AniMask: Anime-Informed Masking for MAE
Semantic-Aware Representation Learning on Stylized Anime Imagery

Team:
Tingting Du (tdu35) Minyuan Zhu (mzhu257)
Xin Chen (xchen2232) Frank Sun (jsun373)
CS566 Project Proposal

1) Problem

We study *semantic-aware masking* for masked image modeling (MIM) on anime artworks —i.e., using semantic cues specific to anime visual characteristics rather than purely random selection to decide *which* patches to mask, *how much* to mask, and *in what schedule*. Concretely, we augment Masked Autoencoders (MAE) [1] with three strategies: (S1) masking focusing on character features (eyes, hair, facial features); (S2) character/background semantic masking accounting for anime's distinct foreground-background separation; (S3) part-aware curriculum masking that progressively learns from simple elements (line art, solid colors) to complex details (shading, effects, fine details). We target regimes common in course projects: **special data** (anime-captions) and **limited compute** (single-GPU).

2) Why it matters

Random masking is the default in MIM [1, 2], but anime images possess highly structured visual hierarchies with distinct semantic elements. Unlike natural images, anime artwork features stylized, spatially concentrated semantic regions (character faces, eyes, distinctive hairstyles) and clear foreground-background separation with minimal texture in backgrounds. Guiding masks with anime-specific semantics (character-focused attention, line art saliency, anatomical parts) may (i) improve representation quality under the same budget by prioritizing semantically rich regions over flat backgrounds, (ii) speed up convergence to target linear-probe accuracy by learning character-defining features early, and (iii) yield more robust features for downstream tasks common in anime domains (character recognition, style classification, attribute tagging). The outcome is valuable for reproducible education settings, for practitioners constrained by time/GPU, and for the growing application of self-supervised learning in creative and artistic domains where labeled data is scarce but semantic structure is pronounced.

3) State-of-the-art (gap)

Pixel-regression MIM such as MAE/SimMIM [1, 2] and token-prediction MIM such as BEiT/iBOT [3, 4] are strong baselines, primarily validated on natural image datasets (ImageNet). Several works explore semantic guidance: attention-guided masking (AttMask) [5], semantic-part masking with curriculum (SemMAE) [6], and learnable/curriculum masking [7]. However, these approaches have not been systematically evaluated on artistic/stylized domains like anime, where visual semantics differ fundamentally from natural images—featuring exaggerated, localized features (large expressive eyes, distinctive hair), minimal photographic texture, and strong line-art structure. Additionally, there is limited systematic comparison in the small-data, small-compute regime with a unified training/evaluation setup, which is particularly relevant for anime where large-scale curated datasets are scarcer than natural image corpora. We fill this gap by adapting and evaluating semantic-aware masking strategies specifically for anime imagery, providing insights into how self-supervised learning can be optimized for non-photorealistic, highly stylized visual domains.

4) Approach: baseline → three semantic masking variants

Baseline. We re-implement MAE [1] with ViT-S/16 on anime-captions (or other anime dataset), masking ratio $r \in \{0.5, 0.75\}$, lightweight decoder, and report linear-probe/top-k on the frozen encoder.

Definition (Semantic-aware masking). Instead of masking patches uniformly at random, we use semantic signals to select patches: attention heatmaps, saliency/weak segmentation masks, or self-supervised part maps.

S1: Attention-guided masking (AttMask-style). Compute teacher ViT attention (e.g., DINO/iBOT/MAE teacher). Sample masks biased to *high-attention* ("mask-hard") or *low-attention* ("mask-easy") areas [5].

- **H1 (falsifiable).** Under equal pre-train budget, *mask-high-attention* yields $\geq$ **+1.0** absolute Top-1 (linear probe) vs. random masking on anime-captions with ViT-S/16 @ 75% mask; and reaches 75% Top-1 with $\geq$ **10% fewer epochs**.

S2: Foreground/background semantic masking. Obtain a binary semantic map M via saliency or weak segmentation. Apply *foreground-heavy masking* (force model to reconstruct objects from context) and *background-heavy masking* (reduce reconstruction noise) as two distinct curricula.

- **H2 (falsifiable).** *Background-heavy* masking improves **stability** (lower variance across seeds) and **time-to-threshold** (epochs to 70% Top-1) vs. random masking, at equal final accuracy; *foreground-heavy* improves final linear probe by $\geq$ **+0.5** Top-1 when $r = 0.75$.

S3: Part-aware curriculum masking (SemMAE-style). Derive *semantic parts* from ViT attention (self-supervised) and gradually increase task difficulty: start by masking a portion *within* each part, then mask entire parts [6].

- **H3 (falsifiable).** Part-aware curriculum reduces **epochs-to-X%** by $\geq$ **15%** vs. random masking, and improves robustness under synthetic occlusion (Cutout $p = 0.3$) by **+1.0** Top-1 in linear probe.

Ablations (shared across S1–S3). Masking ratio $r \in \{0.5, 0.75, 0.9\}$; decoder depth $\in \{1, 2, 4\}$; attention source (DINO vs. MAE-teacher); saliency quality (coarse vs. refined); curriculum schedule (epochs per stage).

5) Evaluation plan

Datasets. Anime-captions (default), Anime_Characters (optional), few-shot-anime-face (fallback).

Metrics. (i) *Representation*: Top-1/Top-5 linear probe (frozen encoder), k-NN accuracy; (ii) *Efficiency*: wall-clock hours, epochs-to-threshold (70/75% Top-1), peak memory; (iii) *Robustness*: accuracy under Cutout occlusion; (iv) *Qualitative*: reconstructions and attention maps.

Baselines. MAE random masking [1]; SimMIM [2]. (Optional token-prediction refs: BEiT [3], iBOT [4] for discussion.)

Compute. Single 32 GB GPU; ViT-S/16; pre-train 50–100 epochs (anime-captions).

Deliverables. Reproducible code & scripts, `wandb`/TensorBoard logs, figures (learning curves, ablations), and a short demo notebook.

6) Timeline (8 weeks)

W1	Setup & Baseline: data pipelines (anime-captions), MAE baseline (random masking) runs $\geq$30 ep; fix optimizer/schedule; draft plotting scripts.
W2	Implement S1 (attention-guided): extract teacher attention (DINO or MAE-teacher); run *mask-high/low* pilots @ 50 ep; choose r and decoder depth.
W3	S1 Ablations: sweep $r \in \{0.5, 0.75, 0.9\}$ and decoder depths; 3 seeds for stability; measure epochs-to-threshold (70%).
W4	Implement S2 (FG/BG masking): integrate saliency/weak segmentation; run FG-heavy vs BG-heavy curricula; collect reconstructions.
W5	S2 Ablations & Robustness: stability (3 seeds), Cutout robustness, variance analysis; pick best S1/S2 settings.
W6	Implement S3 (part-aware curriculum): derive part maps from attention; stage-wise curriculum; pilots @ 50 ep.
W7	Consolidation: full runs of S1–S3 with chosen configs; optional small few-shot-anime-face confirmation; finalize all curves/tables.
W8	Writing & Demo: compile qualitative figs (attn + recon); ablation tables; threats-to-validity; code cleanup; demo notebook/video.

Risks & fallbacks

R1 Saliency/part quality too noisy → start with attention-only S1 (no external model), or use simple salient object detector; **R2** Training unstable/time-constrained → shrink to ViT-T/16 and 30–50 ep; **R3** Effects marginal → increase mask ratio to 0.9 (harder task) or switch to CL-MAE-style learnable/curriculum masking [7] for a clearer delta. We also need to do adaptation of teacher ViT attention on anime images, because they are trained on natural images. Moreover, we have to redesign the loss function to give less weight to pure color areas.

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *CVPR*, 2022.

[2] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, “Simmim: A simple framework for masked image modeling,” in *CVPR*, 2022.

[3] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: Bert pre-training of image transformers,” *arXiv:2106.08254*, 2021.

[4] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, “ibot: Image bert pre-training with online tokenizer,” *arXiv:2111.07832*, 2021.

[5] I. Kakogeorgiou, N. Nikolaidis, E. Papalexakis, N. Papamarkos, A. Tefas, and I. Pitas, “What to hide from your students: Attention-guided masked image modeling,” in *ECCV*, 2022. “AttMask” attention-guided masking.

[6] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng, “Semmae: Semantic-guided masking for learning masked autoencoders,” in *NeurIPS*, 2022.

[7] N. Madan, K. Duarte, Y. S. Rawat, and M. Shah, “Cl-mae: Curriculum-learned masked autoencoders,” in *WACV*, 2024.