EASL: MULTI-EMOTION GUIDED SEMANTIC DISENTANGLEMENT FOR EXPRESSIVE SIGN LANGUAGE GENERATION

Yanchao Zhao[3], *Jihao Zhu*[4], *Yu Liu*[1,2†], *Weizhuo Chen*[1,2], *Yuling Yang*[1,2], *Kun Peng*[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]School of Rehabilitation Science and Engineering, University of Health and Rehabilitation Sciences
[4]School of Language, Literature, Music and Visual Culture, The University of Aberdeen

ABSTRACT

Large language models have revolutionized sign language generation by automatically transforming text into high-quality sign language videos, providing accessible communication for the Deaf community. However, existing LLM-based approaches prioritize semantic accuracy while overlooking emotional expressions, resulting in outputs that lack naturalness and expressiveness. We propose EASL (Emotion-Aware Sign Language), a multi-emotion-guided generation architecture for fine-grained emotional integration. We introduce emotion-semantic disentanglement modules with progressive training to separately extract semantic and affective features. During pose decoding, the emotional representations guide semantic interaction to generate sign poses with 7-class emotion confidence scores, enabling emotional expression recognition. Experimental results demonstrate that EASL achieves pose accuracy superior to all compared baselines by integrating multi-emotion information and effectively adapts to diffusion models to generate expressive sign language videos.

Index Terms— Sign Language Generation, Large Language Models, Multimodal Learning, Emotion Expression.

1. INTRODUCTION

Sign language generation represents a critical challenge in multimodal artificial intelligence, aiming to convert written or spoken language into visual expression through gestures, postures, and facial expressions[1]. With advances in deep learning and computer vision, significant progress has been made in semantic accuracy and generation quality[2]. Current research approaches falls into two categories: end-to-end and pose-driven approaches. The former has evolved from CNN/GAN to Transformer/diffusion models, emphasizing semantic alignment and intelligibility; the latter employs keypoint sequences to drive virtual characters[3], highlighting naturalness and editability. The prevailing paradigm integrates pose-based intermediate representations with diffusion/Transformer architectures.

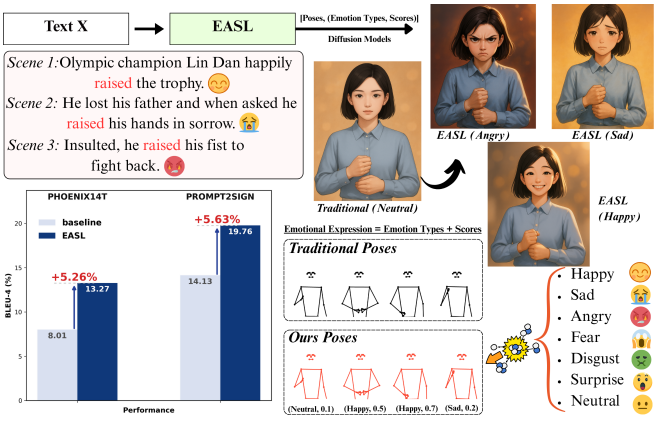

Fig. 1: EASL demonstrating accuracy improvements over baseline methods (**bottom left**) while applying appropriate emotional expressions to identical semantic content across different contextual scenes (**raised**) through multi-emotion guidance with confidence scores.

Despite these advances, three critical challenges persist in current sign language generation systems: **First**, insufficient emotional expression results in monotonous and unvivid generated videos that lack the natural expressiveness of human communication[4, 5]. **Second**, inadequate integration of emotion and semantics leads to a disconnect between emotional expression and semantic content[6, 7]. **Third**, reliance on static input makes it difficult to dynamically adjust emotional intensity and variability[8], particularly in emotional scenarios where nuanced expression is crucial.

To address these limitations, we propose **EASL** (**E**motion-**A**ware **S**ign **L**anguage), a multi-emotion-guided, confidence-driven architecture for sign language generation. Our framework features two core modules: **DESE** (**D**isentangled **E**motion-**S**emantic **E**ncoder) uses gated attention to separate semantic and emotional representations, while **EGSID** (**E**motion-**G**uided **S**emantic **I**nteraction **D**ecoder) leverages

†Corresponding author

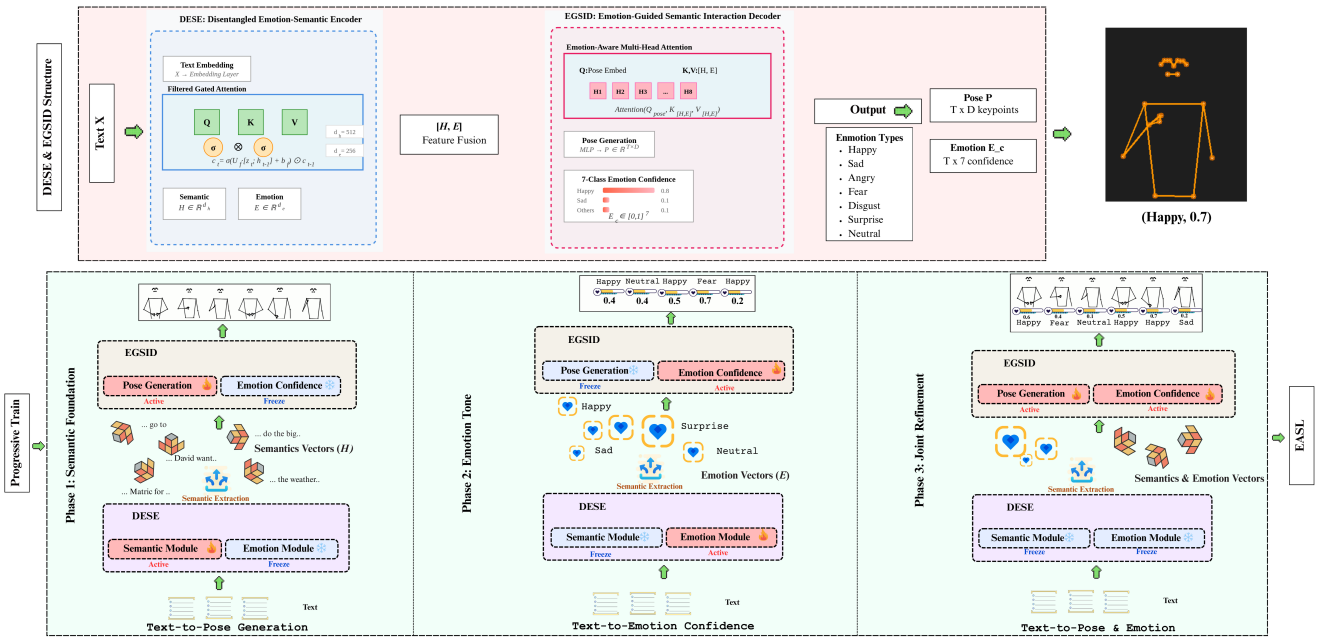

Fig. 2: EASL architecture with three-phase training strategy.

emotional features to guide semantic decoding, generating poses with seven-class emotion confidence scores. A three-phase progressive training strategy sequentially builds semantic understanding, extracts emotional features, and achieves joint refinement, preventing feature entanglement while ensuring semantic accuracy and emotional expressiveness.

Our main contributions are as follows:

1. We propose **EASL**[1], a multi-emotion guided semantic disentanglement framework that decouples semantic and emotional features, effectively adapting to diffusion models to generate emotionally expressive sign language videos.

2. We design an emotion-semantic disentanglement architecture with two modules: **DESE** for separating semantic and emotional representations, and **EGSID** for emotion-guided semantic interaction.

3. We adopt three-phase progressive training strategy with parameter freezing across semantic foundation, emotional feature extraction, and joint refinement stages, achieving integration of emotional expression with sign language generation.

4. Experimental results demonstrate that EASL achieves pose accuracy superior to all compared baselines by integrating multi-emotion information and effectively adapts to diffusion models to generate expressive sign language videos.

[1]Codes: https://github.com/ycz-res/SLP/

2. METHODOLOGY

2.1. Problem Formulation

Given text sequence $\mathcal{X} = \{x_t\}_{t=1}^{T}$, we aim to learn a mapping function $\mathcal{F} : \mathcal{X} \to (\mathcal{P}, \mathcal{E})$ that generates pose sequence $\mathcal{P} = \{p_i\}_{i=1}^{M}$ with $p_i \in \mathbb{R}^n$, and an emotion-confidence sequence $\mathcal{E} = \{e_i\}_{i=1}^{M}$, where $e_i = (e_i^{(1)}, \ldots, e_i^{(K)}) \in [0,1]^K$. Here, K is the number of emotion categories, and the superscript (k) indexes the k-th category.

2.2. Architecture Overview

As shown in Figure 2, EASL employs a dual-module architecture with complementary streams. The upper stream maps text to pose sequences with multi-emotion confidence scores: DESE (Disentangled Emotion–Semantic Encoder) disentangles semantic and emotional representations, which EGSID (Emotion-Guided Semantic Interaction Decoder) integrates via emotion-guided semantic interactions for decoding. The lower stream illustrates the three-stage training of DESE and EGSID that underpins and optimizes the upper pipeline.

2.3. DESE: Disentangled Emotion-Semantic Encoder

DESE extracts a sequence of frame-wise semantic representations $H = \{h_t\}_{t=1}^{T}$ with $h_t \in \mathbb{R}^{d_h}$ (semantic dimension d_h) through filtered gated attention:

$$\tilde{c}_t = \sigma(U_f \cdot [z_t; h_{t-1}] + b_f) \odot \tilde{c}_{t-1}, \quad (1)$$

$$h_t = \mathcal{A}(W_q \tilde{c}_t,\, W_k[z_t; h_{t-1}],\, W_v[z_t; h_{t-1}]), \tag{2}$$

where σ is the sigmoid activation, $[;]$ denotes concatenation, and $\odot$ is element-wise multiplication. z_t is the input embedding, h_{t-1} the previous hidden state, and $\tilde{c}_{t-1}$ the previous filtered context. U_f, b_f are gating parameters, and W_q, W_k, W_v are the query/key/value projection matrices for attention $A(\cdot)$.

The emotional representation is extracted as a sequence $E = \{e_t\}_{t=1}^{T}$ with $e_t \in \mathbb{R}^{d_e}$ (emotion dimension d_e):

$$e_t = \sigma(W_u h_t + b_u) \odot e_{t-1}, \tag{3}$$

where W_u and b_u are the emotion-gating parameters, and e_{t-1} is the previous emotional state. Equivalently, $H \in \mathbb{R}^{T \times d_h}$ and $E \in \mathbb{R}^{T \times d_e}$.

2.4. EGSID: Emotion-Guided Semantic Interaction Decoder

EGSID leverages emotion-aware multi-head attention, using E to guide the interaction over H, to generate the pose sequence P and, for each frame, outputs a seven-class emotion confidence vector $E_c \in [0,1]^7$:

$$\{P, E_c\} = L_\Theta\Big(A\big(U_q(P_{\text{embed}} + P_{\text{pos}}),\, U_k(H,E),\, U_v(H,E)\big)\Big), \tag{4}$$

where P_{embed} and P_{pos} denote the query-side pose embedding and positional encoding, respectively; U_q, U_k, U_v are learnable Q/K/V projections; L_Θ is the task-specific readout operator.

2.5. Three-Phase Training Strategy

To achieve emotion–semantic disentanglement and global emotion guidance, we adopt a three-stage training paradigm. Let the parameter set be $\Theta = \{\Theta_s, \Theta_e, \Theta_c\}$: $\Theta_s$includes DESE's semantic encoder and EGSID parameters; Θ_e includes DESE's emotion encoder and EGSID parameters; Θ_c includes only EGSID parameters. **Stage 1** (Semantic Foundation, Θ_s): text→pose, learning DESE's semantic representation. **Stage 2** (Emotion Tone, Θ_e, freezing Θ_s): text→multi-emotion confidence scores, learning DESE's emotion representation. **Stage 3** (Joint Refinement, Θ_c, freezing Θ_s and Θ_e): train EGSID only, jointly refining text→pose and multi-emotion confidence scores.

Loss function. We use Mean Absolute Error (MAE) to evaluate pose and emotional generation accuracy with three main loss functions:

$$L_{\text{pose}} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{D}\sum_{d=1}^{D}\left|\hat{x}_{t,d} - x_{t,d}\right|, \tag{5}$$

$$L_{\text{emo}} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{7}\sum_{k=1}^{7}\left|\hat{p}_{t,k} - p_{t,k}^{\text{GT}}\right|, \tag{6}$$

$$L = \lambda_{\text{pose}} \cdot L_{\text{pose}} + \lambda_{\text{emo}} \cdot L_{\text{emo}}, \tag{7}$$

where T denotes sequence length, D the pose dimension, $\hat{x}$ and x the predicted and ground-truth poses, $\hat{p}$ and p^{GT} the predicted and ground-truth emotions, and λ_{pose}, λ_{emo} the corresponding weights.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. We evaluate EASL on two benchmark sign language datasets: **PHOENIX14T** [9] and **Prompt2Sign** [10].

Baselines. We compare EASL with four method categories: pre-trained Transformer-based methods (**PT-base** [11], **PT-FP&GN** [11], **GEN-OBT** [6]), pose detection/regression methods (**DET**), traditional deep learning generation methods (**NSLP-G** [12], **Fast-SLP** [13]), and diffusion/generative model methods (**NSA** [14], **SignLLM-M** [15]).

Evaluation Metrics. We report **BLEU-1/2/3/4** (n-gram precision), **ROUGE-L** (longest common subsequence), and **MAE** (Mean Absolute Error). BLEU-4 serves as primary semantic accuracy metric, while MAE evaluates emotional expression quality.

Implementation Details. We use **OpenPose** [16] for pose keypoint extraction and leverage **CLIP** [17] to annotate frame-level emotion confidence scores across seven categories (Happy, Sad, Angry, Fear, Disgust, Surprise, Neutral). We use two available pre-trained BERT models (semantic and emotion) to track DESE/EGSID's emotional representations and the evolution of metrics, while pose quality is evaluated via back-translation.

3.2. Main Results

Table 1 presents EASL's performance on datasets. Incorporating multi-emotion guidance, EASL achieves 5.26% BLEU-4 improvement over GEN-OBT on PHOENIX14T and outperforms SignLLM-M by 5.63% on Prompt2Sign. Lower ROUGE-L on Prompt2Sign reflects our emphasis on semantic-affective accuracy rather than word matching.

3.3. Ablation Study

Table 2 validates each component's effectiveness. Removing the three-stage training causes the most severe degradation (BLEU-4 -3.55, ROUGE-L -5.06, MAE +3.44), confirming that progressive semantic-emotion disentanglement prevents

Table 1: Performance Comparison on Benchmark Datasets (BLEU1-4 and ROUGE-L).

(a) PHOENIX14T

Method	B-1	B-2	B-3	B-4	R-L
PT-base	9.47	3.37	1.47	0.59	8.88
PT-FP&GN	13.35	7.29	5.33	4.31	13.17
DET	17.18	10.39	7.39	5.76	17.64
GEN-OBT	23.08	14.91	10.48	8.01	23.49
EASL	**37.46**	**20.13**	**14.97**	**13.27**	**33.00**
	(+14.38)	**(+5.22)**	**(+4.49)**	**(+5.26)**	**(+9.51)**

(b) Prompt2Sign

Method	B-1	B-2	B-3	B-4	R-L
NSLP-G	17.55	11.62	8.21	5.75	31.98
Fast-SLP	39.46	23.38	17.35	12.85	46.89
NSA	41.31	25.44	18.25	13.12	47.55
SignLLM-M	43.40	25.72	19.08	14.13	51.57
EASL	**50.00**	**32.61**	**24.51**	**19.76**	**47.00**
	(+6.60)	**(+6.89)**	**(+5.43)**	**(+5.63)**	**(-4.57)**

Table 2: Ablation Study Results on Test Sets.†

Configuration	BLEU-4	ROUGE-L	MAE*
Full Model	**16.15**	**38.69**	**5.05**
w/o Three-Phases	12.60 (-3.55)	33.63 (-5.06)	8.49 (+3.44)
w/o E_{DESE}	14.81 (-1.34)	35.78 (-2.91)	6.01 (+0.96)
w/o E_{EGSID}	14.77 (-1.38)	36.49 (-2.20)	6.47 (+1.42)
w/o E_{DESE} & E_{EGSID}	13.56 (-2.59)	34.57 (-4.12)	7.52 (+2.47)

†The current ablation experiments fully mix both datasets to simulate real-world scenarios.
*MAE represents the average across 7 mixed emotion categories.
Note: E_{DESE} denotes the emotion component in DESE; E_{EGSID} denotes the emotion-guided multi-head attention component in EGSID.

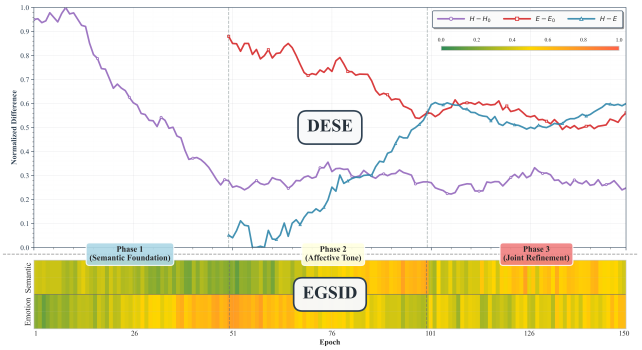

Fig. 3: Emotion similarity $\rho(E_t, E_{BERT})$ across epochs.

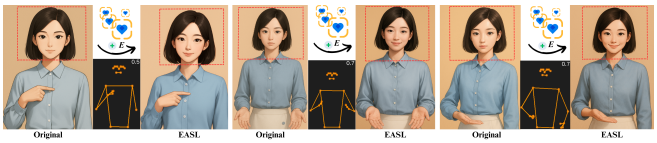

Fig. 4: Case: I am very pleased ... to meet you today.

feature entanglement. Ablating DESE's emotion branch reduces BLEU-4 by 1.34 and increases MAE by 0.96, indicating global emotion representations provide essential decoding constraints. Removing EGSID's emotion-guided attention decreases BLEU-4 by 1.38 with MAE increasing by 1.42, demonstrating its primary role in emotion regression refinement. Simultaneously removing both emotion mechanisms results in substantial performance drops (BLEU-4 -2.59, ROUGE-L -4.12, MAE +2.47), confirming their complementary necessity—DESE enables global emotional modeling while EGSID facilitates local emotional refinement, together achieving optimal semantic-emotion integration.

3.4. Analysis

We evaluate emotion-semantic disentanglement effectiveness using cosine similarity $\rho = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{z}_t^\top \mathbf{Z}_{\text{BERT}}}{\|\mathbf{z}_t\|_2 \, \|\mathbf{Z}_{\text{BERT}}\|_2}$, where scores ($H - H_0$, $E - E_0$, $H - E$) are normalized to [0,1].

Figure 3 illustrates progressive training effectiveness. During **Phase 1**, DESE's semantic representation H aligns with BERT semantic embeddings. In **Phase 2**, the emotion representation E converges toward emotion-specific BERT embeddings. After freezing DESE parameters in **Phase 3**, the divergence between H and E stabilizes, confirming emotion-semantic disentanglement. The lower panel shows EGSID's attention patterns evolve throughout training: initially focusing on semantic features, gradually incorporating emotional signals in **Phase 2**, and achieving balanced semantic-emotion integration in the final phase. These convergence patterns validate our training strategy's effectiveness.

3.5. Case Study

We deploy EASL to **Stable Diffusion** [18]. Figure 4 shows representative frames from the generated sequences that demonstrate accurate poses and, compared to traditional methods, vividly convey the semantic context with appropriate emotional expression (Happy).

4. CONCLUSION

This paper introduces EASL, a multi-emotion guided semantic disentanglement framework with progressive three-stage training for expressive sign language generation. Experimental results demonstrate EASL significantly outperforms existing methods in pose accuracy. Ablation studies validate the progressive training effectiveness and confirm the complementary roles of DESE (global emotional constraints) and EGSID (fine-grained emotion-semantic interaction). Furthermore, EASL successfully adapts to diffusion models, generating accurate poses that vividly convey semantic context with appropriate emotional expressions, significantly enhancing naturalness over traditional methods.

5. ACKNOWLEDGEMENTS

This research is supported by the National Key R&D Program of China (Grant No. 2021YFF0901502).

6. REFERENCES

[1] Nada Shahin and Leila Ismail, "From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: survey, taxonomy and performance evaluation," *Artificial Intelligence Review*, vol. 57, no. 10, pp. 271, 2024.

[2] Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto, "Deep learning, reinforcement learning, and world models," *Neural Networks*, vol. 152, pp. 267–275, 2022.

[3] Jan Zelinka and Jakub Kanis, "Neural sign language synthesis: Words are our glosses," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 3395–3403.

[4] Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani, "Including facial expressions in contextual embeddings for sign language generation," in *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, 2023, pp. 1–10.

[5] Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea, "Towards ai-driven sign language generation with non-manual markers," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

[6] Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang, "Gloss semantic-enhanced network with online back-translation for sign language production," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5630–5638.

[7] Xiaohan Ma, Rize Jin, and Tae-Sun Chung, "Multi-channel spatio-temporal transformer for sign language production," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 11699–11712.

[8] Pan Xie, Taiying Peng, Yao Du, and Qipeng Zhang, "Sign language production with latent motion transformer," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3024–3034.

[9] Oscar Koller, Jens Forster, and Hermann Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," in *Computer Vision and Image Understanding*, 2015, vol. 141, pp. 108–125.

[10] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden, "Adversarial training for multi-channel sign language production," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[11] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden, "Progressive transformers for end-to-end sign language production," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer International Publishing, 2020, pp. 687–705.

[12] Eui Jun Hwang, Jong Hwan Kim, and Jong C Park, "Non-autoregressive sign language production with gaussian space," in *BMVC*, 2021, p. 197.

[13] Shiwei Fang, Chenyang Sui, Yutong Zhou, et al., "Signdiff: Diffusion models for american sign language production," *arXiv preprint arXiv:2308.16082*, 2023.

[14] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, et al., "Neural sign actors: a diffusion model for 3d sign language production from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1985–1995.

[15] Shiwei Fang, Lexing Wang, Chuanxi Zheng, et al., "Signllm: Sign languages production large language models," *arXiv preprint arXiv:2405.10718*, 2024.

[16] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.