MM-DiT ベースの Stable Diffusion 3 による ゼロショット領域分割

山口 廉斗[1,a)]　柳井 啓司[1,b)]

概要

本研究では, U-Net ベースの Stable Diffuison v1, v2 よりも高品質な画像生成能力を持つ Stable Diffusion 3 を用いた, 学習なしのゼロショット領域分割手法を提案し, MM-DiT (Multimodal Diffusion Transformer) ベースのアーキテクチャを持つモデルによるゼロショット領域分割の潜在能力を示す. 現状では U-Net ベースの SD を用いた手法を上回る結果は得られていないが，引き続き，改良を行い高精度化を目指す予定である.

1. はじめに

コンピュータビジョンにおける領域分割は, 画像内の各ピクセルに適切なクラスラベルを割り当てるタスクであり, 画像編集や自動運転, 医療画像解析など幅広いドメインで重要な役割を果たしている. しかし, 画像のピクセルを正確に分類するためには, 従来の教師あり学習アプローチが抱えている大規模なラベル付きアノテーションデータセットが必要という課題がある. ラベルデータの手作業での生成には高いコストがかかり, 実用的な利用が難しいことが頻繁に指摘されている.

この問題に対処するため, 学習なしでの領域分割手法が近年注目を集めている. さらに, モデルが学習データに含まれていない未知のクラスに対する汎化性能の低さも課題として挙げられ, この問題を克服するために少数データやゼロショット学習による領域分割手法の研究が進展している. 特に, ゼロショット学習アプローチは, 未知クラスに対する汎化性能を持ち, ラベルデータの不足を補う効果が期待されている.

最近では, 大規模なテキストと画像のペアデータで事前学習された CLIP (Contrastive Language-Image Pretraining) [8] や Stable Diffusion [9] のようなモデルの事前知識を活用し, テキストと画像の二つのモダリティを組み合わせた革新的な領域分割手法が提案されている. これらのモデルは, それぞれのモダリティの強みを組み合わせることで, 従来の手法では困難であったタスクを効率よく実行する能力を持っている.

U-Net ベースの Stable Diffusion の学習なし領域分割の手法は提案されているが, Stable Diffusion 3 では MM-DiT (Multimodal Diffusion Transformer) が用いられているために, 既存の手法をそのまま適用することができない. そこで, 本研究では, MM-DiT から Attention Map を取り出すことによって Stable Diffusion 3 [2] を用いて, 学習なしのゼロショット領域分割を行う手法を提案し, その有効性と潜在能力を評価する.

2. 関連研究

2.1 大規模モデルを用いたゼロショット領域分割

大規模に事前学習されたテキストと画像の二つのモダリティを扱えるモデルを活用したゼロショット領域分割手法に関する研究が多くなされており, Segment Anything [4] のように, 大規模なアノテーションデータを利用した事前学習済みの領域分割基盤モデルの取り組みや, 事前学習済みの CLIP を使った手法 [1,6], Diffusion Models の高精度な画像生成能力を活用した手法 [3,10–13] が提案されている.

特に, 本部ら [3,13] の StableSeg では, 事前学習済みの Stable Diffusion の U-Net に含まれる Attention Map に注目することで, 学習なし領域分割手法を提案しており, 本研究では, StableSeg における Cross Attention Map の扱い方に着想を得て, Stable Diffusion 3 における領域分割手法を提案する.

2.2 Stable Diffusion 3

Stable Diffusion v1, v2 で Diffusion Model のノイズ除去を担うモデルとして Attention 機構を取り入れた U-Net が用いられていた一方で, Transformer ベースのノイズ除去モデル DiT (Diffusion Transformer) [7] が提案されており, Stable Diffusion v3 [2] では, Flow Matching [5], DiT を活用した MM-DiT (Multimodal Diffusion Transformer) 採用されている. また, 入力プロンプトに対する生成画像の忠実性を向上させるために, 2 つの CLIP モデルと 1 つの

[1] 電気通信大学
[a)] yamaguchi-r@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

T5 モデルの合計 3 つのテキストエンコーダを用いて, 画像とテキストを共に埋め込み, Joint-Attention を計算することで, 高品質な画像生成を実現している.

3. 手法

3.1 手法概要

本研究では, 本部らの StableSeg [3] で行われている Cross-Attention Map を利用するアイディアに触発され, Stable Diffusion 3 の MM-DiT を活用して, 自由な語彙でのゼロショット領域分割を行う手法を提案する. Stable Diffusion 3 では, v1, v2 系列のモデルよりも入力プロンプトの特徴を高度に捉え, テキストに忠実な画像を生成するために 3 つのテキストエンコーダを用い, 画像とテキストの埋め込みを利用した Joint-Attention を通じた MM-DiT の推論を繰り返す構造になっている. 本手法では, 大きく分けて (1) 3 つのテキストエンコーダによるテキスト埋め込みの生成 (2) Joint-Attention を利用した領域分割の 2 つの手順により, MM-DiT の推論を 1 ステップを行うことによって, ゼロショット領域分割を実現する.

3.2 テキスト埋め込みの生成

Stable Diffusion 3 において, テキストエンコーダには CLIP/L-14, CLIP/G-14, T5-XXL の 3 つのモデルが用いられている. 本項では, DiT の推論に利用するためのテキスト埋め込みを生成する方法について述べる. 領域分割の対象を表す k 個のクラスラベルの集合 $C = [c_0, c_1, \dots c_{k-1}]$ が与えられた時, はじめに, 各クラスラベルに “a photo of” のようなプレフィックスプロンプト $p_0, p_1, \dots p_{k-1}$ を付与したものを $C' = [p_0 + c_0, p_1 + c_1, \dots p_{k-1} + c_{k-1}] = [p'_0, p'_1, \dots, p'_{k-1}]$ とする. 次に, 各プロンプト p'_i に対して, CLIP/L-14, CLIP-G14, T5-XXL の 3 つのモデルを用いて, 3 つのテキスト埋め込みを生成する.

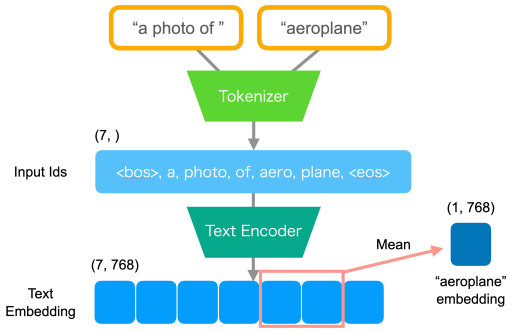

図 1 各クラスカテゴリの埋め込み生成方法

これらの埋め込みから, 図 1 のように各クラスラベル c_i に対応するテキスト埋め込みのみを取得する. CLIP モデルによるトークナイザでは, bos, eos の特殊トークンが含まれることを考慮し, 各プレフィックスプロンプト p'_i のトークン数を n_i とすると, 以下のような処理でクラスに対応するテキスト埋め込みを取得する.

$$\mathcal{E}_{\text{CLIP-L},i} = \text{Sum}\left(\text{CLIP-L}(p'_i)[n_i + 1 : -1]\right) \tag{1}$$

$$\mathcal{E}_{\text{CLIP-G},i} = \text{Sum}\left(\text{CLIP-G}(p'_i)[n_i + 1 : -1]\right) \tag{2}$$

$$\mathcal{E}_{\text{T5},i} = \text{Sum}\left(\text{T5}(p'_i)[n_i : -1]\right) \tag{3}$$

それぞれのモデルにおいて, 0 からクラス数 k 番目までのトークンまで, 上記の処理を繰り返し, k+1 番目から 77 番目までのトークンは pad トークンの埋め込みを挿入したものを, それぞれ, $\mathcal{E}_{\text{CLIP-L}}$, $\mathcal{E}_{\text{CLIP-G}}$, $\mathcal{E}_{\text{T5}}$ とする.

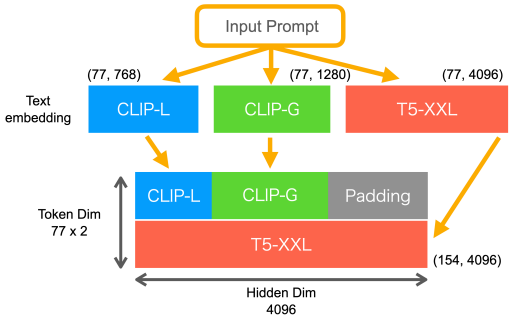

図 2 テキスト埋め込みの生成概要図

これら 3 つの埋め込みに対し, CLIP による 2 つの埋め込みを, 埋め込み次元方向に結合し, T5 の埋め込み次元の大きさに合わせてパディングした埋め込みと, T5 の埋め込みをトークン次元方向に結合した埋め込みを, 最終的に DiT へのテキスト埋め込み $\mathcal{E}'$ とする.

3.3 Joint-Attention を利用した領域分割

入力画像を 1024×1024 の画像とすると, Stable Diffusion 3 の VAE エンコーダによって, 16 チャネルの潜在変数 z が生成される. この潜在変数 z に対して, 2×2 のパッチを適用し, 位置埋め込みを付与した画像特徴を x とする.

Stable Diffusion 3 では, 画像特徴同士やテキスト特徴同士の Self-Attention と画像テキスト特徴間の Cross-Attention を明示的に計算しておらず, それぞれを結合して一つの Query, Key, Value とし, 結合した Joint-Attention を計算している. 画像特徴とテキスト特徴に対応する各 3 つの線形変換層をそれぞれ, l_{IQ}, l_{IK}, l_{IV}, l_{TQ}, l_{TK}, l_{TV} とすると, 以下のような処理で JAMap (Joint-Attention Map) が計算される. ここで, 各線形出力層の入力には, 条件付きテキスト埋め込みのみが使われる.

$$Q = \text{Concat}(l_{IQ}(x),\ l_{TQ}(\mathcal{E}')) \tag{4}$$

$$K = \text{Concat}(l_{IK}(x),\ l_{TK}(\mathcal{E}')) \tag{5}$$

$$V = \text{Concat}(l_{IV}(x),\ l_{TV}(\mathcal{E}')) \tag{6}$$

$$\text{JA Map} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{7}$$

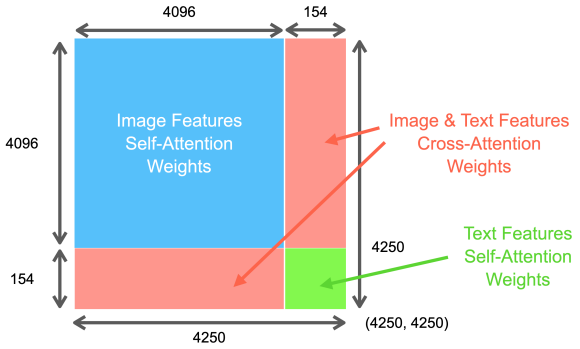

図 3 Joint-Attention Map の概要図

Joint-Attention Map は, 図 3 のように 4 象限に分けることができ, 画像特徴同士の Self-Attention, テキスト特徴同士の Self-Attention, 画像テキスト特徴間の 2 つの Cross-Attention から構成されている. Joint-Attention Map から, 画像とテキスト間の Cross-Attention Map に相当する 2 つの領域を取り出し, 一方を転置させて平均を取ったものを, i 層目の DiT による CAMap (Cross-Attention Map) とする.

$$\mathrm{CAMap}_i = \frac{\mathrm{JAMap}[:, d_i :, : d_i] + \mathrm{JAMap}[:, : d_i, d_i :]^T}{2} \quad (8)$$

ここで, d_i は画像特徴の次元数を表す. i 層目の DiT による Cross-Attention Map は 24 個の Multi-Head Attention によって計算される. 今, 各トークナイザによるトークン次元と画像特徴の次元をそれぞれ d_t, d_i とすると, Head 次元で平均を取った Cross-Attention Map は, $(2 \times d_t, d_i^2)$ の形状をもつため, この Map を $(2, d_t, d_i, d_i)$ に reshape する. ここで, 最初の次元は それぞれ CLIP と T5 による CAMap を表している.

今, k 個のクラスラベルがあるとしたとき, 最終的な Cross-Attention によるセグメンテーションマップ M は以下のように計算される. ここで, CLIP と T5 からクラスラベルに相当するトークンのインデックスを取得する際は, 先述した特殊トークンの影響を考慮する必要があることに留意する.

$$\mathrm{CAMap}_{\mathrm{CLIP},i} = \mathrm{CAMap}[0, 1 : k+1] \quad (9)$$

$$\mathrm{CAMap}_{\mathrm{T5},i} = \mathrm{CAMap}[1, : k] \quad (10)$$

CLIP と T5 による CAMap を取得したのちに, 24 個ある Multi-head Attention の平均をとり, 最後にクラス次元で Argmax を取ることで領域分割マスクを生成する. ここで, 3 つのテキストエンコーダから由来する CAMap の平均マップを CAPM (Cross Attention Probability Map) と呼ぶ.

$$\mathrm{CAPM}_i = \frac{\mathrm{CAMap}_{\mathrm{CLIP}} + \mathrm{CAMap}_{\mathrm{T5}}}{2} \quad (11)$$

$$M = \mathrm{Argmax}\left(\left(\sum_{i=1}^{24} \mathrm{CAPM}_i\right)/24\right) \quad (12)$$

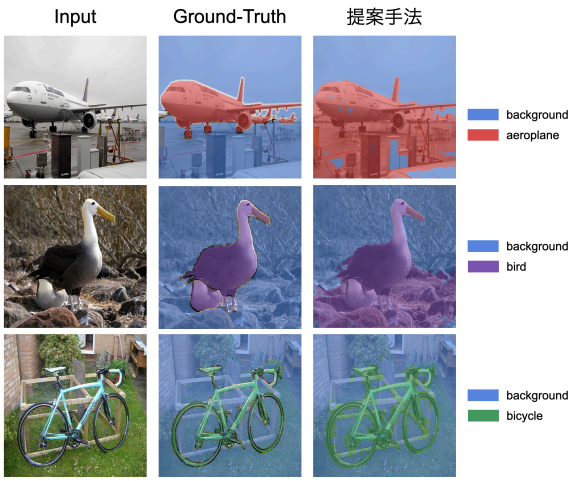

図 4 Pascal VOC による定性評価

4. 実験

4.1 実験設定

実験は学習済みの Stable Diffusion 3 Medium の重みを用いて行い, 本部ら [3] と同様にタイムステップとして, $t = 1$ を採用した. 特に指定がない限り, 入力画像は全て 1024×1024 で固定し, クラスラベルは既知であるとした.

4.2 Pascal VOC による評価

表 1 Pascal VOC による定量評価

手法	提案手法	MaskCLIP [1]	StableSeg [3]
mIoU	36.0	44.7	50.3

Pascal VOC 2012 の領域分割データセットの Val サブセットを用いて, 提案手法の定量評価を行った結果を表 1 に, 定性評価の結果を図 4 に示す. この図を見ると, bicycle のような細かい領域分割が可能であることが分かる. MM-DiT は 128×128 の解像度の潜在変数をパッチで分けた 64×64 の空間で全ての層から Attention Map を抽出することができる. これは, Stable Diffusion v1 で用いられている U-Net の最大スケールが 64×64 であり、本部ら [3] の実験からも 32, 64 の比較的大きいスケールの Attention Map が意味のある情報を含んでいないことが多いことと比較しても, MM-DiT による Attention Map は物体の細かなディティールや外観を捉えることができると考える.

4.3 MM-DiT の層別の定性評価

全 24 層の MM-DiT の中で, 各層の Cross-Attention Map を可視化し他結果, 図 5 のようになった. MM-DiT の前半層と後半層では, 比較的ノイズの多い領域分割が行われていることに対し, 9 層目から 13 層目のような中間層では, クラスラベルに対応した領域分割が行われていることがわかる. これは, 本部ら [3] の研究で U-Net の中間層を領域

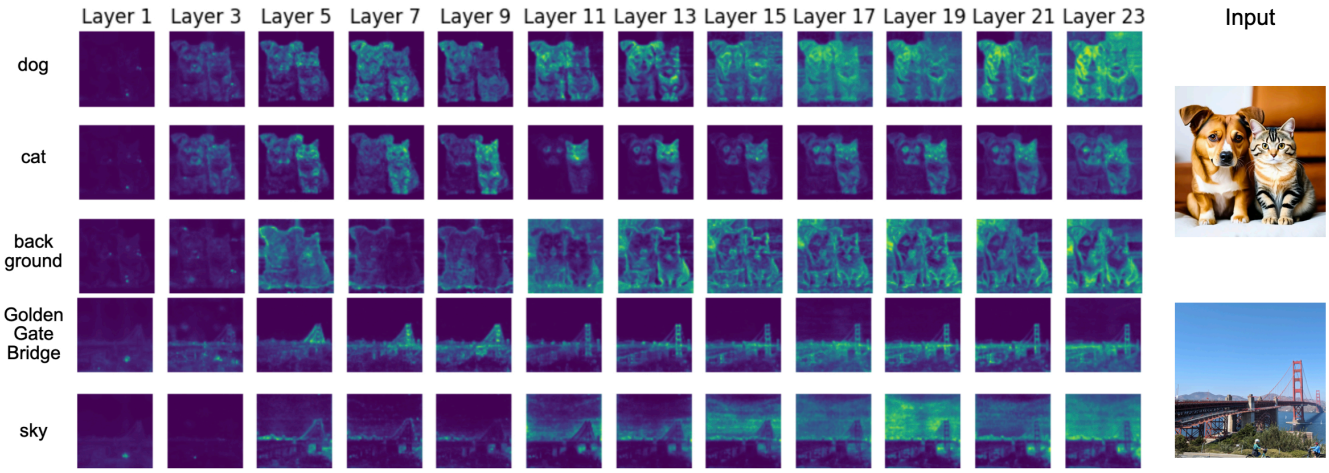

図 5 DiT の層別の領域分割の結果

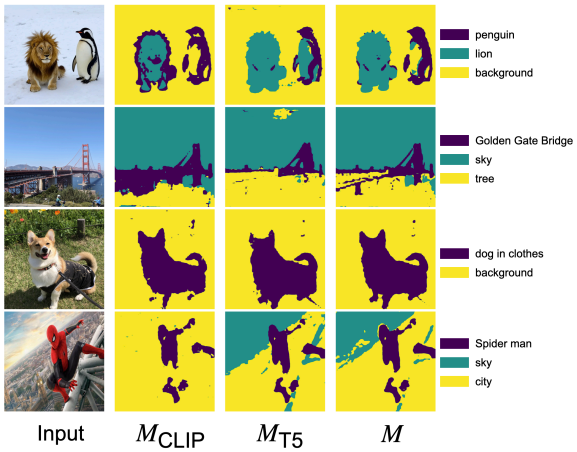

図 6 テキストエンコーダごとの領域分割の結果

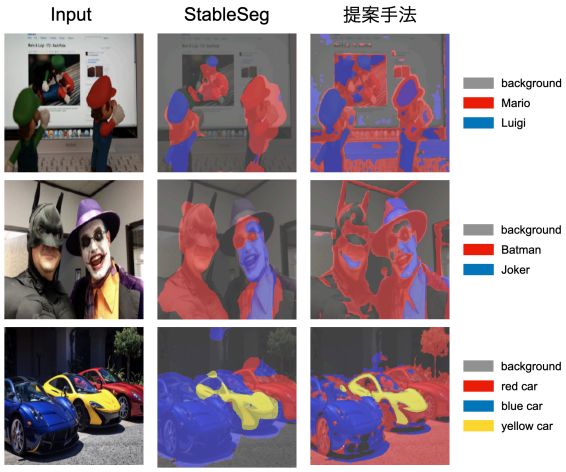

図 7 固有名詞や自由な語彙における既存手法との領域分割結果の比較

分割に利用していることからも分かるように, DiT の中間層ではより画像特徴の意味を含む情報があるからであると推測できる.

4.4 ゼロショット領域分割の定性評価

図 6 に, CLIP の Cross-Attention Map $\text{CAMap}_{\text{CLIP}}$, T5 の Cross-Attention Map CAMap_{T5}, 両者の平均を取った Cross-Attention Map M の領域分割の例を示す. ラベルとして図の上部では, “bear”, “bird”,“background”, 下部では “dog”, “cat”, “background” が与えられている. この結果より, Stable Diffusion 3 の学習が広範な画像テキストペアから学習されているために, 固有名詞や形容詞と名詞の組み合わせといった既存のクラスに縛られない自由な語彙で領域分割が行えていることが分かる.

また, Stable Diffusion v1 モデルを用いた学習なしゼロショット領域分割手法である StableSeg との Web データの領域分割における定性評価の結果を図 7 に示す. StableSeg と同様に “Mario” や “Batman” といった固有名詞の領域分割や, “red car” や “blue car” といった色を表す形容詞の違いを区別する能力があることも示している. CLIP と MM-DiT の学習に多様な Web データが用いられていることから, このような既存のクラスラベルに縛られない特殊な語彙での領域分割を可能にしている.

5. まとめ

本研究では, Stable Diffusion 3 を用いたゼロショット領域分割手法を提案し, MM-DiT ベースのアーキテクチャを持つモデルによるゼロショット領域分割の潜在能力を示した. U-Net ベースの Stable Diffusion v1, v2 と同様に, Stable Diffusion 3 においても, DiT 内の Joint-Attention から, 画像特徴とテキスト特徴間の Cross-Attention を利用した領域分割が可能であることを示した. 今後の課題として, DiT の層別・ヘッド別の領域分割結果の定量的な分析や, より高品質な領域分割マスクの作成のためのプロンプトエンジニアリング, Self-Attention を活用した手法の検討することで領域分割の精度を向上させることが挙げられる.

参考文献

[1] Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D. et al.: Maskclip: Masked self-distillation advances contrastive language-image pretraining, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 10995–11005 (2023).

[2] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F. et al.: Scaling rectified flow transformers for high-resolution image synthesis, *Proc. of International Conference on Machine Learning* (2024).

[3] Honbu, Y. and Yanai, K.: Training-Free Region Prediction with Stable Diffusion, *Proc. of the International Multimedia Modeling Conference (MMM)* (2024).

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al.: Segment anything, *Proc.of IEEE International Conference on Computer Vision*, pp. 4015–4026 (2023).

[5] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. and Le, M.: Flow Matching for Generative Modeling, *Proc. of International Conference on Learning Representation* (2022).

[6] Lüddecke, T. and Ecker, A.: Image segmentation using text and image prompts, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 7086–7096 (2022).

[7] Peebles, W. and Xie, S.: Scalable diffusion models with transformers, *Proc.of IEEE International Conference on Computer Vision*, pp. 4195–4205 (2023).

[8] 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, *International conference on machine learning*, pp. 8748–8763 (2021).

[9] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-resolution image synthesis with latent diffusion models, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 10684–10695 (2022).

[10] Tian, J., Aggarwal, L., Colaco, A., Kira, Z. and Gonzalez-Franco, M.: Diffuse, Attend, and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion, *Proc.of IEEE Computer Vision and Pattern Recognition* (2024).

[11] Wu, W., Zhao, Y., Shou, M. Z., Zhou, H. and Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models, *Proc.of IEEE International Conference on Computer Vision*, pp. 1206–1217 (2023).

[12] 吉橋亮太，大塚雄也，土井賢治，田中智大：アテンションはアノテーションの代わりになるか？：テキスト－画像生成モデルの注意機構を利用した領域分割の弱教師あり学習，画像の認識・理解シンポジウム (MIRU) (2023).

[13] 本部勇真，山口廉斗，柳井啓司：StableSeg: Stable Diffusion によるゼロショット領域分割，画像の認識・理解シンポジウム (MIRU) (2023).