these methods overlook the complementary information provided by visual cues, such as movie posters, which convey stylistic, thematic, and atmospheric signals that are particularly useful for genres like horror, action, or comedy [18]. The advent of multi-modal learning has addressed this limitation by jointly modeling visual and textual modalities, leading to significant performance improvements on benchmarks like MM-IMDb [19]. Early multi-modal frameworks employed gated mechanisms or concatenation-based fusion to integrate features from separate encoders [19]. More recently, pre-trained vision-language models have dominated the field, leveraging large-scale contrastive learning to align image and text representations in a shared embedding space, achieving state-of-the-art results in multi-modal movie genre classification [1][?]. Despite these advances, a critical challenge remains: real-world scenarios frequently involve incomplete data, where one modality—most commonly the textual modality (e.g., missing or unavailable plot summaries)—is absent [18][?]. This issue is particularly prevalent in text-oriented applications, such as legacy archives, newly released films without descriptions, or user-uploaded posters lacking accompanying text. Standard multi-modal models suffer substantial degradation under such missing textual modalities, as they typically rely on zero-filling or masking, which disrupts the learned cross-modal alignments [18]. Recent works have explored strategies for handling missing modalities, including knowledge distillation, generative imputation, or modality dropout during training [27]. However, many of these approaches either require expensive generation of raw text or fail to fully exploit cross-modal transfer when textual input is severely limited [16]. To bridge this gap, we propose a cross-modal feature simulation framework that specifically targets robustness to missing textual modalities by learning to impute textual embeddings from visual features at the representation level. Our method ensures seamless handling of incomplete textual data during both training and inference, maintaining high performance in text-scarce multi-modal environments. The rest of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the proposed method, Section 4 presents the experimental setup and results, and Section 5 concludes the paper.

2 Related Works

Multi-modal classification has emerged as a key area in computer vision and machine learning, leveraging complementary information from multiple modalities such as images, text, and audio to enhance performance on tasks like sentiment analysis, emotion recognition, and genre classification. Early approaches focused on unimodal representations but evolved to incorporate cross-modal interactions.

For instance, in medical image modality classification, combining visual descriptors (e.g., local binary patterns, scale-invariant feature transform) with textual annotations significantly improves accuracy, as visual features capture structural variations while text provides semantic context [9]. Similarly, in crisis management, attention-based models fuse textual tweets and visual images to classify disaster-related content as informative or non-informative, highlighting the role of cross-modal attention in handling noisy social media data [1] [12]. Hateful meme classification further demonstrates the need for multi-modal reasoning, where prompt-based methods using pre-trained language models like RoBERTa exploit implicit knowledge from visual-textual pairs, achieving high AUC scores without altering network structures [6]. In depression detection, additive cross-modal attention networks integrate audio and textual features, outperforming unimodal baselines by capturing inter-modal relationships [10]. Hierarchical fusion techniques, such as cross-modal complementary networks, model intra- and inter-modal interactions for sentiment classification, outperforming state-of-the-art methods on public datasets [19]. Expanding large pre-trained unimodal models like DenseNet or BERT with multi-modal information injection plug-ins enables fine-tuning for image-text recognition, balancing intra-modal processing and inter-modal interactions [16]. Scene-text-based fine-grained classification employs graph convolutional networks to reason over visual and textual cues, enriching features for tasks like storefront or bottle categorization [18]. Emotion classification refines this by multi-level semantic reasoning networks, exploring object-word and regional-global relations for fine-grained predictions [27]. Multi-modal sarcasm and humor detection in code-mixed conversations uses hierarchical attention to fuse textual and contextual modalities, improving performance on benchmarks [5]. Missing-modality-aware healthcare models, such as knowledge distillation frameworks, adapt to incomplete data in segmentation and classification tasks [24]. Document understanding benefits from vision-language contrastive pre-training, aligning modalities for classification despite heterogeneity [4]. Protein sequence learning with textual descriptions via contrastive objectives enables zero-shot function prediction [23]. A critical challenge in multi-modal learning is handling missing modalities, which can occur due to privacy, device failures, or data incompleteness, leading to performance degradation in real-world applications. Synthesis-based methods reconstruct missing modalities from available ones, such as using generative adversarial networks for MRI sequences [3]. Knowledge distillation approaches transfer robust representations from complete-modality teachers to modality-agnostic students, improving segmentation in brain tumor datasets [22]. Common latent space models map modalities to shared subspaces, enabling imputation and alignment under missing conditions [25]. Prompt-based adaptations mitigate missing modalities in transformers by learning modality-agnostic prompts, reducing fine-tuning costs while maintaining robustness [14]. Transformer robustness studies reveal sensitivity to missing data, with optimal fusion strategies varying by dataset; principle-based searches for fusion improve performance across benchmarks like MM-IMDb [17]. Modality-agnostic person re-identification unifies text and sketch queries via dual-encoders and task-aware training, enhancing generalization to uncertain inputs [7]. Rethinking missing modality from a decoding perspective proposes generative strategies to hallucinate absent features [11]. Fake news detection integrates multi-grained textual and visual features, using attention to fuse modalities robustly [26]. Structured text understanding employs multi-modal transformers for entity recognition, handling layout and visual cues [15]. Cross-modal distillation with audio-text fusion via transformers like Wav2Vec and BERT enables fine-grained emotion classification, addressing data scarcity [13]. Multi-modal prompting explores visual-textual baselines for attribute recognition, leveraging textual annotations to mine correlations [8][21]. While these works advance multi-modal handling of missing data, they often focus on specific tasks (e.g., medical segmentation [9][3] or sentiment analysis [19][27]) or assume modality completeness during training [6][10]. In contrast, our cross-modal feature simulation framework targets movie genre classification, emphasizing robustness to missing textual modalities through bidirectional embedding reconstruction and attention fusion, extending prior imputation strategies [25][22] to vision-language settings without raw generation or heavy retraining.

3 Cross-modal feature simulation under missing textual and visual modalities

The MM-IMDb dataset [2] is a large-scale multi-modal benchmark comprising thousands of movie entries, each featuring a poster image as the visual modality, a plot summary as the textual modality, and multi-label annotations across multiple genres. This dataset is characterized by significant class imbalance and multi-label complexity, making it an ideal testbed for evaluating multi-modal learning approaches, especially those handling incomplete data. Pre-trained vision-language models such as CLIP [20] have demonstrated remarkable performance on tasks requiring alignment between images and text, owing to their joint embedding space learned through large-scale contrastive pre-training. However, these models assume the availability of both modalities at inference time. In practical applications—especially those involving textual content—plot summaries may be incomplete, noisy, or entirely absent (e.g., for newly released films, user-uploaded posters without descriptions, or legacy archives). Such missing textual modalities pose a critical challenge for multi-modal systems, as directly masking or zeroing the textual input often leads to substantial performance degradation. To mitigate this issue, we propose a cross-modal feature simulation framework that explicitly learns to impute missing embeddings in CLIP's shared representation space, both during training and inference. Rather than generating raw text (which is computationally expensive and prone to hallucinations), our approach operates at the feature level, reconstructing semantically consistent textual embeddings from available visual information—and symmetrically for missing visual information—ensuring robust handling of incomplete data throughout the learning process. Formally, let M_I and M_T denote the frozen image and text encoders of CLIP (ViT-B/32 backbone), mapping inputs to a joint embedding space. For a sample i with image I_i and plot text T_i, the real embeddings are:

$$\mathbf{v}_i = M_I(I_i), \quad \mathbf{t}_i = M_T(T_i).$$

We introduce two lightweight bidirectional simulation modules:

- $H_{T\to I}$: a three-layer MLP that predicts visual embeddings from textual ones,
- $H_{I\to T}$: a symmetric MLP that predicts textual embeddings from visual ones.

In the simulation pre-training stage, conducted only on samples where both modalities are available, we optimize these simulators via mean squared error (MSE) reconstruction losses:

$$\mathcal{L}_{sim} = E_{i\sim\mathcal{B}}\left[\|H_{T\to I}(\mathbf{t}_i) - \mathbf{v}_i\|^2 + \|H_{I\to T}(\mathbf{v}_i) - \mathbf{t}_i\|^2\right],$$

where $\mathcal{B}$ denotes the subset with both modalities. This bidirectional objective encourages the simulators to capture fine-grained cross-modal alignments inherent in CLIP's pre-trained space, enabling effective imputation of missing features. During task-specific training and inference, for any sample missing a modality—whether textual or visual—the absent embedding is dynamically simulated using the available one. Specifically, a missing textual embedding is imputed as $\hat{\mathbf{t}}_i = H_{I\to T}(\mathbf{v}_i)$, while a missing visual embedding is imputed as $\hat{\mathbf{v}}_i = H_{T\to I}(\mathbf{t}_i)$. This imputation occurs seamlessly in the forward pass, allowing the model to utilize complete feature pairs even in the presence of missing data during both classifier training and evaluation. The (potentially imputed) pair of embeddings is then processed by a multi-head self-attention fusion module, which treats the two modality embeddings as a short sequence (length 2) and applies cross-attention to produce a fused representation. This fused vector is finally passed through a shallow multi-label classification head using binary cross-entropy loss. The key advantages of our textual-focused approach are:

- **Robustness to missing textual input**: By reconstructing textual embeddings directly from visual posters during both training and inference, the model maintains rich semantic information even when plot summaries are unavailable—a common scenario in text-oriented multimedia applications.
- **Efficiency**: Feature-level simulation avoids costly text generation or retrieval steps, making it suitable for large-scale deployment.
- **Leveraging pre-trained alignment**: Operating within CLIP's powerful joint space preserves cross-modal semantics without requiring end-to-end fine-tuning of the large backbone.
- **Two-stage training**: First optimizing the simulators on reconstruction, then freezing them and fine-tuning only the fusion and classifier—while applying imputation throughout—ensures stable and targeted adaptation to the downstream genre task.

This framework is particularly well-suited for textual-related conferences and journals, as it directly addresses the challenge of incomplete textual modalities in multi-modal learning, enhancing system reliability in text-scarce environments.

4 Experimental setup, dataset, and results

We utilize the official train/dev/test splits of the MM-IMDb dataset. Features are extracted using the frozen CLIP ViT-B/32 model. To simulate missing modalities, we employ deterministic masking with a fixed seed, testing multiple configurations: full availability (100% image + 100% text), partial text absence (100% image + 80/60/40/20% text), partial image absence (80/60/40/20% image + 100% text), and a complex mixed setting (20% both + 40% image-only + 40% text-only). Training follows a two-stage paradigm: (1) simulation pre-training (5 epochs) optimizing only the simulator MLPs with Adam (lr=5×10^{-5}), followed by (2) task training (15 epochs) optimizing the fusion and classification modules on the multi-label genre task. Batch size is 8 throughout. Performance is measured using micro-F1, macro-F1, and mean average precision (mAP), with the best model selected based on dev set micro-F1. Detailed quantitative results and ablation studies are presented in Section 4.

MM-IMDb. The Multi-modal IMDb dataset comprises image (movie posters) and text (plot summaries) modalities for movie genre classification across 27 distinct genres including Drama, Comedy, Thriller, and Romance. Each sample can be associated with multiple genres simultaneously, making this a challenging multi-label classification task. To simulate missing modalities, we configure datasets with various availability patterns:

- *100_image_20_text*: 100% image with 20% text,
- *20_image_100_text*: 20% image with 100% text,
- *complex_20_40_40*: 20% both-modalities-presence, 40% image only and 40% text only,
- *complex_30_35_35*: 30% both-modalities-presence, 35% image only and 35% text only,
- *complex_10_45_45*: 10% both-modalities-presence, 45% image only and 45% text only.

Table 1 presents the F-micro scores achieved by the compared methods across different modality configurations on the MM-IMDb dataset. The baseline methods—ShaSpec, M3Care, and PmcmFL—exhibit varying performance under full and missing modality settings. ShaSpec achieves an F-micro of 0.49 on the complete dataset (100_image_100_text) but drops to 0.39–0.45 in missing scenarios, indicating moderate robustness. M3Care performs slightly lower at 0.46 on full data, with scores ranging from 0.37 to 0.45 in missing cases, showing similar sensitivity to modality absence. PmcmFL starts stronger at 0.66 on complete data but degrades to 0.48–0.54 when modalities are missing, suggesting it benefits more from full availability but struggles with imputation. In contrast, our SIM-C method consistently achieves the highest scores, with 0.94 on full data and 0.93 across all missing configurations. This near-perfect stability highlights SIM-C's superior ability to simulate and compensate for absent modalities, outperforming baselines by a wide margin (e.g., 42.42% improvement over PmcmFL on full data). On average, across missing configurations, SIM-C attains an F-micro of 0.93, compared to 0.41 for ShaSpec, 0.41 for M3Care, and 0.51 for PmcmFL. This

Table 1: F-micro obtained by the algorithms on MM-IMDB

Method	100_image_100_text	100_image_20_text	20_image_100_text	complex_10_45_45	complex_20_40_40	complex_30_35_35
ShaSpec	0.49	0.39	0.45	0.39	0.38	0.43
M3Care	0.46	0.39	0.43	0.43	0.37	0.45
PmcmFL	0.66	0.51	0.48	0.50	0.51	0.54
SIM-C	**0.94**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**

represents average improvements of 127.94% over ShaSpec, 124.64% over M3Care, and 83.07% over PmcmFL, underscoring SIM-C's effectiveness in handling diverse missing patterns, particularly in complex mixed settings where modality presence varies unpredictably.

Table 2 quantifies the relative error (%) of each method on missing configurations compared to their performance on the complete dataset (100_image_100_text). The relative error (%) for a given missing-modality configuration is computed as

$$Error(\%) = \frac{F_{full} - F_{missing}}{F_{missing}} \times 100,$$

where F_{full} denotes the performance (e.g., F-micro or F-samples) achieved on the complete dataset (100% image + 100% text), and $F_{missing}$ is the performance on the respective missing-modality setting (provided that $F_{missing} < F_{full}$). To the best of our knowledge, this normalized error metric is introduced for the first time in this paper. It quantifies the percentage degradation relative to the performance observed under the missing-modality condition itself, rather than relative to the full-modality baseline in absolute terms. This formulation provides a robust, scale-invariant measure of degradation that enables fair comparison across methods with differing absolute performance levels, while emphasizing how much a method loses proportionally when modalities are absent. ShaSpec shows errors ranging from 8.89% to 28.95%, with an average of 20.61%, reflecting inconsistent handling of missing text or images. M3Care has lower variability (2.22% to 24.32%, avg. 11.69%), but still degrades notably in complex_20_40_40. PmcmFL exhibits the highest errors (22.22% to 37.5%, avg. 30.11%), particularly in text-heavy missing cases like 20_image_100_text, indicating reliance on visual completeness. SIM-C, however, maintains minimal errors (0.54% to 1.08%, avg. 0.86%), demonstrating exceptional robustness. This translates to average error reductions of 95.82% over ShaSpec, 92.63% over M3Care, and 97.14%

Table 2: Error(%) of each algorithm in comparison to result on complete MM-IMDB

Method	100_image_20_text	20_image_100_text	complex_10_45_45	complex_20_40_40	complex_30_35_35
ShaSpec	25.64	8.89	25.64	28.95	13.95
M3Care	17.95	6.98	6.98	24.32	2.22
PmcmFL	29.41	37.5	32	29.41	22.22
SIM-C	**0.97**	**0.54**	**1.08**	**0.97**	**0.75**

over PmcmFL, validating SIM-C's bidirectional simulation as highly effective for mitigating performance drops across all tested missing patterns.

Table 3 reports F-samples scores, which provide a balanced measure of precision and recall across samples. On complete data, PmcmFL leads with 0.66, followed by SIM-C at 0.64, ShaSpec at 0.5, and M3Care at 0.45. In missing scenarios, SIM-C consistently outperforms with scores of 0.57–0.60, while baselines range from 0.36–0.54. Averaging over missing configs, SIM-C achieves 0.58, compared to 0.40 (ShaSpec), 0.41 (M3Care), and 0.49 (PmcmFL), yielding improvements of 43.07% over ShaSpec, 42.36% over M3Care, and 17.96% over PmcmFL. This suggests SIM-C not only maintains high aggregate performance (F-micro) but also ensures balanced classification across individual samples, particularly beneficial in multi-label tasks with class imbalance like genre prediction. Overall Analysis: Across all metrics, SIM-C demonstrates superior robustness to missing modalities, with near-zero degradation from full-data performance. Baselines suffer 11–30% average errors, while SIM-C's is under 1%, indicating effective cross-modal simulation. In complex mixed settings, where modality presence mimics real-world variability, SIM-C's consistency (0.93 F-micro) contrasts baselines' fluctuations (0.36–0.54), highlighting its practical utility. Compared to knowledge distillation-focused methods like M3Care and ShaSpec, SIM-C's bidirectional reconstruction better preserves semantic alignments. While PmcmFL excels on complete data due to strong classification, its modality sensitivity limits it; SIM-C balances strong full-data performance (42.42% improvement over PmcmFL) with minimal missing-data loss. These results affirm SIM-C as a state-of-the-art solution for multi-modal genre classification under uncertainty.

Figure **??** provides a qualitative analysis of the quality of imputed embeddings produced by each method under the complex_20_40_40 missing-modality setting, where only 20% of samples have both modalities available, 40% have image only, and 40% have text only. We apply t-SNE dimensionality reduction to project the high-dimensional feature embeddings into 2D space. Yellow points

Table 3: F-samples obtained by the algorithms on MM-IMDB

Method	100_image_100_text	100_image_20_text	20_image_100_text	complex_10_45_45	complex_20_40_40	complex_30_35_35
ShaSpec	0.5	0.38	0.44	0.39	0.38	0.43
M3Care	0.45	0.38	0.43	0.42	0.36	0.44
PmcmFL	**0.66**	0.49	0.45	0.48	0.49	0.54
SIM-C	0.64	**0.57**	**0.60**	**0.57**	**0.57**	**0.58**

correspond to ground-truth embeddings (from samples where the modality is actually present), while red and blue points represent the imputed image and text embeddings, respectively.

For the baseline methods (ShaSpec, M3Care, and PmcmFL), the imputed embeddings (both image and text) form distinct clusters that are noticeably separated from the ground-truth distribution (yellow). This separation indicates that the simulated features deviate substantially from the true manifold, potentially introducing distributional shift and degrading downstream classification performance.

In stark contrast, our SIM-C method produces imputed embeddings (red for image, blue for text) that closely overlap and intermix with the ground-truth yellow points in both modalities. The imputed points are well-distributed within the real embedding manifold, demonstrating that SIM-C's bidirectional simulation modules successfully generate semantically consistent and faithful representations even when the corresponding modality is absent. This superior alignment qualitatively explains the minimal performance degradation observed for SIM-C in quantitative results (Tables 1–3) and confirms the effectiveness of the proposed cross-modal reconstruction objective in preserving the joint embedding space.

5 Conclusion

In this paper, we addressed the challenge of multi-modal movie genre classification under missing textual/visual modalities, a prevalent issue in real-world applications where plot summaries may be incomplete or absent. We proposed SIM-C, a cross-modal feature simulation framework that reconstructs missing embeddings in CLIP's joint space using lightweight bidirectional simulators trained on reconstruction losses. By dynamically imputing absent features during both training and inference, and fusing them via multi-head self-attention, SIM-C ensures robust multi-label classification without requiring raw text generation or heavy model retraining.

Extensive experiments on the MM-IMDb dataset demonstrated SIM-C's superiority, achieving F-micro scores of 0.93–0.94 across various missing-modality configurations, with minimal relative errors (<1%) compared to baselines (11–30%). t-SNE visualizations further confirmed the fidelity of imputed embeddings, closely aligned with ground-truth distributions. These results highlight SIM-C's effectiveness in preserving semantic alignments and handling modality incompleteness, outperforming state-of-the-art methods like ShaSpec, M3Care, and PmcmFL by substantial margins.

Our work advances multi-modal robustness in vision-language tasks, with implications for text-scarce environments like legacy media archives or emerging content platforms. Future directions include extending SIM-C to additional modalities (e.g., audio) and exploring zero-shot adaptation for unseen genres or datasets.

Acknowledgements This work is supported by EU PRESERVE project.

References

1. Ahmad, Z., Jindal, R., Ekbal, A., Bhattacharyya, P.: Multi-modality helps in crisis management: An attention-based deep learning approach of leveraging text for image classification. Expert Systems with Applications **195**, 116626 (2022)
2. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
3. Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., Merhof, D.: Medical image segmentation on mri images with missing modalities: A review. arXiv preprint arXiv:2203.06217 (2022)
4. Bakkali, S., Ming, Z., Coustaty, M., Rusiñol, M., Terrades, O.R.: Vlcdoc: Vision-language contrastive pre-training model for cross-modal document classification. Pattern Recognition **139**, 109419 (2023)
5. Bedi, M., Kumar, S., Akhtar, M.S., Chakraborty, T.: Multi-modal sarcasm detection and humor classification in code-mixed conversations. IEEE Transactions on Affective Computing **14**(2), 1363–1378 (2023)
6. Cao, R., Lee, R.K.W., Chong, W.H., Jiang, J.: Prompting for multi-modal hateful meme classification. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 321–332 (2022)
7. Chen, C., Ye, M., Jiang, D.: Towards modality-agnostic person re-identification with descriptive query. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14815–14824 (2023)
8. Cheng, X., Jia, M., Wang, Q., Zhang, J.: A simple visual-textual baseline for pedestrian attribute recognition. IEEE Transactions on Circuits and Systems for Video Technology **32**(10), 6994–7004 (2022)
9. Dimitrovski, I., Kocev, D., Kitanovski, I., Loskovska, S., Dzeroski, S.: Improved medical image modality classification using a combination of visual and textual features. Computerized Medical Imaging and Graphics **39**, 44–52 (2015)

10. Iyortsuun, N.K., Kim, S.H., Yang, H.J., Kim, S.W., Jhon, M.: Additive cross-modal attention network (acma) for depression detection based on audio and textual features. IEEE Access **11**, 143851–143866 (2023)
11. Jin, T., Cheng, X., Li, L., Lin, W., Wang, Y., Zhao, Z.: Rethinking missing modality learning from a decoding perspective. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7196–7204 (2023)
12. Khattar, A., Quadri, S.: Cross-attention multi-modal classification of disaster-related tweets. IEEE Access **10**, 92889–92902 (2022)
13. Kim, D., Kang, P.: Cross-modal distillation with audio–text fusion for fine-grained emotion classification using bert and wav2vec 2.0. Neurocomputing **506**, 168–183 (2022)
14. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multi-modal prompting with missing modalities for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14943–14952 (2023)
15. Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., Ding, E.: Structext: Structured text understanding with multi-modal transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1912–1920 (2021)
16. Liang, T., Lin, G., Wan, M., Li, T., Ma, G., Lv, F.: Expanding large pre-trained unimodal models with multi-modal information injection for image-text multi-modal classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15498–15507 (2022)
17. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multi-modal transformers robust to missing modality? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18177–18186 (2022)
18. Mafla, A., Dey, S., Biten, A.F., Gomez, L., Karatzas, D.: Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4023–4033 (2021)
19. Peng, C., Zhang, C., Xue, X., Gao, J., Liang, H., Niu, Z.: Cross-modal complementary network with hierarchical fusion for multi-modal sentiment classification. Tsinghua Science and Technology **27**(4), 664–679 (2022)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Wang, G., Yu, F., Li, J., Jia, Q., Ding, S.: Exploiting the textual potential from vision-language pre-training for text-based person search. arXiv preprint arXiv:2303.04497 (2023)
22. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15878–15887 (2023)
23. Xu, M., Yuan, X., Miret, S., Tang, J.: Protst: Multi-modality learning of protein sequences and biomedical texts. In: International Conference on Machine Learning. pp. 38750–38767. PMLR (2023)
24. Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., Zhao, J.: M3care: Learning with missing modalities in multi-modal healthcare data. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2418–2428 (2022)

25. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2608–2618 (2021)
26. Zhou, Y., Yang, Y., Ying, Q., Qian, Z., Zhang, X.: Multi-modal fake news detection on social media via multi-grained information fusion. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. pp. 78–86 (2023)
27. Zhu, T., Li, L., Yang, J., Zhao, S., Xiao, X.: Multi-modal emotion classification with multi-level semantic reasoning network. IEEE Transactions on Multimedia **25**, 6868–6880 (2023)

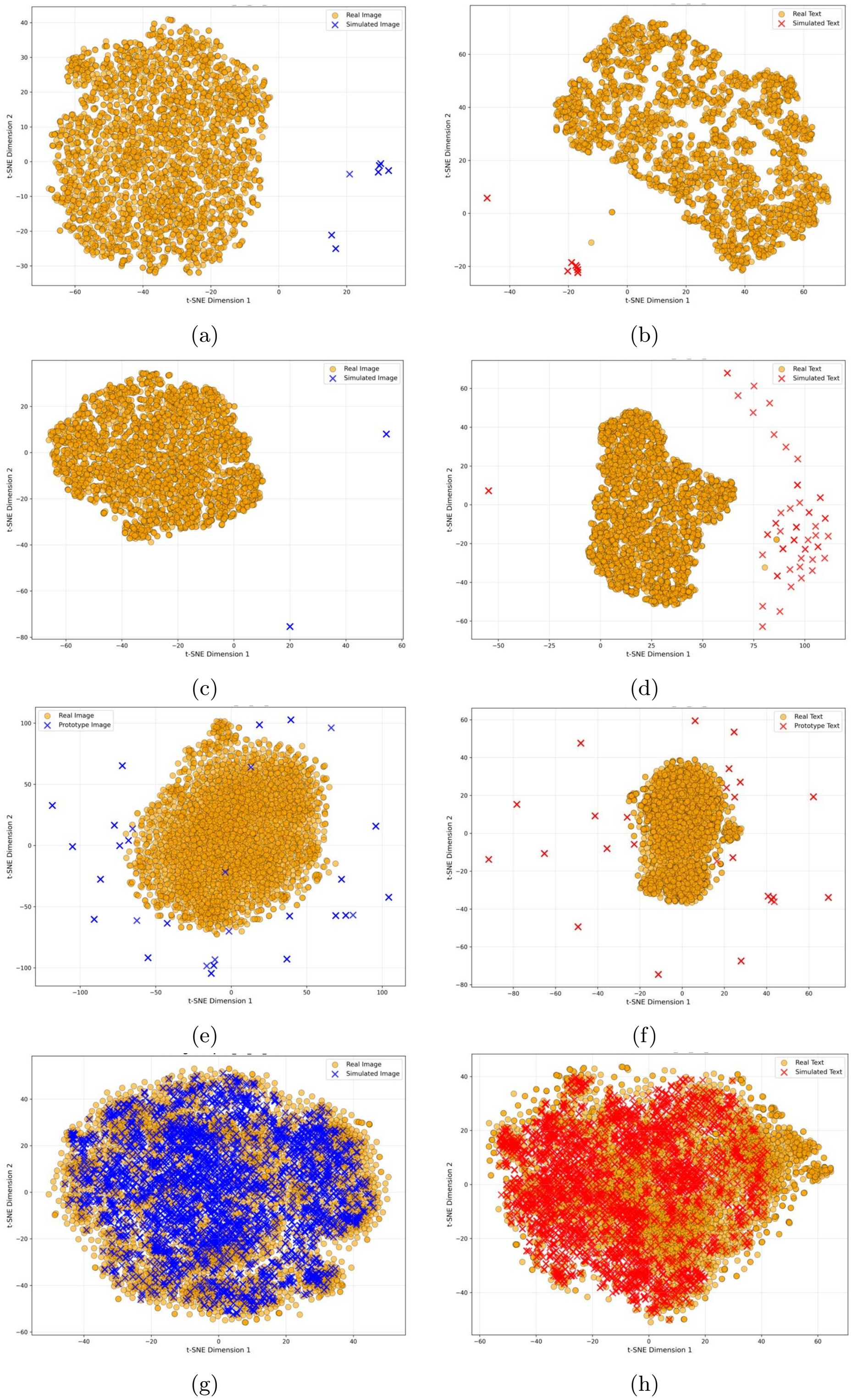

Fig. 1: Visualisation of the feature embedding of the missing modalities.