# Generate, Transduct, Adapt: Iterative Transduction with VLMs

## Supplementary Material

## 9. Additional Ablations

We explore additional ablative studies over GTA-CLIP and its various components in this section.

**Per Dataset Results.** Table 6 breaks down the Top-1 accuracies across reported in Table 4 of the main papers across individual datasets namely CUB, Stanford Cars, FGVC Aircraft, Flowers102, and Food101 datasets using the ViT-B/16 architecture. We observe similar trends for each dataset for all the ablations considered.

**Using MetaCLIP as the base VLM.** MetaCLIP introduces better CLIP architectures by curating training data and scaling training. We switch the base VLM from CLIP to MetaCLIP to take advantage of this and test the generalization of our approach to new architectures. Table 7 presents the accuracies for the inductive version of MetaCLIP, TransMetaCLIP (the TransCLIP method applied to MetaCLIP), and GTA-MetaCLIP (our method applied to MetaCLIP). The experiments are conducted using the ViT-B/16 architecture of MetaCLIP across CUB, Stanford Cars, FGVC Aircraft, Flowers102, and Food101 datasets. We observe consistent improvements in the case of MetaCLIP too. On average, over the five datasets, we see an improvement of 6.8% over MetaCLIP, and an improvement of 2.7% over TransMetaCLIP on using our method. This is similar to our improvements of 7.0% and 3.3% on the corresponding baselines with CLIP.

**Effect of the LLM Model in GENERATEATTRIBUTES.** For all the results in the main paper, we used Llama-3.1 as the LLM model for dynamic attribute generation. Now we explore using GPT4o as the LLM in Table 8. We observe that the accuracy remains similar on average over the CUB, Stanford Cars, FGVC Aircraft, Flowers102, and Food101 datasets on ViT-B/16. Thus, using Llama-3.1 is a more cost-effective choice for dynamic attribute generation due to its open-source nature.

**Removing the internal call to TRANSDUCT in GENERATEATTRIBUTES.** We remove the internal transductive inference update call (see § 3.2) in GENERATEATTRIBUTES and present the results over five datasets using the ViT-B/16 architecture in Table 9. We observe that on average the accuracy drops on removing this call to TRANSDUCT. For Aircraft, we observe that the accuracy slightly improves on dropping this TRANSDUCT call, however for Cars we see a significant decrease.

## 10. Sensitivity Analysis

**Top-$k$ Selection and Number of Iterations $T$.** In Tab. 10 we show the performance of GTA-CLIP when varying top-$k$ and $T$ selection. The table is divided into two sections: first we fix $T$ and sweep over $k$, and secondly we fix $k$ and sweep over $T$. We find that increasing $T$ has the strongest correlation with performance, with average performance across benchmarks monotonically increasing for $k = 8$ when going from $T = 1$ to $T = 50$. Furthermore, we find that Flowers and Food are the most insensitive to changes in hyperparameters, keeping mostly the same value irrespective of $k$ and $T$. Overall, we find the performance guarantees to be quite high even in the worst case (65.89 with $k = 8, T = 1$), still being higher than default TransCLIP (64.26) or TransCLIP with static fine-grained attributes (65.60).

**Probability Threshold $\alpha$.** Similarly, in Tab. 11 we show the performance of GTA-CLIP when varying the probability threshold for determining confusing pairs of classes, $\alpha$. For the whole experiment, we fix $k = 8, T = 30$ and sweep over $alpha$. We find that each benchmark has it's

Table 6. **Per-dataset results of Ablation Study.** For five datasets on the ViT-B/16 architecture, we present the effect of various components of GTA-CLIP. We use the same conventions as Table 4.

| ATTRIBUTES | TRANSDUCT | ADAPT | CUB | Cars | Aircraft | Flower | Food | **Average** |
|---|---|---|---|---|---|---|---|---|
| Ø | ✗ | ✗ | 55.20 | 65.38 | 24.75 | 71.38 | 86.10 | 60.56 |
| S | ✗ | ✗ | 57.70 | 65.65 | 24.78 | 73.33 | 86.50 | 61.59 |
| Ø | ✓ | ✗ | 62.23 | 68.87 | 26.88 | 76.17 | 87.15 | 64.26 |
| S | ✓ | ✗ | 64.15 | 69.83 | 26.73 | 80.06 | 87.25 | 65.60 |
| S | ✓ | ✓ | 65.86 | 71.33 | 28.62 | 80.67 | 87.30 | 66.76 |
| D | ✓ | ✗ | 64.20 | 69.53 | 26.58 | 80.23 | 87.27 | 65.56 |
| D | ✓ | ✓ | **66.76** | **72.09** | **29.31** | **82.05** | **87.38** | **67.52** |

Table 7. **Performance with MetaCLIP.** We change the base VLM from CLIP to MetaCLIP for TransCLIP and GTA-CLIP and observe consistent improvements over the baselines on ViT-B/16

| Method | CUB | Cars | Aircraft | Flower | Food | **Average** |
|---|---|---|---|---|---|---|
| CLIP | 55.20 | 65.38 | 24.75 | 71.38 | 86.10 | 60.56 |
| TransCLIP | 62.23 | 68.87 | 26.88 | 76.17 | 87.15 | 64.26 |
| GTA-CLIP | **66.76** | **72.09** | **29.31** | **82.05** | **87.38** | **67.52** |
| MetaCLIP | 68.67 | 74.49 | 28.65 | 73.81 | 84.01 | 65.93 |
| TransMetaCLIP | 74.02 | 79.01 | 31.56 | 80.15 | 85.53 | 70.05 |
| GTA-MetaCLIP | **78.36** | **82.30** | **35.58** | **81.57** | **85.98** | **72.76** |

Table 8. **Effect of LLM model on accuracy of GTA-CLIP.** We switch the LLM model used by GENERATEATTRIBUTES from Llama-3.1 to GPT4o and observe similar performance on ViT-B/16.

| LLM | CUB | Cars | Aircraft | Flower | Food | **Average** |
|---|---|---|---|---|---|---|
| GPT4o | 66.50 | 72.13 | 29.89 | 81.55 | 87.36 | 67.49 |
| Llama-3.1 | 66.76 | 72.09 | 29.31 | 82.05 | 87.38 | 67.52 |

Table 9. **Removing the internal transductive update step in GENERATEATTRIBUTES**, thereby making only a single call to TRANSDUCT per iteration reduces the accuracy on average over five datasets on the ViT-B/16 architecture.

| LLM | CUB | Cars | Aircraft | Flower | Food | **Average** |
|---|---|---|---|---|---|---|
| GTA-CLIP *single* TRANSDUCT | 66.72 | 69.89 | 29.55 | 81.32 | 87.32 | 66.96 |
| GTA-CLIP *original* | 66.76 | 72.09 | 29.31 | 82.05 | 87.38 | 67.52 |

Table 10. **Sensitivity analysis over the top-$k$ and $T$ selection** of GTA-CLIP using the CLIP ViT-B/16 architecture without the dynamic GENERATEATTRIBUTES component (ie. TransCLIP$^{FT}$ in Tab. 1) as given in Algorithm 1. We pick $k = 8, T = 30$ even though there exist better performing alternatives. We fix this hyperparameter selection to ablate on the remaining parameters of GTA-CLIP.

| **top-$k$** | $T$ | **CUB [47]** | **Cars [17]** | **Aircrafts [23]** | **Flowers [31]** | **Food [5]** | **Average** |
|---|---|---|---|---|---|---|---|
| 1 | 30 | 65.93 | 71.55 | **29.43** | 81.04 | 87.36 | 67.06 |
| 3 | 30 | 65.64 | **71.97** | 28.95 | **82.01** | **87.43** | **67.20** |
| 5 | 30 | 65.64 | **71.97** | 28.74 | 81.28 | 87.39 | 67.00 |
| 8 | 30 | 65.86 | 71.33 | 28.62 | 80.67 | 87.30 | 66.76 |
| 10 | 30 | 65.84 | 71.45 | 28.53 | 81.04 | **87.43** | 66.86 |
| **20** | 30 | **66.09** | **71.97** | 28.29 | 81.04 | 87.36 | 66.95 |
| 8 | 1 | 63.98 | 69.87 | 27.48 | 80.76 | 87.37 | 65.89 |
| 8 | 10 | 65.05 | 70.55 | 28.17 | 81.04 | 87.36 | 66.43 |
| 8 | 20 | 65.48 | 71.11 | 28.47 | 81.04 | **87.43** | 66.71 |
| 8 | 30 | 65.86 | 71.33 | 28.62 | 80.67 | 87.30 | 66.76 |
| 8 | 40 | **66.14** | 72.63 | 28.80 | 81.04 | 87.34 | 67.19 |
| 8 | 50 | **66.14** | **72.71** | **28.98** | **82.01** | **87.43** | **67.46** |

Table 11. **Ablation over the probability threshold** $\alpha$ of the GENERATEATTRIBUTES implementation of GTA-CLIP as given in Sec. 4 using $k = 8, T = 30$ as determiend from Tab. 10. Like Tab. 10, even though there are better performing selection, we choose $\alpha = 10\%$.

| $\alpha$ | **CUB [47]** | **Cars [17]** | **Aircrafts [23]** | **Flowers [31]** | **Food [5]** | **Average** |
|---|---|---|---|---|---|---|
| 2.5% | 65.67 | 71.68 | 28.50 | **83.23** | 87.27 | 67.27 |
| 5.0% | 66.69 | 71.67 | 28.50 | 81.04 | 87.40 | 67.06 |
| 7.5% | **66.98** | 71.56 | 28.98 | 80.67 | 87.32 | 67.10 |
| 10.0% | 66.76 | 72.09 | **29.31** | 82.05 | 87.38 | 67.52 |
| 12.5% | 65.48 | 72.64 | 28.47 | 82.01 | **87.50** | 67.22 |
| 15.0% | 66.90 | 71.74 | 28.65 | 82.05 | 87.41 | 67.35 |
| 17.5% | 66.83 | 72.65 | 28.83 | 82.42 | 87.29 | **67.61** |
| 20.0% | 66.72 | **72.99** | 28.95 | 80.88 | 87.33 | 67.37 |

own ideal $\alpha$ value, namely that no two benchmark's max performances share an common $alpha$. Surprisingly, we see that $\alpha = 17.5\%$, which does not perform the best on any benchmark, has the highest average value. We also conclude that GTA-CLIP has a greater insensitivity to the choice of $\alpha$ as compared to $T$ but similar to $k$. Namely we find that the spread of $\alpha$ to be $67.61 - 67.06 = 0.55$, $T$ to be $67.46 - 65.89 = 1.57$, and $k$ to be $67.20 - 66.76 = 0.44$. Finally, we find that the minimum performance increase by introducing GENERATEATTRIBUTES is at $\alpha = 5.0\%$ with a gain of $67.10 - 66.76 = 0.34$. In other words, adding any amount of comparative attribute generation improves performance.

## 11. Evolution of Attribute Space

In Fig. 3 through Fig. 7, we depict the evolution of the set of attributes for a given class over the course of our method. GTA-CLIP begins with a list of static fine-grained attributes (depicted in blue) and through iterations of the method generates additional comparative attributes between confusing classes (red). We embed these attributes with the CLIP text tower and use t-SNE to visualize the relative locations of these attributes. The specific prompt generated for a given point is indicated within the figure. We see that attributes within the reduced space often form tight clusters grouped by similar concepts (eg. "habitat" or "appearance"). When dynamically generated attributes (red points) are close to the initial static attributes (blue) we see more similar semantic meaning. Finally, one can notice that the newly added attributes occupy different regions of the space, namely that using dynamic generation effectively expands the list of fine-grained details on a given class.

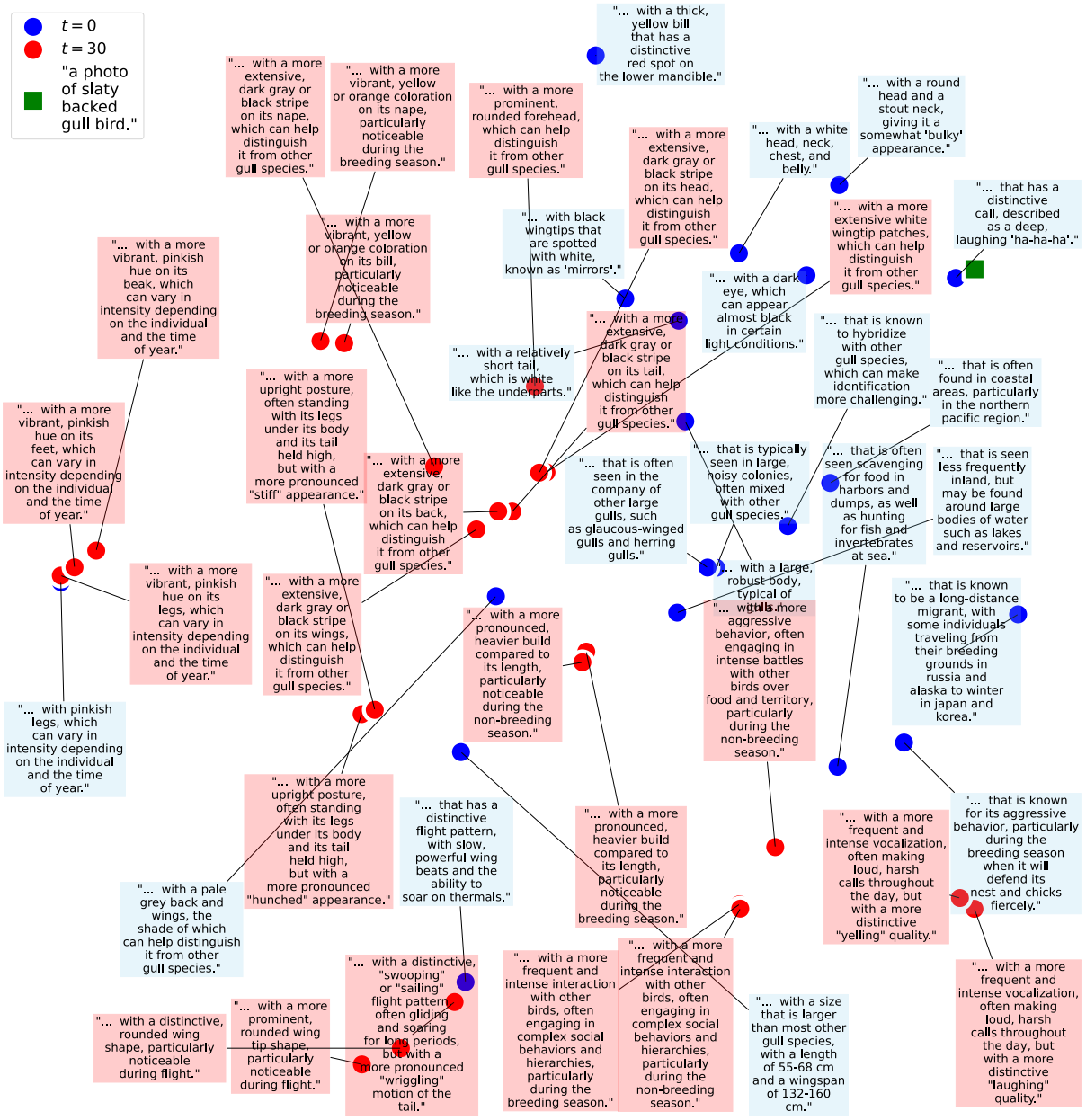


Figure 3. **Slaty-backed Gull (vs. Western Gull) Annotated T-SNE Plot.**

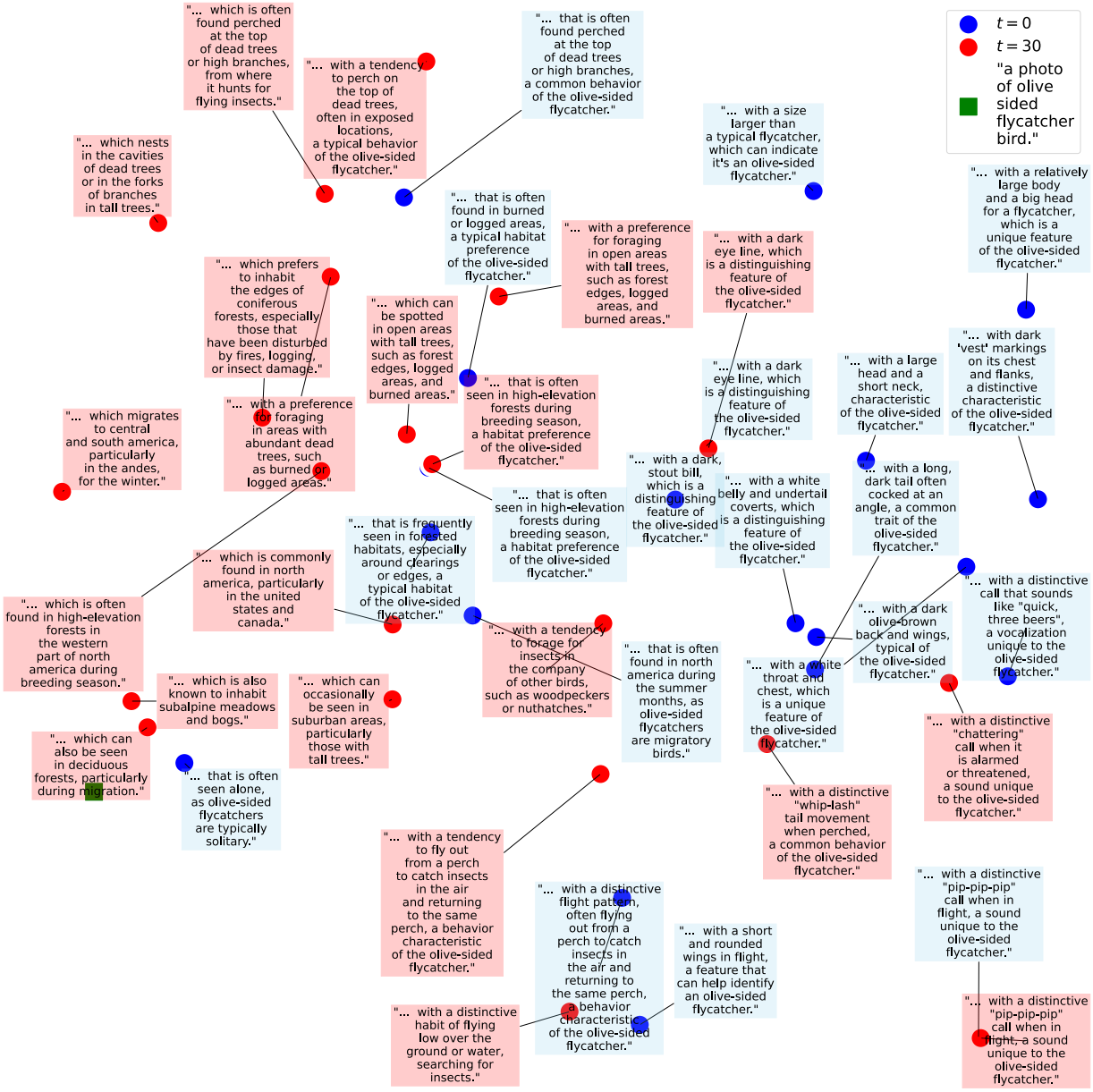


Figure 4. **Olive-sided Flycatcher (vs. Least Flycatcher) Annotated t-SNE Plot.**

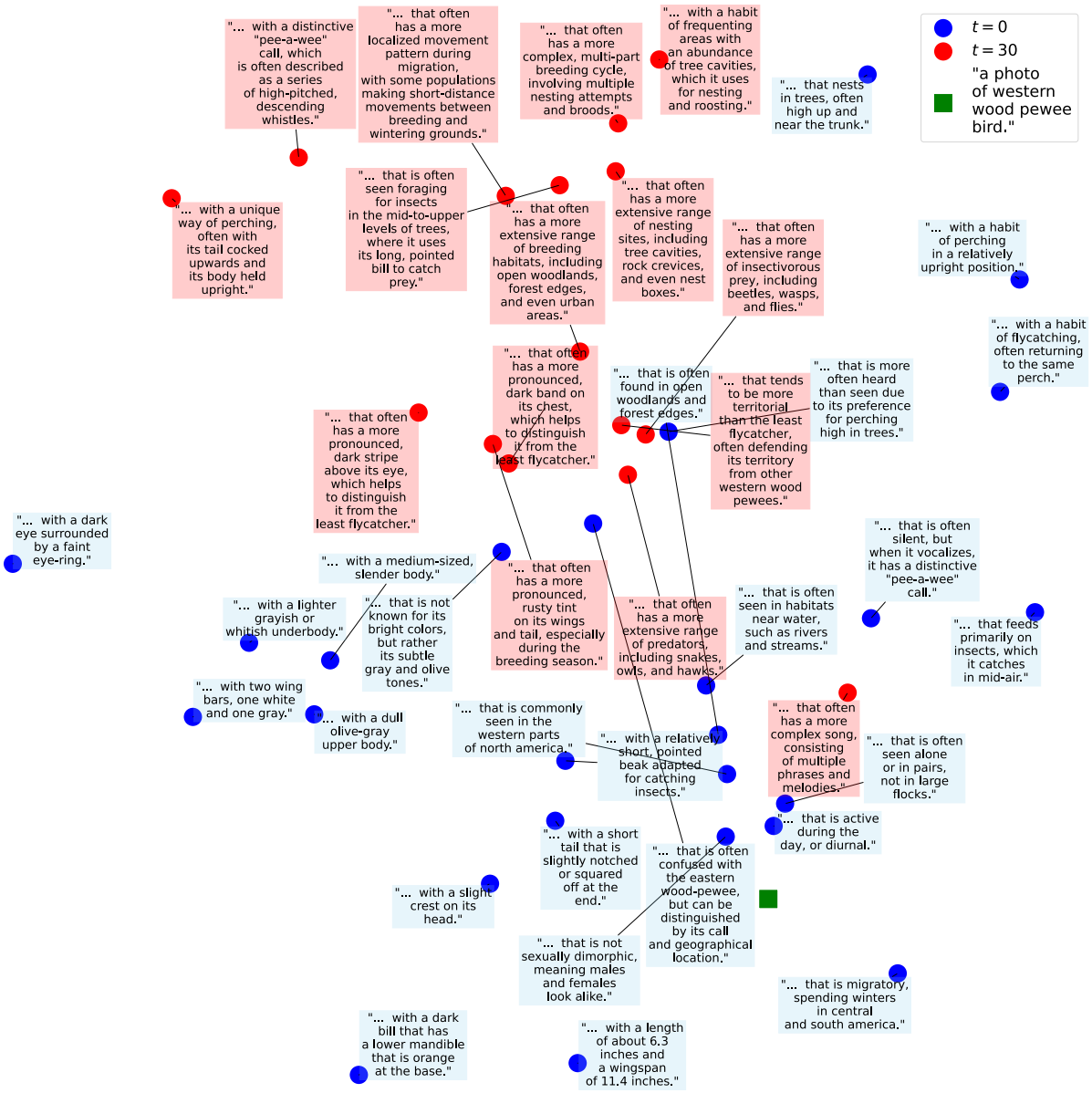


Figure 5. **Western Wood-Pewee (vs. Least Flycatcher) Annotated t-SNE Plot.**

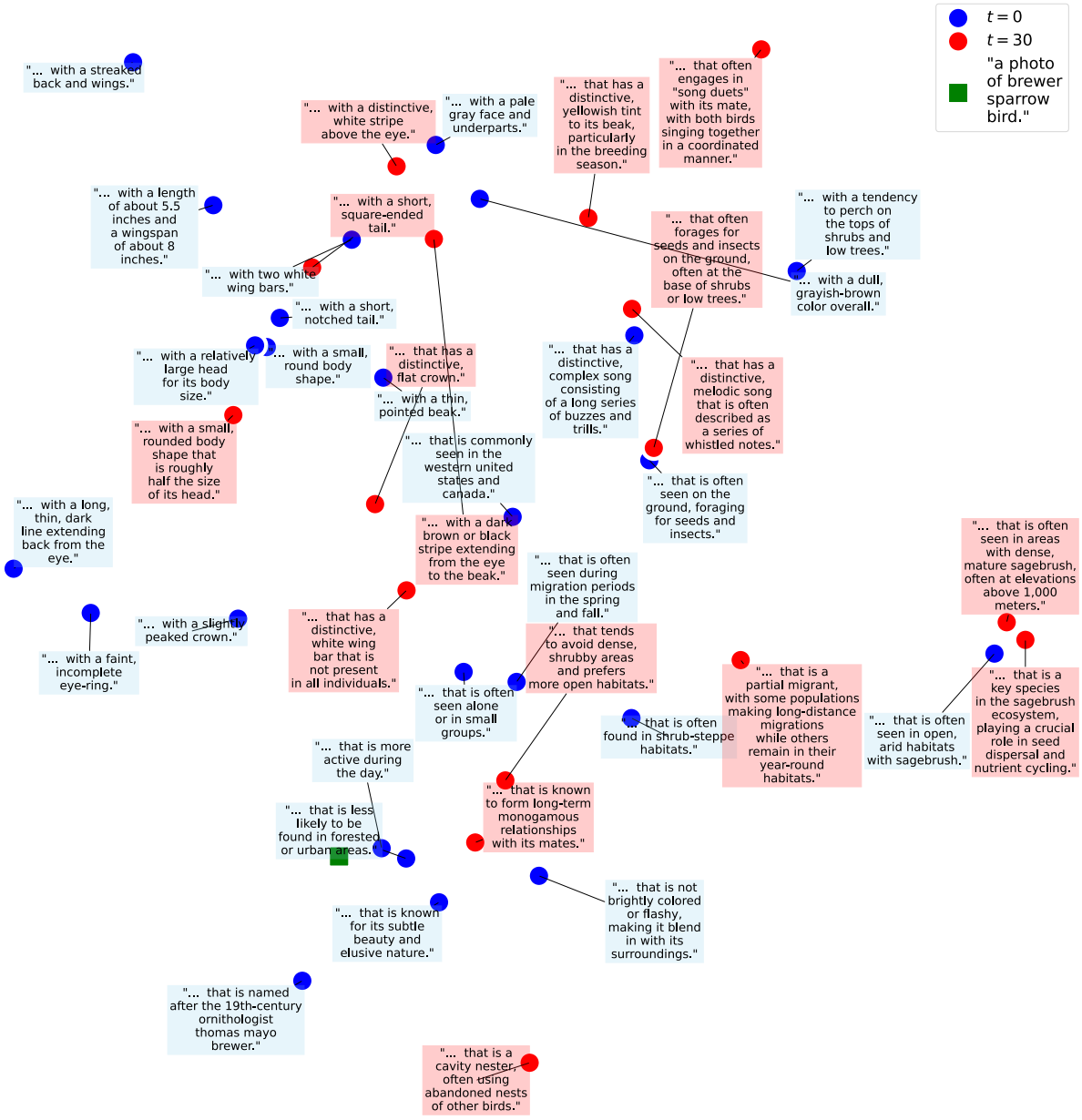


Figure 6. **Brewer's Sparrow (vs. Harris' Sparrow) Annotated t-SNE Plot.**

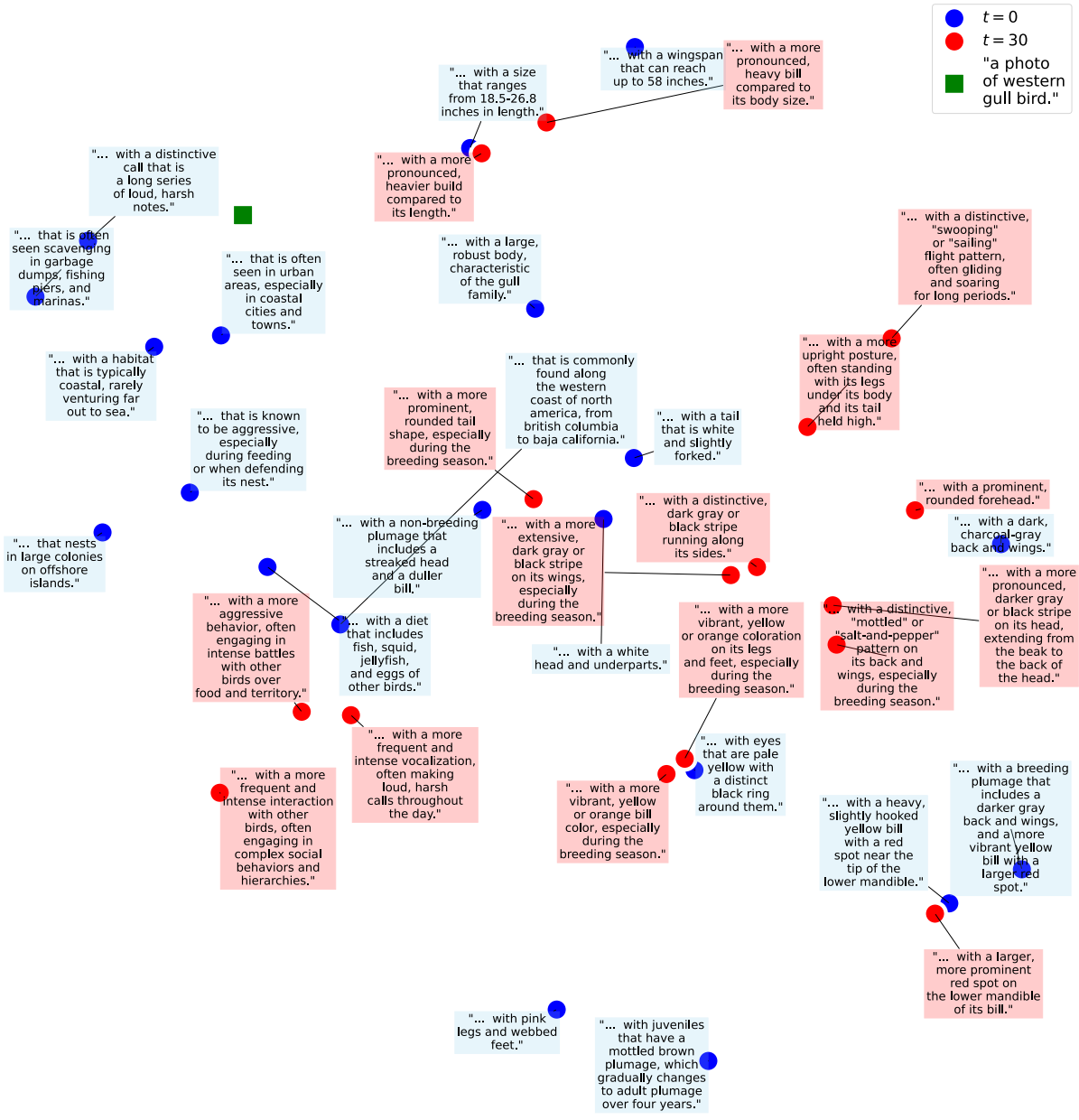


Figure 7. **Western Gull (vs. California Gull) Annotated t-SNE Plot.**