Multilingual Communication System with Deaf Individuals Utilizing Natural and Visual Languages

Tuan-Luc Huynh[†1,2], Khoi-Nguyen Nguyen-Ngoc[†1,2], Chi-Bien Chu[†1,2], Minh-Triet Tran[1,2], and Trung-Nghia Le[‡1,2]

[1]Faculty of Information Technology, University of Science, VNU-HCMC, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

***Abstract*—According to the World Federation of the Deaf, more than two hundred sign languages exist. Therefore, it is challenging to understand deaf individuals, even proficient sign language users, resulting in a barrier between the deaf community and the rest of society. To bridge this language barrier, we propose a novel multilingual communication system, namely MUGCAT, to improve the communication efficiency of sign language users. By converting recognized specific hand gestures into expressive pictures, which is universal usage and language independence, our MUGCAT system significantly helps deaf people convey their thoughts. To overcome the limitation of sign language usage, which is mostly impossible to translate into complete sentences for ordinary people, we propose to reconstruct meaningful sentences from the incomplete translation of sign language. We also measure the semantic similarity of generated sentences with fragmented recognized hand gestures to keep the original meaning. Experimental results show that the proposed system can work in a real-time manner and synthesize exquisite stunning illustrations and meaningful sentences from a few hand gestures of sign language. This proves that our MUGCAT has promising potential in assisting deaf communication.**

***Index Terms*—Sign Language Recognition, Text-to-Image Synthesis, Image Captioning**

I. Introduction

Communication with deaf people is mainly based on sign language, a combination of hand gestures, facial expressions, and postures to convey semantic information. However, these visual communication systems are difficult to learn and remember, leading to barriers between the deaf community and the rest of society; this problem has not been fully solved until now.

Although technologies have been developed to understand the behaviors of deaf people, such as sign language translation via cameras and sensory gloves, they still have several issues. Sign language translation via camera systems [1], [2] needs a fixed camera and a simple background to recognize sign gestures accurately. Besides that, translating sign languages to human-understandable languages often leads to unnatural results. It causes difficulties in jobs that require smoothness in words, such as explaining new concepts or telling stories. The birth of sensory devices like smart gloves [3], [4] is a big step forward in this field. However, it still does not solve the problem of unnatural translations. In addition, the more modern sensors will come with high prices, which makes it difficult to reach the deaf community.

Combined with natural language, which can be expressed as text or voice, visual language can reform the communication between ordinary people and deaf/dumb people. Indeed, visual cues (*e.g.*, images, videos, 3D models) are the best aid to express new concepts intuitively. Visual cues play an important role in deaf communication, especially in literacy education for deaf children. Visual communication efficiently bridges ordinary people with the deaf community, regardless of different nationalities or different languages.

To assist communication with deaf individuals, we propose a **MU**ltilin**G**ual **C**ommunic**AT**ion system (MUGCAT). Inspired by the adage "A picture is worth a thousand words," our system supports diverse cues, such as sign languages, natural languages, and visual languages, to help deaf people express their thoughts more clearly. The proposed MUGCAT system consists of two main phases: converting sign language to intermediate language, which ordinary people can understand, and enriching the translated information by reconstructing a meaningful sentence aided by illustrations. We first recognize and translate these hand gestures of deaf individuals (*i.e.*, sign language) into human-understandable language (*i.e.*, textual words or phases). Illustrations are then synthesized via a text-to-image model for visual communication. By transforming sign languages into pictures - universal mediums of expressiveness - our system significantly help deaf individuals convey their thoughts. Due to the limitation of sign languages, it is challenging to translate hand gestures to complete meaningful sentences for ordinary people. Therefore, we propose using an image captioning method to assist the incompleted translated text. Furthermore, our MUGCAT

system can measure the semantic similarity of generated image captions with the intermediate translated text to keep the original meaning of the sign language. In this way, our MUGCAT system can help to express the intentions of deaf communicators more intuitively and clearly.

Experimental results on the WLASL dataset [5] show the potential of our MUGCAT system in assisting natural communication with deaf individuals. The proposed system can recognize sign gestures with an accuracy of 46.8% in real time. In addition, meaning sentences for humans are generated with corresponding exquisite, beautiful, and stunning illustrations. We expect our MUGCAT system to benefit both the deaf community and the sign language research community.

Our main contributions are summarized as follows:

- We propose a novel system, namely MUGCAT, to support multilingual communication for deaf individuals. Our simple yet efficient system utilizes both natural and visual languages to enhance the interpretation of deaf communicators.
- Our MUGCAT system accurately recognizes and translates sign languages to human-understandable text. The proposed system also can transform the translated text into illustrative and expressive images in real-time performance.
- The synthesized images might be misleading; hence, we propose to use an image captioning model to select the image that best fit the translated text, further improving the efficiency of MUGCAT.

II. Related Work

A. Deaf Communication

Communicating with deaf individuals mainly occurs through auditory (*e.g.*, lip reading) and visual (*e.g.*, sign language) modes. However, sign language is more popular than lip reading because understanding speech by visually interpreting the movements of the lips, face, and tongue is extremely challenging, even for deaf people.

With the development of modern technology, many technological devices have been invented to translate sign language into text, speech, etc. Some special sensors were made to detect hand movements. The translation glove products (EnableTalk [3] and SignAloud [4]) work on the integration of sensors that attach to the finger to record hand posture and movements and then convert sensor signals into speech through an independent processing unit. However, these products are difficult to access widely due to their high cost and difficulty to use in daily life.

B. Sign Language Recognition

Recently, many computer vision algorithms have been proposed to recognize sign language from video only, thus avoiding the dependence on costly sensor devices. Given a video, besides RGB frames, we can also obtain other modalities of input such as image depth [6], [7] and optical flow [8], [9] (pixel-wise motions between consecutive video frames). For RGB input only, 3D ConvNets were widely applied [10], [11] to extract spatial-temporal information from videos. Lin *et al*. [12] inserted a Temporal Shift Module into 2D ConvNets to get an accuracy commensurate with 3D ConvNets while keeping the complexity of 2D ConvNets. Komkov *et al*. [13] combined the learned knowledge from multiple single-modality models with mutual learning technique [14] to obtain the best model on each input modality.

C. Text-to-Image

In the last couple of years, text-to-image models [15] have attracted big tech companies' attention and thus have received rapid and massive improvements. Classifier-free guided diffusion models have recently been shown to be highly effective at high-resolution image generation, and they have been widely used in large-scale diffusion frameworks, including GLIDE [16], DALL-E 2 [17], and Imagen [18]. Nevertheless, the latest development of diffusion model-based text-to-image model, namely Stable Diffusion [19], has been the most significant impact since its release. Stable Diffusion offers excellent image quality while significantly lowering the computation cost. What makes Stable Diffusion exceptionally attractive compared to other competitors due to its open-source. On the other hand, Google and OpenAI do not intend to open-source Imagen [20] and DALL-E 2 [17], respectively.

While the artificial intelligence community has dominantly used text-to-image models to create beautiful artworks, there is little attention on using these models on real-world problems. In this work, we customized Stable Diffusion to generate meaningful and expressive images from the translated sign language text in a real-time manner to help visualize conversation with or between sign language users.

D. Image Captioning

Research on image captioning in recent years generally uses the encoder-decoder architecture. The encoder extracts the visual information from images for the decoder, which generates an acceptable description. In the early, the encoder was a CNN backbone [21], [22]. Later, it was replaced by an object detector such as Faster R-CNN to extract object-level features [23]. This proved more efficient and improved performance because the object information and their relationships are very useful in describing an image. However, due to the high computational cost of the object detection model, it is hard to apply in a problem that requires high speed, such as communication. Besides that, Transformer applications in the encoder to extract features or the decoder for caption generating task [24], [25] also demonstrated surprising efficiency improvements.

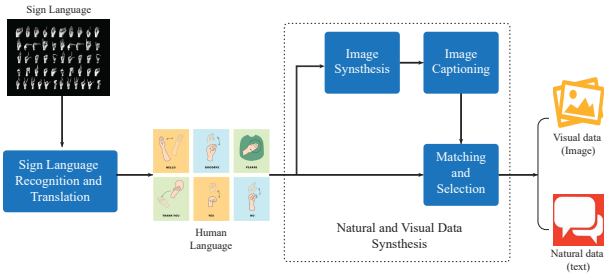

Fig. 1. Pipeline of our multilingual communication (MUGCAT) system

In this study, with the aim of balancing accuracy and efficiency, we used a recent state-of-the-art method [26] which proposed a Transformer-only neural architecture utilizing dual visual features to improve performance and increase speed.

III. Proposed System

A. Overview

Figure 1 illustrates the pipeline of our proposed multilingual communication (MUGCAT) system, which consists of two main components: Sign language recognition and translation (SLRT), Natural and visual data synthesis. First, words obtained from SLRT are illustrated by the text-to-image module resulting in several images. Then, image captioning is carried out to achieve complete descriptions of all synthesized images. Finally, with each image and its description, we compare its semantic similarity with the translated keywords from SLRT to choose the most suitable image and description. Unlike conventional SLRT systems that need to correctly recognize the whole sentence to express the meaning, our system generates a suggested image and complete description to represent the keywords made in sign language to overcome the disadvantage of missed recognized sign language.

B. Sign Language Recognition and Translation

Sign language recognition aims to predict the sequence of signs performed in a video, while sign language translation further translates the signs into spoken/written languages. To synthesize images that fully convey the meaning of a signer, we only need to identify some keywords from the hand gesture sequence. Therefore, the recognition task is more suitable for our MUGCAT system because it is simple but still responsive to the system. Due to the lack of suitable datasets in this domain, we treat the problem as an action recognition task, where the objective is to identify single words from short clips. This simplifies the problem and meets our requirement for indicating visual keywords.

In this work, we employed and compared several action recognition methods [10], [12], [13] on WLASL dataset [5], the largest video dataset of word-level American sign language (ASL). The main ideas of employed methods are summarized as follows:

Two-Stream Inflated 3D ConvNets (I3D) [10] combines two 3D ConvNets (one for RGB image stream, one for optical-flow stream) to both take advantage of pre-trained ImageNet weights and force the model to learn motion features directly. Two 3D ConvNets are trained separately, and their predictions are averaged at test time.

Temporal Shift Module (TSM) [12] inserted TSM module into 2D ConvNets to capture temporal relationships between video frames. Feature maps are shifted along the temporal dimension to maintain 2D ConvNet's complexity while achieving the performance of 3D ConvNets.

Mutual Modality Learning (MML) [13] ensembled knowledge from single-modality models into a single model to obtain the best single-modality model for each modality. The algorithm can be summarized in three steps: train two separate networks A_1, A_2 on the RGB modality; respectively initialize two networks B_1, B_2 with the weights of A_1, A_2, then train B_1, B_2 together using mutual learning technique on RGB modality; from B_1's weights, initialize N models C_1, C_2, ..., C_N corresponding to N different modalities (RGB, optical flow, and depth), then train these N models together using mutual learning.

C. Natural and Visual Data Synthesis

1) Text-To-Image Synthesis: Stable Diffusion [19], a state-of-the-art diffusion-based text-to-image model, is the core component of our system, which strives to actualize the adage "A picture is worth a thousand words." This method can offer excellent image quality while significantly lowering computation costs.

However, the sequential sampling process of diffusion-based models is time-consuming. As a result, the text-to-image module is also the bottleneck of our system. To overcome this limitation, we customized hyperparameters of Stable Diffusion to retain high-quality images while significantly reducing the sampling process time.

Another issue that affects our system performance is the relevancy of synthesized images. Prompt engineering (*i.e.*, prompt modifiers) is necessary for guiding the text-to-image models to generate superior-quality art. However, in our proposed system, the prompt text for Stable Diffusion is limited keywords from the SLRT module. Therefore, it is unavoidable that the prompt's quality is limited, which leads to potential drops in generated image relevancy. We addressed this issue by introducing the image captioning model in the system's next stage, which serves as a filter to select the most relevant image.

2) Image Captioning: Image captioning methods are classified into two main approaches: grid features and region features. Methods based on grid features directly extract object features from high-layer feature maps of

the whole image. Thus, generated captions can contain information about the whole image. Meanwhile, methods based on region features [23] rely on detecting objects in the image and then extracting local features of image regions to infer results. However, detected objects cannot represent the overall context of the image nor the relationships of objects that affect generated captions.

In this work, we used Grid- and Region-based Image captioning Transformer (GRIT) [26], a state-of-the-art image captioning method, which uses both types of mentioned features to enhance both contextual information and object-level information. Grid features are extracted using a standard self-attention Transformer, and region features are extracted by Deformable DETR detector [27]. Then, the extracted features are fed to a caption generator based on Transformer to generate the final caption. In this step, we employed Parallel Cross-Attention [26] to relate between dual visual features and caption words.

3) Matching and Selection: SLRT generates incomplete keywords; thus, the images synthesized from the previous step inevitably are not completely consistent with each other and the communicator's expression. Therefore, we propose an extra step of matching and selecting the caption whose meaning is closest to the input keywords. From there, our system is able to recover the complete sentence that the communicator wanted to express from just the discrete words.

Concretely, given K results $\{I_1, I_2, \ldots, I_K\}$ of the previous step, the goal of the matching and selection is to find the most suitable pair of image $\widehat{I}$ and its caption $q_{\widehat{I}}$. For each caption sentence, we measure its semantic similarity with keywords obtained from SLRT. We used Sentence Transformers [28], denoted by $\psi(\cdot)$, to compute sentence embeddings and evaluate them with cosine similarity. Mathematically, it performs a maximization expressed as:

$$\{\widehat{I}, q_{\widehat{I}}\} = \underset{i \in \{1,2,\ldots,K\}}{\operatorname{argmax}} D(\psi(q_{I_i}), \psi(Q)), \qquad (1)$$

where Q is a sentence that includes keywords received from SLRT, $D(\cdot)$ is a cosine similarity function.

IV. Experiments

In this section, we elaborate on the extensive experiments conducted on our proposed system. All experiments were tested on a machine with a single Nvidia V100 GPU.

A. Sign Language Recognition

As shown in Table IV-A, we compared several action classification methods [10], [12], [13] on the WLASL [5] test set. In detail, for TSM [12], we trained a model from scratch and fine-tuned another model, which was pre-trained on the Kinetic [10] dataset. For MML [13], we trained a model from scratch with only RGB input as WLASL [5] dataset does not provide optical flow or depth annotations. All the networks above use an ImageNet-pre-trained ResNet50 as the backbone. Lastly, we reused two public I3D [10] with different pretraining datasets [5], [29] for our experiment.

TABLE I
Accurary and efficiency on the WLASL [5] test set. All the compared methods utilize a pre-trained backbone on ImageNet, and then they were finetuned on the WLASL dataset. The FPS was measured on a Nvidia V100 GPU.

Method	**Pretraining Dataset**	**Accuracy (%)**	**FPS** (infer only)	**FPS** (infer & load data)
I3D [10]	BSL1K [29]	**46.8**	**1429**	95
	Kinetic [10]	32.5		
TSM [12]	✗	20.8	357	60
	Kinetic [10]	13.9		
MML [13]	✗	20.8	323	**104**

We first evaluated top-1 accuracy on the WLASL [5] test set. The main challenge of this dataset is the number of words to classify up to 2,000, while the number of videos in the training set is just over 14,000. Therefore compared methods only achieve acceptable accuracy. I3D achieved the top-1 accuracy of 46.8%, which is also state-of-the-art top-1 accuracy on the WLASL test set.

We then evaluated the efficiency of methods by measuring the execution time and the number of processed frames on the whole test set to obtain the average FPS. All methods can run in a real-time manner. Especially, MML [13] can achieve 104 FPS counting all initialization steps, such as loading the model, preparing the dataset, etc. We also tried to compute FPS in the practice scenario, where SLRT methods directly process the video stream and ignore the initialization steps. All methods can achieve more than 300 FPS, and I3D [10] achieved a surprising speed of 1429 FPS. The results show that these models have the potential to be deployed on mobile devices and embedded systems while still achieving real-time speed.

B. Stable Diffusion Hyperparameters Adjustment

The default settings of Stable Diffusion [19] hardly achieve near real-time performance. This section discusses our extensive experiments in various settings to discover the optimal trade-off point between execution time and image quality. Specifically, we focus on the number of sampling steps, the desired resolution, and the number of samples. We used the public checkpoint *sd-v1-4.ckpt* in our experiments.

The number of sampling steps is the most crucial hyperparameter that directly controls the quality of generated images and positively correlates with the execution duration. The default hyperparameter of 50 sampling steps using PLMS sampler [30] generates high-quality images. Since our system ideally should work in real-time performance, we figured that 20 sampling steps could speed up 2.4 times while having subtle drops

TABLE II
Performance of Stable Diffusion on a single Nvidia V100 GPU. The prompt is "A beautiful flower garden on a sunny day with a valley background." Resolution is 512×512. The FID score [31] was calculated using 50 sampling steps as the real distribution, 128 images per distribution. The optimal hyperparameter is 20 sampling steps, which can keep the image quality but speed up 2.4 times.

Sampling steps	FID Score ↓	Seconds per Batch ↓
50	0	35.50
45	33.43	32.05
40	30.44	28.66
35	31.70	25.24
30	31.55	21.79
25	33.19	18.39
20	33.51	14.97
15	40.33	12.25

Fig. 2. The renowned "a photograph of an astronaut riding a horse." The top and down rows are 20 and 50 sampling steps, respectively.

in contextual information, as demonstrated in Fig. 2. Indeed, Table IV-B shows that setting the sampling steps to 20 is optimal with the FID of 33.5, approximately equivalent to higher sampling steps but can process a batch of 128 images in only 15 seconds. Going lower than 20 sampling steps results in a surge of FID scores.

Image resolution is another element that significantly affects Stable Diffusion's running time. Following tips of Suraj *et al*. [32], we tried reducing the resolution and came up with seven different resolutions in decreasing execution time, namely 512×512, 512×448, 448×448, 512×384, 448×384, 512×320, and 384×384. As illustrated in Fig. 4, the first two images on the top row have a valley background. As the resolution decreases, contextual information in the prompt will gradually become less constrained.

Fully utilizing GPU's capability is another technique to enhance the performance of the model. We tried setting the largest batch size on each respective resolution and recorded the run time accordingly. Figure 3 illustrates the benchmark result of the highest resolution, lowest resolution, and median one. The experimental result shows that the optimal number of synthesis images (*i.e*., K in Eq. 1) is either 8 or 16. The reason is twofold: Firstly, a small batch size can easily fit into a conventional GPU; Secondly, since the image captioning

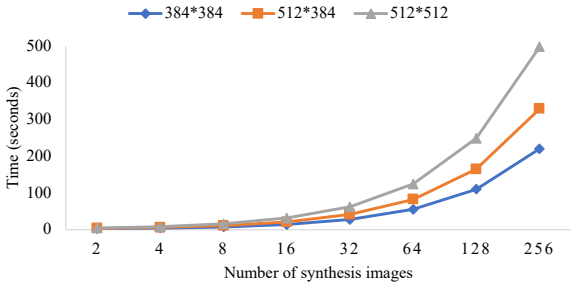

Fig. 3. Stable Diffusion's execution time in different resolutions using a single Nvidia V100 GPU. The batch size is maximized for each respective resolution, and the hyperparameter of sampling steps is 20. The optimal numbers of synthesis images are from 8 to 16.

Fig. 4. Images generated by Stable Diffusion using the same prompt "A beautiful flower garden on a sunny day with a valley background" in decreasing resolution order (from left to right, top to bottom).

module's execution time scales linearly with the number of generated images, selecting a small batch size can thus improve both modules' performance.

C. Image Captioning Visualization

We employed GRIT [26] model that uses the pre-trained of object detector on four datasets: COCO [33], Visual Genome, Open Images [34], Object365 [35], and applied Parallel Cross-Attention [26] for image captioning. We can achieve the per-batch inference time of about 0.75s when setting the batch size of 16 and 0.87s with the batch size of 8 on a single Nvidia V100 GPU.

Example results are visualized in Fig. 5. With the developed text refinement mechanism, our MUGCAT system obviously generates an illustration and complete caption with high semantic similarity with the original sentence from the keywords, as shown in Fig.5.

V. Conclusion

We have proposed a **MU**ltilin**G**ual **C**ommunic**AT**ion system (MUGCAT), which integrates sign language recognition and translation, text-to-image, and image captioning methods. The proposed system harmonizes three different methodologies to help overcome the difficulty of communicating with deaf individuals. Leveraging the latest development in text-to-image synthesis and image captioning to transform written text into visual images, we strive to lift the language barrier that has always existed in the sign language community.

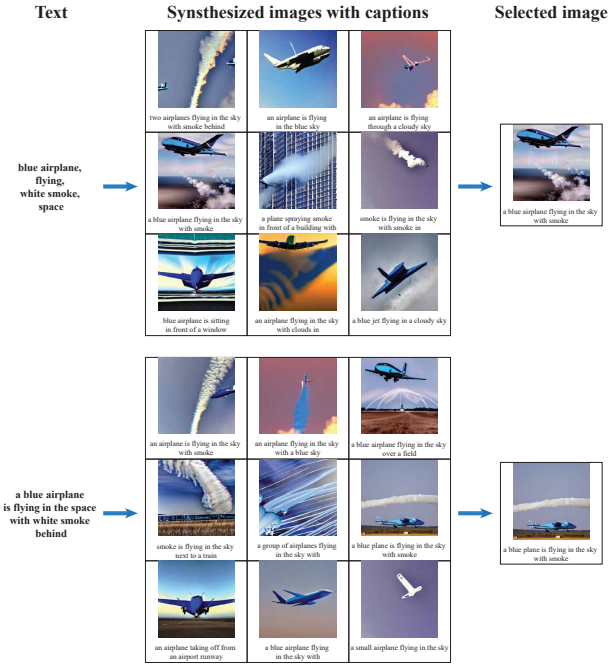

Fig. 5. Example of text refinement on a complete sentence (below) and only on keywords (above) giving similar results.

Experiments show the potential of our proposed system in practice. In the future, it would be interesting to modify our system's camera to first-person. We want to explore the possibility of sign language recognition and translation methods from a first-person perspective since this would overcome the problem of requiring standing in front of a fixed camera.

Acknowledgment. This research is funded by University of Science, VNU-HCM, under grant number CNTT 2022-15.

REFERENCES

[1] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *ICCV*, 2021.

[2] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *ICCV*, 2021, pp. 11 542–11 551.

[3] F. Lardinois, "Ukrainian students develop gloves that translate sign language into speech," https://techcrunch.com/2012/07/09/enable-talk-imagine-cup, 2012.

[4] "UW undergraduate team wins $10,000 lemelson-mit student prize for gloves that translate sign language," https://www.washington.edu/news/2016/04/12/uw-undergraduate-team-wins-10000-lemelson-mit-student-prize-for-gloves-that-translate-sign-language, 2016.

[5] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *WACV*, 2020, pp. 1459–1469.

[6] A. Gurram, A. F. Tuna, F. Shen, O. Urfalioglu, and A. M. López, "Monocular depth estimation through virtual-world supervision and real-world sfm self-supervision," *IEEE T-PAMI*, 2021.

[7] J. Wang, Y. Zhong, Y. Dai, S. Birchfield, K. Zhang, N. Smolyanskiy, and H. Li, "Deep two-view structure-from-motion revisited," *CVPR*, 2021.

[8] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li, "FlowFormer: A transformer architecture for optical flow," *ECCV*, 2022.

[9] S. Bai, Z. Geng, Y. Savani, and J. Z. Kolter, "Deep equilibrium optical flow estimation," in *CVPR*, 2022.

[10] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE T-PAMI*, vol. 35, no. 1, pp. 221–231, 2013.

[12] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *ICCV*, 2019.

[13] S. Komkov, M. Dzabraev, and A. Petiushko, "Mutual modality learning for video action classification," *arXiv preprint arXiv:2011.02543*, 2020.

[14] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *CVPR*, 2018, pp. 4320–4328.

[15] S. S. Baraheem, T.-N. Le, and T. V. Nguyen, "Text-to-image synthesis via aesthetic layout," in *ACM MM*, 2020.

[16] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[17] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[18] C. Meng, R. Gao, D. P. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," *arXiv preprint arXiv:2210.03142*, 2022.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[20] Google, "Imagen," https://imagen.research.google/, 2022.

[21] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[22] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017.

[23] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[24] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *ICCV*, 2019.

[25] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *CVPR*, 2020.

[26] V.-Q. Nguyen, M. Suganuma, and T. Okatani, "Grit: Faster and better image captioning transformer using dual visual features," *ArXiv preprint arXiv:2207.09666*, 2022.

[27] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[28] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *ArXiv preprint arXiv:1908.10084*, 2019.

[29] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," in *ECCV*, 2020.

[30] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *ICLR*, 2022.

[31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, vol. 30, 2017.

[32] S. Patil, P. Cuenca, N. Lambert, and P. von Platen, "Stable diffusion with diffusers," https://huggingface.co/blog/stable_diffusion.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[34] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.

[35] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *ICCV*, 2019.