Attention to Trajectory: Trajectory-Aware Open-Vocabulary Tracking

Yunhao Li[1,2] Yifan Jiao[1,2] Dan Meng [4] Heng Fan[3†] Libo Zhang[1†*]
[1]Institute of Software Chinese Academy of Sciences [2]University of Chinese Academy of Sciences
[3]University of North Texas [4]OPPO Research Institute

Abstract

*Open-Vocabulary Multi-Object Tracking (OV-MOT) aims to enable approaches to track objects without being limited to a predefined set of categories. Current OV-MOT methods typically rely primarily on instance-level detection and association, often overlooking trajectory information that is unique and essential for object tracking tasks. Utilizing trajectory information can enhance association stability and classification accuracy, especially in cases of occlusion and category ambiguity, thereby improving adaptability to novel classes. Thus motivated, in this paper we propose **TRACT**, an open-vocabulary tracker that leverages trajectory information to improve both object association and classification in OV-MOT. Specifically, we introduce a Trajectory Consistency Reinforcement (**TCR**) strategy, that benefits tracking performance by improving target identity and category consistency. In addition, we present **TraCLIP**, a plug-and-play trajectory classification module. It integrates Trajectory Feature Aggregation (**TFA**) and Trajectory Semantic Enrichment (**TSE**) strategies to fully leverage trajectory information from visual and language perspectives for enhancing the classification results. Extensive experiments on OV-TAO show that our TRACT significantly improves tracking performance, highlighting trajectory information as a valuable asset for OV-MOT. We will release TRACT at* https://github.com/Nathan-Li123/TRACT.

1. Introduction

Multi-Object Tracking (MOT) is an important task in computer vision, focusing on the detection and tracking of objects within video sequences. It has many key applications, such as autonomous driving, intelligent surveillance, and robotics. Early MOT research primarily concentrates on a few common categories, *e.g.*, pedestrians and vehicles, and later shifts toward tracking a broader range of categories. Recently, as the demand for practical applications grows, Open-Vocabulary MOT [12] is introduced to enable tracking across arbitrary categories, overcoming the limitations imposed by pre-defined tracking categories in training data.

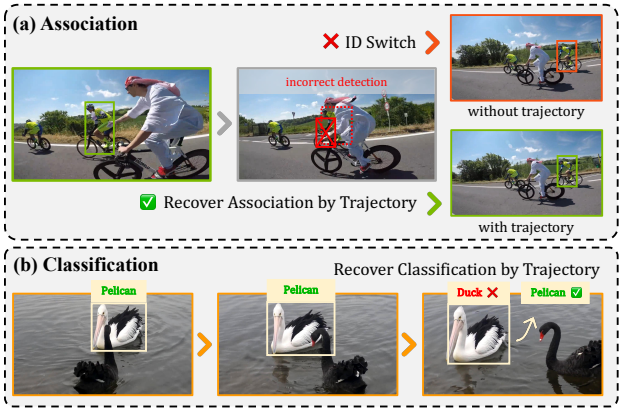


Figure 1. Trajectory information can enhance both association and classification by helping to recover associations disrupted by inaccurate or missed detections (as shown in (a)) and by correcting incorrect classifications (as shown in (b)).

Despite great advancements, current OV-MOT methods are often constrained by a critical limitation: an overwhelming focus on *instance-level* information, with limited attention to *trajectory-level* insights. Specifically, although recent methods have introduced innovative association strategies for open-vocabulary scenarios, they fail to incorporate trajectory information, which is an important cue in videos and widely utilized in classic MOT approaches. This oversight may prevent current OV-MOT approaches from fully leveraging contextual continuity offered by trajectory that is essential to effective tracking, and thus leads to degradation in association and classification (see Fig. 1).

In this context, we rethink the role of trajectory in OV-MOT and apply it for improvement. Currently, OV-MOT

Yunhao Li and Yifan Jiao make equal contributions.
†Equal Advising and Co-last Authors.
*Corresponding author: Libo Zhang (libo@iscas.ac.cn).
Acknowledgement Libo Zhang was supported by National Natural Science Foundation of China (No. 62476266). Heng Fan and his employer were not supported by any financial support for this work.

Please note that, in this paper trajectory information refers to all data related to the trajectory during the tracking process, including its position and classification results from previous frames, among other details.

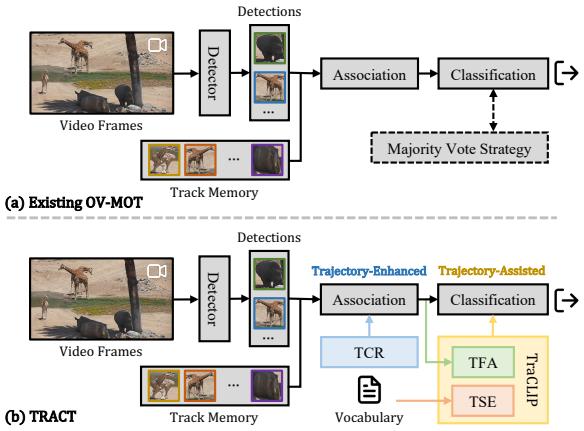


Figure 2. Comparison of the overall pipeline between existing OV-MOT approaches and our TRACT. We introduce three strategies, *i.e.*, TCR, TFA, and TSE strategies, to utilize trajectory information in association and classification.

usually contains three steps, including *localization*, *association*, and *classification*. Since the localization mainly depends on the performance of external detectors, it is hard to directly use trajectory information for enhancing localization (see Sec 4.5 for detailed analysis). However, trajectory information can largely benefit both association and classification, especially in cases with novel classes. For association, the instability of open-vocabulary detection often leads to inaccurate or missed detections in certain frames. In such cases, trajectory information helps recover matches, reducing identity switches (see Fig. 1 (a)). For classification, open-vocabulary systems struggle with frequent blurring and occlusion in objects, causing misclassifications. In this situation, trajectory information can aid in distinguishing between incorrect categories accurately (see Fig. 1 (b)).

Motivated by the above, we propose a novel *Trajectory-aware OV-Tracker* (**TRACT**), a method that comprehensively utilizes trajectory information to improve both object association and classification. We demonstrate the comparison of the overall pipeline between existing OV-MOT approaches and our TRACT in Fig. 2. Built on the top of current mainstream [12–14], it functions as a two-stage tracker that tracks on arbitrary detection results. To better adapt to OV-MOT task, we further divide the tracking stage to association and classification steps (see Section 3.1 for more details). Consequently, the design of TRACT is structured in two key steps: a *Trajectory-Enhanced Association* step and a *Trajectory-Assisted Classification* step. In the initial step, we introduce *Trajectory Consistency Reinforcement* (**TCR**) strategy, to enhance appearance-based matching models to better capture trajectory dynamics. Specifically, we construct a set of feature banks and category banks to retain memory of previous trajectory information, namely, target visual features and category predictions. Such design strengthens model's ability to maintain *identification* and *category* consistency, thereby aiding in association and indirectly supporting open-vocabulary classification.

On the other hand, in the trajectory assisted classification step we introduce **TraCLIP**, a plug-and-play method that leverages trajectory information to directly improve classification accuracy. In video sequences, occlusion and blurriness frequently lead to incomplete visual cues, which complicates tracking and especially classification. Trajectory information, however, captures target features under varying occlusion and blur conditions, thus complementing these incomplete visual cues. Therefore, we first propose *Trajectory Feature Aggregation* (**TFA**) strategy to integrate trajectory features derived from the corresponding detection features. Additionally, since trajectories provide information from multiple viewpoints, trajectory-assisted classification has the potential to offer a more detailed and nuanced understanding of the target compared to image-based classification. In this context, vanilla category names may not be fully or accurately defined, as is typically assumed. We propose *Trajectory Semantic Enrichment* (**TSE**) strategy, which incorporates attribute-based descriptions as an alternative to relying solely on category names, thereby enriching the semantic context and improving classification precision. With TFA and TSE, TraCLIP leverages the image-text alignment capabilities of CLIP [17] to comprehensively utilize trajectory information for classification.

Thorough experiments on the popular open-vocabulary tracking benchmark OV-TAO [12] show the effectiveness of our method, showing satisfactory enhancements in tracking performance in open-vocabulary scenarios. This indicates that trajectory information can effectively contribute to OV-MOT, providing a new research direction. Additionally, this paper aims to encourage researchers to approach the OV-MOT task from a comprehensive video perspective rather than focusing solely on instance-level information.

In summary, in this paper we make the following major contributions: **(i)** We develop an effective open-vocabulary tracker, termed TRACT, which leverages trajectory-level information to enhance association and classification without bells and whistles; **(ii)** We propose a plug-and-play trajectory classification method, termed TraCLIP, and introduce the concept of using trajectory itself for classification in OV-MOT; **(iii)** Extensive experiments demonstrate that our method effectively improves the performance on OV-TAO, in-depth analysis is conducted to provide guidance for future algorithm design.

2. Related Works

2.1. Multi-Object Tracking

Multi-object tracking (MOT) involves detecting and tracking multiple moving objects in a video sequence while maintaining consistent identities across frames. A popu-

lar paradigm in MOT is the "*tracking-by-detection*". This method [2, 6, 16, 21, 28] first performs object detection and then associates detections across frames, forming the basis of many representative methods. In this context, MOT methods often improve their performance by enhancing the detection and matching effectiveness. Another common paradigm is "*joint-detection-and-tracking*" [20, 22, 27], which integrates the tracking and detection into a unified process. Recently, Transformers [19] have been introduced into MOT [7, 18, 26, 29], significantly surpassing previous trackers in terms of performance.

2.2. Open-Vocabulary Detection

Open-Vocabulary Detection (OVD) is an emerging task in object detection that aims to identify and localize object categories that are not encountered during the training phase, particularly in few-shot and zero-shot scenarios. In recent years, significant progress has been made in the field of OVD, leading to the proposal of various new algorithms. OVR-CNN [25], as one of the pioneering works in OVD, successfully applies pretrained vision-language models to detection frameworks, improving recognition capabilities for unseen categories through the integration of image and text. ViLD [8] and RegionCLIP [30] utilize the CLIP [17] model, employing knowledge distillation to learn visual region features from classification-oriented models, thus enhancing adaptability in open-world environments. OV-DETR [24], a novel open-vocabulary detector based on the DETR architecture, reformulates the classification task into a binary matching problem between input queries and referent objects to achieve object detection.

2.3. Open-Vocabulary Multi-Object Tracking

Open-Vocabulary Multi-Object Tracking (OV-MOT) aims to identify, locate, and track dynamic objects unseen during training. Li et al.[12] introduce OVTrack, leveraging vision-language models for classification and association, and enhancing appearance learning via diffusion-based data augmentation. They also restructure the TAO dataset[5] into base and novel classes to establish an OV-MOT benchmark. Building on this, MASA [13] utilizes the Segment Anything Model (SAM)[10] to generate instance-level correspondences through unsupervised learning. More recently, SLAck[14] proposes a unified framework integrating semantic, positional, and appearance cues for early association, removing the need for complex post-processing.

3. Methodology

3.1. Preliminary

In real-world applications, object categories typically follow a long-tailed distribution with a vast vocabulary, reflecting the remarkable diversity that no single dataset can fully encompass. To address this limitation, Li et al. [12] introduced Open-Vocabulary MOT, aiming to bridge the gap between conventional MOT and real-world complexity. The mainstream two-stage implementation process of OV-MOT is demonstrated in Fig. 2. For convenient understanding, in this paper we formulate it as follows.

Following the TBD paradigm [2], we broadly divide the process into two stages, *i.e.*, detection and tracking. In the first stage, given a video with N frames, a replaceable open-vocabulary detector is first utilized to generate a set of detection results $\mathcal{R} = \{\mathbf{b}_i, \mathbf{c}_i, \mathbf{f}_i\}_{i=1}^{N}$, where $\mathbf{b}_i$, $\mathbf{c}_i$, and $\mathbf{f}_i$ respectively denotes the set of bounding boxes, category predictions, and extracted target features of the i^{th} frame.

The second stage is tracking. Unlike conventional MOT, OV-MOT typically involves a highly diverse vocabulary $\mathcal{V}$ of categories, which presents significant challenges for classification. Consequently, open-vocabulary trackers often perform association in a class-agnostic manner, deferring final classification until the acquisition of the complete trajectory. We define the former as the association step and the latter as the classification step. Notably, although trackers perform association in a class-agnostic manner, the classification prediction for each detection is preserved for later processing. Concretely, trackers obtain a set of trajectories $\mathcal{T}$ after the association step. Each trajectory $\mathbf{t} = \{b_i, c_i, f_i\}_{i=1}^{n} \in \mathcal{T}$ consists of a series of linked detection results, where $b = [x, y, w, h]$ denotes the 2D bounding box coordinates, f denotes the visual feature, c is the category prediction, and n is the length of $\mathbf{t}$. Subsequently, in the second step existing trackers utilize the category prediction set $\{c_i\}_{i=1}^{n}$ to decide the final classification result.

3.2. Overview

In this paper we present the *Trajectory-aware OV-Tracker* (**TRACT**), to utilize trajectory information in OV-MOT without bells and whistles. As shown in Fig. 3, we address two steps of its design: 1) *Trajectory-Enhanced Association*: we show how to employ trajectory information while associating the detections in Section 3.3. Note that in this step, trajectory information refers to the temporarily stored trajectory segments during the association process. 2) *Trajectory-Assisted Classification*: existing methods determine the category of a trajectory by voting based on the reserved classification results $\{c_i\}_{i=1}^{n}$. In this paper, we aim to further leverage the trajectory representations $\{f_i\}_{i=1}^{n}$ to assist obtain the classification results. Therefore, we propose TraCLIP to achieve trajectory-level classification, as described in Section 3.4. Lastly, we introduce the training strategy of TRACT in Section 3.5

3.3. Trajectory-Enhanced Association

As analyzed in Section 3.1, in this step all reserved detections are send to association module to obtain object tra-

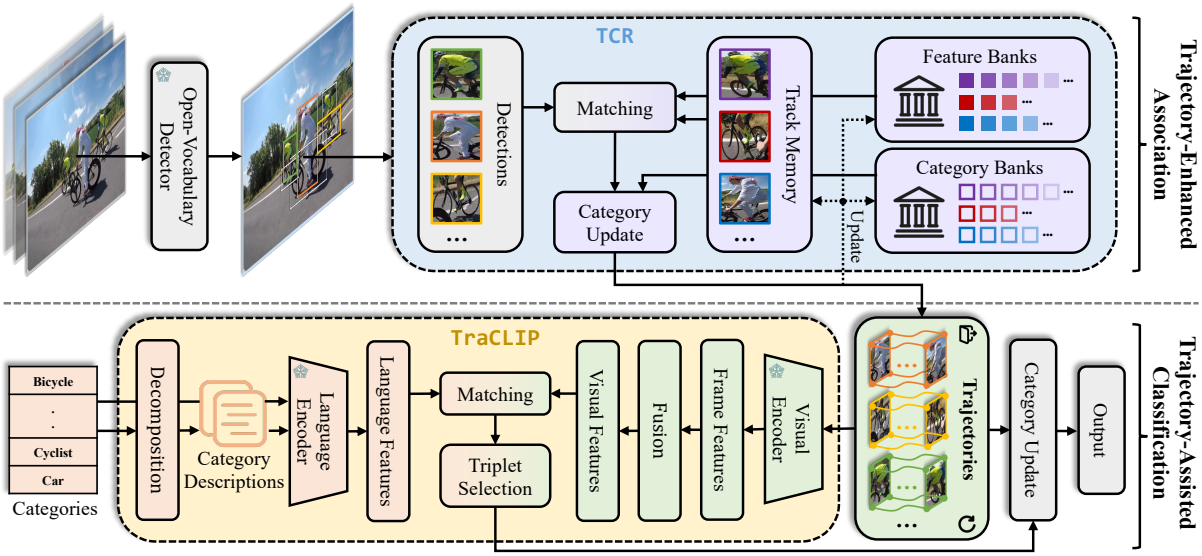


Figure 3. The overall architecture of the proposed TRACT. A replaceable open-vocabulary detector is used to generate boxes of arbitrary categories, and these detection results are used for trajectory association. TRACT leverages trajectory information in both the trajectory-enhanced association and trajectory-assisted classification steps.

jectories $\mathcal{T}$. Based on [12, 13], we adopt an appearance-based matching approach as the core association module of TRACT. Please notice, Although OV-MOT methods, including our TRACT, perform class-agnostic association, they retain the classification prediction for each detection within the trajectory (see Section 3.1). These retained classification predictions are used to determine the final classification result of the trajectory later in the second step. Building upon this, we propose *Trajectory Consistency Reinforcement* (**TCR**) strategy, a method designed to incorporate trajectory information during association. We decompose its functionality into two aspects:

1) Identification consistency. For each trajectory alive in the i^{th} frame, we maintain not only a commonly used trajectory memory $\mathbf{f}$, but also a feature bank $\bar{\mathbf{f}} = \{f_{i-j}\}_{j=1}^{n_{\text{bank}}}$, recording the target feature embeddings f associated to the trajectory from its previous n_{bank} frames. We update trajectory memory $\mathbf{f}$ in a commonly used manner as follows:

$$\mathbf{f}_i = \alpha \times f_i + (1-\alpha) \times \mathbf{f}_{i-1} \tag{1}$$

where $\mathbf{f}_i$ and f_i represents the trajectory memory and target feature of the target in i^{th} frame, and α is the weighting parameter. We then calculate the similarity between each active trajectory $\mathbf{t} \in \mathcal{T}_i$ and each candidate object $r \in \mathcal{R}_i$:

$$\mathtt{S}(\mathbf{t}, r) = \alpha \cdot \Psi(f_i, \mathbf{f}) + (1-\alpha) \cdot \frac{1}{n_{\text{bank}}} \sum_{j=1}^{n_{\text{bank}}} \Psi(f_i, f_{i-j}) \tag{2}$$

where α is the weighting parameter, and $f_i \in \mathbf{f}_i$ denotes the extracted object feature of r. We use both cosine similarity and bi-directional softmax for the similarity calculation function $\Psi(\cdot)$ as in [12]. We derive a similarity matrix between each candidate target r and existing trajectories $\mathcal{T}_i$, from which we extract the maximum similarity $s_{\max}$ and its corresponding trajectory $\mathbf{t}_{\max}$. If $s \geq \tau_{match}$, we assign r to $\mathbf{t}_{\max}$. If r does not have a matching track, we create a new trajectory for r if its confidence score $p_r \geq \tau_{new}$, otherwise we discard it.

2) Category consistency. As mentioned above, TRACT retains the classification predictions for individual detections during the association process. However, due to the complexity of the OVD task, the classification accuracy achieved by current methods is often suboptimal. Therefore, in TRACT we aim to leverage trajectory information, specific to video-based tasks, to assist this association process. For the i^{th} frame, similar to the approach applied in association, we maintain a category bank $\bar{\mathbf{c}} = \{c_{i-j}\}_{j=1}^{n_{\text{clip}}}$ for each trajectory to store the category predictions c of the previous n_{clip} frames. When a detected object r with category prediction c is successfully matched to a trajectory $\mathbf{t}$, we first consider its classification prediction reliable if its confidence $p_r \geq \tau_{\text{high}}$. If the confidence falls below τ_{high} but remains above τ_{low}, we add it to the corresponding category bank. Lastly, if the confidence $p_r < \tau_{\text{low}}$, the classification prediction is deemed unreliable. In the first case, the final recorded classification prediction $\mathbf{c}$ is set as c, while in the latter two cases, the classification is determined by a voting mechanism. The process can be depicted as follows:

$$\mathbf{c} = \begin{cases} c & \tau_{\text{high}} \leq p_r \\ \mathtt{Vote}(\bar{\mathbf{c}} \cup \{c\}) & \tau_{\text{low}} \leq p_r < \tau_{\text{high}} \\ \mathtt{Vote}(\bar{\mathbf{c}}) & p_r < \tau_{\text{low}} \end{cases}$$

where $\mathtt{Vote}(\cdot)$ stands for the major vote strategy. Note that,

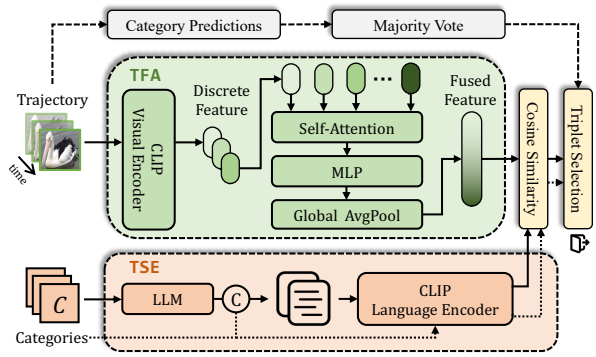


Figure 4. The architecture of the proposed TraCLIP. It approaches both language and visual aspects, making full use of trajectory information to assist classification.

the retained classification prediction $\mathbf{c}$ is not only used for subsequent trajectory classification but also for updating the corresponding category bank.

In summary, during this trajectory-enhanced association step, TRACT predicts object trajectories within a given video and uses TCR to enhance the consistency of identification and classification throughout the association process.

3.4. Trajectory-Assisted Classification

Furthermore, we propose a plug-and-play trajectory classification approach, termed **TraCLIP**, as illustrated in Fig. 4. Specifically, TraCLIP takes N trajectories $\mathcal{T}$ and vanilla vocabulary $\mathcal{V}$ (category names) as input, processes both visual and language information to obtain visual trajectory features and category language features, and then matches them to produce the final trajectory classifications. We address three perspectives of its design:

1) Visual process. To leverage the features provided by trajectories under varying occlusion and blur conditions, we introduce *Trajectory Feature Aggregation* (**TFA**) strategy. Concretely, given a input trajectory $\mathbf{t} \in \mathcal{T}$, we first sample them based on the detection confidence, obtaining a sample clip $\dot{\mathbf{t}}$ of length n_{clip}. If the trajectory length is already less than n_{clip}, no sampling is performed. We then use the CLIP visual encoder to extract its 2D feature $\dot{\mathbf{f}} \in \mathbb{R}^{\text{n} \times d}$ frame by frame, where n is the length of the sample and feature dimension $d = 768$. We consider $\dot{\mathbf{f}}$ as a sequential data, and use self-attention and MLP to get self-enhanced feature $\tilde{\mathbf{f}}$:

$$\ddot{\mathbf{f}} = \dot{\mathbf{f}} + \texttt{SA}(\texttt{LN}(\dot{\mathbf{f}})) \tag{3}$$

$$\tilde{\mathbf{f}} = \ddot{\mathbf{f}} + \texttt{MLP}(\texttt{LN}(\ddot{\mathbf{f}})) \tag{4}$$

where $\texttt{SA}(\mathbf{x})$ denotes self-attention with $\mathbf{x}$ generating query, key, and value as in [19], $\texttt{MLP}(\cdot)$ denotes the multi-layer perception, and $\texttt{LN}(\cdot)$ is a layer normalization function. Finally, we generate the fused trajectory feature by global average pooling $\mathbf{f}^{traj} = \{\texttt{AvgPool}(\tilde{\mathbf{f}}_i)\}_{i=1}^{\text{n}}$. We have explored additional fusion methods, please kindly refer to the **supplementary material** due to limited space.

2) Language process. Since trajectories provide richer feature information, *e.g.*, target characteristics from different perspectives and lighting conditions, relying solely on category names often results in incomplete language features. Therefore, to fully utilize trajectory information, we introduce *Trajectory Semantic Enrichment* (**TSE**) strategy to enhance semantics using attribute information. Given the input vanilla vocabulary $\mathcal{V} = \mathcal{C}^{\text{base}} \cup \mathcal{C}^{\text{novel}}$, we use Large Language Models (LLMs) to decouple them into various attribute descriptions (see Fig.4). Specially, to better employ LLMs to enrich category contexts, we carefully design a prompt template to ensure accurate decomposition, *i.e.*, "*Provide a brief description of the* {category} *focusing on two to three visual attributes*". In this work we prompt ChatGPT to generate attribute answers, and then concatenate them with the corresponding category as follows:

$$\mathcal{A} = \texttt{Concat}(\mathcal{V}, \Phi(\mathcal{V})) \tag{5}$$

where $\Phi(\cdot)$ denotes LLM processing operation. Please refer to **supplementary material** for more details. With the enriched category texts $\mathcal{A}$ available, we use the CLIP language encoder to extract two sets of category language features:

$$\mathcal{F}^{attr} = \texttt{Linear}(\texttt{Enc}(\mathcal{A})) \tag{6}$$

$$\mathcal{F}^{cate} = \texttt{Linear}(\texttt{Enc}(\mathcal{V})) \tag{7}$$

where $\texttt{Enc}(\cdot)$ represents the CLIP language encoder, and $\texttt{Linear}(\cdot)$ stands for a linear projection layer. $\mathcal{F}^{attr}$ and $\mathcal{F}^{cate}$ represent the attribute-assisted language feature and the vanilla language feature, respectively.

3) Triplet selection. At this point, we have obtained two sets of language features $\mathcal{F}^{cate}$, $\mathcal{F}^{attr}$ and a set of visual features $\mathcal{F}^{traj} = \{\mathbf{f}_i^{traj}\}_{i=1}^{n}$ representing each trajectory, where n denotes the length of the trajectory. Together with the classification predictions $\{c_i\}_{i=1}^{n}$ retained in the association step, for each trajectory $\mathbf{t}$ we obtain three classification results along with the corresponding similarity scores. Specifically, we first compute the affinity between its visual features and two types of language features, as follows:

$$\mathcal{Z}(\mathbf{t}) = [\texttt{Cos}(\mathbf{f_t}, \mathcal{F}_1^*), \texttt{Cos}(\mathbf{f_t}, \mathcal{F}_2^*), \cdots, \texttt{Cos}(\mathbf{f_t}, \mathcal{F}_{|\mathcal{V}|}^*)] \tag{8}$$

where $\texttt{Cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$ represents the cosine similarity, $\mathbf{f_t} \in \mathcal{F}^{traj}$ is the visual feature of $\mathbf{t}$, and $\mathcal{F}_i^*$ denotes the i^{th} language feature in $\mathcal{F}^{cate}$ or $\mathcal{F}^{attr}$. We select the classification with the highest similarity score, yielding two classification results $\mathbf{v}_{\text{cate}}$ and $\mathbf{v}_{\text{attr}}$ along with their similarity scores $\mathbf{s}_{\text{cate}}$ and $\mathbf{s}_{\text{attr}}$. Furthermore, we apply a majority vote strategy to obtain the third classification result, represented as $\mathbf{v}_{\text{det}} = \texttt{Vote}(c_1, c_2, \ldots, c_{|\mathcal{V}|})$, and use its proportion as the similarity score $\mathbf{s}_{\text{det}}$. Finally, the result with the *highest* similarity score is selected as the final output.

Table 1. Comparison with state-of-the-art trackers on OV-TAO dataset. The experiments are grouped based on different detectors, which we consider to be more reasonable. The best and second best results within each detection setting are highlighted in **bold** and underline.

Detector	Method		Base				Novel			
Validation Set		**Publication**	**TETA↑**	**LocA↑**	**AssA↑**	**ClsA↑**	**TETA↑**	**LocA↑**	**AssA↑**	**ClsA↑**
ViLD [8]	DeepSORT [21]	ICIP 2017	26.9	47.1	15.8	17.1	21.1	46.4	14.7	2.3
	Tracktor++ [1]	ICCV 2019	28.3	47.4	20.5	17.0	22.7	46.7	19.3	2.2
	OVTrack [12]	CVPR 2023	35.5	49.3	36.9	20.2	27.8	48.8	33.6	1.5
	MASA [13]	CVPR 2024	37.5	**55.2**	37.9	19.3	30.3	**52.8**	35.9	2.3
	TRACT	Ours	**38.5**	55.0	**39.0**	**21.5**	**31.3**	52.7	**37.8**	**3.4**
RegionCLIP [30]	DeepSORT [21]	ICIP 2017	28.4	52.5	15.6	17.0	24.5	49.2	15.3	9.0
	Tracktor++ [1]	ICCV 2019	29.6	52.4	19.6	16.9	25.7	50.1	18.9	8.1
	ByteTrack [28]	ICCV 2019	29.4	52.3	19.8	16.0	26.5	50.8	20.9	8.0
	OVTrack [12]	CVPR 2023	36.3	53.9	36.3	18.7	32.0	51.4	33.2	11.4
	MASA [13]	CVPR 2024	36.7	**54.4**	38.5	17.3	33.6	53.7	35.3	11.8
	TRACT	Ours	**37.9**	54.2	**39.4**	**20.2**	**34.4**	**54.0**	**36.0**	**13.3**
YOLO-World [4]	DeepSORT [21]	ICIP 2017	27.3	47.1	16.5	17.9	21.5	48.9	14.9	3.8
	ByteTrack [28]	ECCV 2022	28.5	46.8	19.2	17.1	22.9	50.1	19.7	3.3
	OC-SORT [3]	CVPR 2023	31.2	51.0	18.8	16.9	24.4	53.3	20.3	3.7
	MASA [13]	CVPR 2024	38.2	**54.9**	**41.0**	18.6	32.2	55.2	37.9	4.4
	TRACT	Ours	**39.4**	**54.9**	40.6	**22.6**	**33.7**	**56.0**	**39.8**	**5.3**
Test Set		**Publication**	**TETA↑**	**LocA↑**	**AssA↑**	**ClsA↑**	**TETA↑**	**LocA↑**	**AssA↑**	**ClsA↑**
ViLD [8]	DeepSORT [21]	ICIP 2017	24.5	43.8	14.6	15.2	17.2	38.4	11.6	1.7
	Tracktor++ [1]	ICCV 2019	26.0	44.1	19.0	14.8	18.0	39.0	13.4	1.7
	OVTrack [12]	CVPR 2023	32.6	45.6	35.4	16.9	24.1	41.8	28.7	1.8
	MASA [13]	CVPR 2024	35.2	**52.5**	37.9	15.3	26.6	47.9	30.6	1.3
	TRACT	Ours	**36.2**	52.3	**39.1**	**17.2**	**27.3**	**48.2**	**30.7**	**3.1**
RegionCLIP [30]	DeepSORT [21]	ICIP 2017	27.0	49.8	15.1	16.1	18.7	41.8	9.1	5.2
	Tracktor++ [1]	ICCV 2019	28.0	49.4	18.8	15.7	20.0	42.4	12.0	5.7
	ByteTrack [28]	ECCV 2022	28.7	51.5	19.9	14.5	20.4	43.0	13.5	4.9
	OVTrack [12]	CVPR 2023	34.8	51.1	36.1	17.3	25.7	44.8	26.2	6.1
	MASA [13]	CVPR 2024	36.5	**53.2**	39.0	17.3	26.8	44.8	29.5	6.2
	TRACT	Ours	**37.3**	53.0	**39.4**	**19.3**	**28.8**	**45.3**	**30.1**	**10.8**
YOLO-World [4]	DeepSORT [21]	ICIP 2017	25.1	43.3	15.6	13.0	16.9	40.5	11.8	8.8
	ByteTrack [28]	ICCV 2019	26.6	44.1	19.3	11.7	18.4	41.3	15.1	5.0
	OC-SORT [3]	CVPR 2023	28.9	49.0	19.1	9.9	20.6	48.3	14.8	5.8
	MASA [13]	CVPR 2024	34.9	**51.8**	39.7	13.2	32.2	51.4	**36.2**	9.2
	TRACT	Ours	**36.1**	51.6	**40.7**	**15.9**	**33.3**	**51.8**	35.9	**12.0**

3.5. Training Strategy

In the trajectory-enhanced association step, both of our proposed trajectory banks are training-free, so rather than designing a specific training method, we turn to the general training approach of appearance-based matching models. In specific, we adopt the training approach from [13] and employ a contrastive learning method.

For the trajectory-assisted classification, we initialize TraCLIP with CLIP [17] weights using ViT-L/14 as the backbone, freezing both the language and visual encoders during training. We adopt CLIP's contrastive loss and use LVIS [9], YouTube-VIS [23], and TAO [5] training set as training data. Specifically, target trajectories and category names from these datasets serve as input and labels. Since LVIS is an image dataset, we generate trajectory data for each target with n_{clip} augmentations, such as random rotation, erasure, and scaling. Note that, during the entire training process, we only used *known* object categories. Please refer to the **supplementary material** for more details.

4. Experiments

4.1. Experimental Setup

Benchmark. We conduct experiments on the large-scale open-vocabulary dataset OV-TAO, extended from TAO [5], which includes 2,907 sequences and over 800 categories. OV-TAO follows the classification scheme inLVIS [9] by dividing categories into *base* (common) and *novel* (rare) classes. This setup mirrors the real-world scenarios and can reflect the adaptability trackers in handling rare categories.

Metrics. Following [12–14], we use Tracking-Every-Thing Accuracy (TETA) metric [11], which disentangles MOT evaluation into three subfactors: Localization Accuracy (LocA), Association Accuracy (AssocA) and Classification Accuracy (ClsA). Please note, all our experiments are conducted under the open-vocabulary setting.

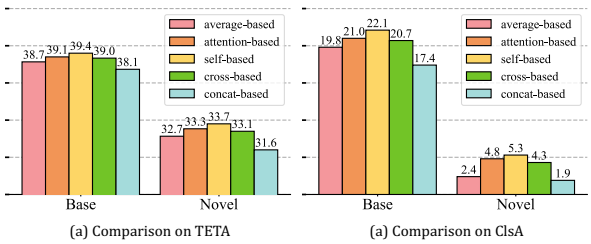


Figure 5. Comparison of different fusion mechanisms on the validation set of OV-TAO [12], using TETA (a) and ClsA (b) metrics.

Table 2. Ablation studies to evaluate the contribution of the proposed strategies in TRACT. The best results are in **bold**.

TCR	TFA	TSE	TETA	LocA	AssA	ClsA	Speed(*s/seq*) ↓
			37.5	**55.0**	40.1	16.9	5.81
✓			37.6	**55.0**	**40.6**	17.3	6.04
✓	✓		38.5	54.9	40.5	19.9	6.64
	✓		38.4	54.9	40.4	19.8	6.57
	✓	✓	38.4	54.9	40.4	19.7	6.71
✓	✓	✓	**38.6**	54.9	**40.6**	**20.3**	6.99

Table 3. Comparison experiments of TSE effect.

	ClsA (*base*)	ClsA (*novel*)		ClsA (*base*)	ClsA (*novel*)
with TSE	20.2	13.3	w/o TSE	21.6	10.9

Implementation details. We conduct experiments with 4 Nvidia Tesla V100 GPUs. We set the batch size to 256 per GPU and use the AdamW optimizer to train the model. Please note again that *only* $\mathcal{C}^{\text{base}}$ categories are used in training. The initial learning rate is set to 1×10^{-4} , and the weight decay is set to 1×10^{-5}. During inference, we set $n_{\text{bank}} = 15$ in the association step and $n_{\text{clip}} = 5$ in the classification step. See **supplementary material** for details.

4.2. Comparison to State-of-the-Art

We conduct experiments on both validation and test sets of TAO. Please note, considering the strong correlation between current OV-MOT and OVD methods, we group the experimental results based on the different OVD models used to ensure fairness in comparison. Concretely, we first use two typically used [12, 13] detector ViLD [8] and RegionCLIP [30], and then explore a state-of-the-art OVD method, YOLO-World [4]. For ViLD and RegionCLIP, we use the same detection results as in OVTrack [12], while for YOLO-World, we utilize the officially provided weights. Given the limited data volume and incomplete annotations [11] in the TAO training set, we refrain from further fine-tuning. Throughout the comparison, we focus primarily on comparisons within each group.

As shown in Tab.1, TRACT consistently achieves top-tier results on nearly all metrics, demonstrating strong results with various detectors. For instance, when using the state-of-the-art open-vocabulary detector YOLO-World, it achieves TETA scores of 39.4% and 33.7% for base and novel classes, respectively, on the validation set, and 35.7% and 33.1% on the test set. Notably, TRACT demonstrates significant improvements on the ClsA metric. Compared to the current leading tracker MASA[13], TRACT shows gains of +2.0% and +4.6% (base/novel classes) on the test set with RegionCLIP and +1.9% and +1.5% on the validation set. These results indicate that incorporating trajectory information in OV-MOT is both beneficial and promising.

Table 4. Ablation studies of n_{bank} on the validation set of OV-TAO . Please note, here we do not use TraCLIP to ensure a clear comparison. The best results are highlighted in **bold**. We use the average time per sequence to measure the model speed.

n_{bank}	TETA	LocA	AssA	ClsA	Speed(*s/seq*) ↓
1	37.51	55.02	41.08	16.91	5.81
5	37.56	**55.04**	41.12	16.52	6.04
10	37.54	55.01	40.64	16.96	6.24
15	**37.62**	**55.04**	40.58	**17.27**	6.36
20	37.51	55.03	40.53	16.97	6.56
25	37.61	55.01	**40.74**	17.09	6.80

Due to limited space, we provide the visualization results in the **supplementary material** to show the effectiveness of TRACT and its superiority in handling object occlusion.

4.3. Analysis on TraCLIP

In this paper, we introduce TraCLIP as not only a trajectory-based classification approach but also a promising new direction for classification research. In this section, we conduct a series of experiments on TraCLIP and provide an in-depth analysis of its strengths, weaknesses, and limitations.

Analysis on feature fusion. In TraCLIP, we introduce the TFA strategy to integrate trajectory visual features into classification. Although similar to video retrieval, our focus is on utilizing complementary information from different perspectives and appearances across trajectories, rather than emphasizing temporal information. In the TFA strategy, the feature fusion module is a key component that generates enhanced trajectory features. In this work, we study five types of feature fusion mechanisms, *i.e.*, average-based fusion, attention-based fusion (using self-attention module), self-based fusion (using self-attention and mlp modules), cross-based fusion (using cross-attention), and a concatenation-based fusion (using concatenation between visual and language features). Please refer to the **supplementary material** for detailed architectures. Fig. 5 shows the results of different fusion mechanisms on the TETA and ClsA metrics. We can see that the second self-based fusion works generally better by achieving the best TETA score (39.4% / 33.7% for base and novel classes) and ClsA score (22.1% / 5.3% for base and novel classes). Therefore, in TRACT we employ the self-based fusion mechanism.

Analysis on running speed. In model design, speed is

Table 5. Ablation studies of n_{clip}. In this study, we only evaluate the running speed of the TraCLIP module.

n_{clip}	TETA	LocA	AssA	ClsA	Speed(*s/seq*) ↓
1	37.96	53.89	40.50	19.09	0.77
5	38.59	54.90	40.51	20.30	1.28
10	38.48	54.90	40.51	20.04	2.55
15	38.51	54.90	40.51	20.13	4.97

Table 6. Ablation studies for the weighting parameter α.

α	TETA	LocA	AssA	ClsA
0.1	37.41	54.96	40.39	**17.60**
0.2	37.49	54.98	40.17	17.56
0.25	**37.62**	**55.04**	**40.58**	17.27
0.3	37.50	55.02	40.45	16.89
0.4	37.17	54.89	39.56	16.12

crucial as it directly affects responsiveness and user experience in real-time applications. In this work, while the additional module designs inevitably introduce some reduction in speed, we believe, as shown in Tab.4 and Tab.5, that TRACT maintains a sufficiently fast rate and achieves a strong balance between efficiency and performance.

4.4. Ablation Study

To further analyze TRACT, we conduct ablations on the validation set of OV-TAO with YOLO-World as the detector.

Ablation on three key strategies. In this paper, we propose three key trajectory-based strategies, *i.e.*, TCR, TFA, and TSE strategies. To assess the impact for them, we compare the performance on the validation set of OV-TAO [12] using the state-of-the-art open-vocabulary detector YOLO-World [4]. As depicted in Tab. 2, we can see that the version incorporating all three strategies achieves the best performance across almost all metrics, especially with a notable +3.4% improvement in the ClsA metric. Besides, as shown in Tab. 3, although the TSE module has a limited impact on overall classification performance, it improves the classification of novel classes, which is a key goal in OV-MOT. Refer tor **supplementary material** for visualizations. Please note that TRACT does not involve adjustments in localization, so the LocA metric shows no significant change.

Ablation on lengths n_{bank} and n_{clip}. To investigate the impact of two key length parameters n_{bank} and n_{clip} of TRACT, we conduct experiments with varying parameter settings. n_{bank} is the maximum length of the feature banks and category banks used in TCR, while n_{clip} is the sample clip length of TraCLIP. From Tab. 4, we can see that, when $n_{\text{bank}} = 15$, the overall best performance is achieved. Please note, in the ablation study of n_{bank}, we exclusively apply the TCR strategy to ensure a fair comparison. We measure the model speed by the average processing time per sequence (*s/seq)*, finding that increasing n_{bank} does not result in a notable increase in time costs (see Tab. 4). Furthermore, as shown in Tab. 5, we do not observe a significant improvement in effectiveness as n_{clip} increases, instead, there is a noticeable decrease in processing speed (see Tab. 5). Therefore, in TRACT we use $n_{\text{clip}} = 5$.

Ablation on weighting parameter α. We propose the TCR strategy, where we use the weighting parameter α to balance the use of the track memory and feature bank. Please note, in this experiment, only TCR strategy is applied. As shown in Tab. 6, we can observe that when $\alpha = 0.25$, the model achieves the overall best performance.

4.5. Discussion

Challenge in OV-MOT. Current OV-MOT faces severe challenges in association due to dense detection results. We find that OV-MOT task, typically evaluated with the TETA metric[11], has a much higher detection density than conventional MOT, which uses the HOTA metric [15]. Please kindly refer to the **supplementary material** for visualization of this situation. This density arises from incomplete annotations in the TAO dataset [5], which covers over 800 categories but contains many *missing* labels. The TETA metric mitigates this by not penalizing unmatchable predictions, but this reduces penalties for false positives, prompting detectors to lower thresholds to capture rare categories. This results in dense, low-quality detections, complicating association further. We argue that the primary issues of current OV-MOT lie in data and evaluation protocol, and hope the community to address these foundational challenges.

Can trajectory improves localization? This paper primarily investigates using trajectory information to enhance *association* and *classification*, but we believe it can also aid in *localization*. In OVD, localizing unknown or rare classes is challenging. However, in OV-MOT, once a target is detected, its appearance can improve localization in subsequent frames. Though preliminary experimental results following MOTRv2 [29] show limited improvement, we believe it is a potential area for future research.

5. Conclusion

In this work, we explore trajectory-level information to improve OV-MOT by enhancing association and classification steps. Our method, TRACT, utilizes trajectory and temporal information to enhance performance compared to instance-level approaches. We introduce the TCR strategy to improve identity and category consistency in trajectory-enhanced association and propose TraCLIP, which employs TFA and TSE strategies for trajectory-assisted classification from visual and language perspectives. Our extensive experiments show that TRACT significantly enhances tracking performance, highlighting the importance of trajectory information in open-vocabulary contexts.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, pages 941–951, 2019. 6

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468. IEEE, 2016. 3

[3] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *CVPR*, pages 9686–9696, 2023. 6

[4] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *CVPR*, pages 16901–16911, 2024. 6, 7, 8, 2

[5] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, pages 436–454. Springer, 2020. 3, 6, 8

[6] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *TMM*, 25:8725–8737, 2023. 3

[7] Ruopeng Gao and Limin Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *ICCV*, pages 9901–9910, 2023. 3

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv*, 2021. 3, 6, 7, 2

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 6

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 3, 2

[11] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *ECCV*, pages 498–515. Springer, 2022. 6, 7, 8, 3

[12] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *CVPR*, pages 5567–5577, 2023. 1, 2, 3, 4, 6, 7, 8

[13] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *CVPR*, pages 18963–18973, 2024. 3, 4, 6, 7, 2

[14] Siyuan Li, Lei Ke, Yung-Hsu Yang, Luigi Piccinelli, Mattia Segù, Martin Danelljan, and Luc Van Gool. Slack: Semantic, location, and appearance aware open-vocabulary tracking. *arXiv*, 2024. 2, 3, 6

[15] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 129:548–578, 2021. 8

[16] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3025–3029. IEEE, 2023. 3

[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 *ICML*, pages 8748–8763. PMLR, 2021. 2, 3, 6

[18] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 3

[19] A Vaswani. Attention is all you need. *NIPS*, 2017. 3, 5

[20] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *ECCV*, pages 107–122. Springer, 2020. 3

[21] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649. IEEE, 2017. 3, 6

[22] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *European Conference on Computer Vision*, pages 733–751. Springer, 2022. 3

[23] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 6

[24] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, pages 106–122. Springer, 2022. 3

[25] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 3

[26] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *ECCV*, pages 659–675. Springer, 2022. 3

[27] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129:3069–3087, 2021. 3

[28] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21. Springer, 2022. 3, 6

[29] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *CVPR*, pages 22056–22065, 2023. 3, 8

[30] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 3, 6, 7, 2

Attention to Trajectory: Trajectory-Aware Open-Vocabulary Tracking

Supplementary Material

In this supplementary material, we provide further elaboration on our proposed method **TRACT**. Specially, **S1** delineates the specific structures of the five feature fusion mechanisms in the proposed **TraCLIP**. Within **S2**, we present details of the loss function deployed for the training of TRACT. In **S3** we provide details of the parameter settings used in our experiments. In **S4**, we analyze the effectiveness of TSE using visualization. **S5** delineates the visualization and detailed analysis of the detection density problem. We demonstrates the visualization results of TRACT in **S6**. Lastly, **S7** details the meticulously crafted 1203 category descriptions we used.

S1 Feature Fusion Mechanisms

In Sec. 4.3, we compare five different feature fusion mechanisms, as shown in Fig. 6. They all take N features $\{f\}_{i=1}^{N}$ belonging to the same trajectory sample as input, and adopt different feature fusion mechanisms to integrate discrete features.

Average-based fusion. The simplest way is to average the features (see Fig. 6 (a)), which has the advantages of being parameter-free and highly efficient. The fusion process is formulated as follows:

$$\breve{f} = \texttt{Avg}(\{f\}_{i=1}^{N}) \tag{1}$$

where $\breve{f}$ represents the output trajectory feature, and $\texttt{Avg}(\cdot)$ represents the average function along the sequence dimension.

Attention-based fusion. As depicted in Fig. 6 (b), the attention-based fusion strategy employs a self-attention module to achieve information interaction within the trajectory. In specific, it takes the input features $\{f\}_{i=1}^{N}$ as a sequential data, and fuse them as follows:

$$\breve{f} = \texttt{Avg}(\texttt{SA}(\{f\}_{i=1}^{N})) \tag{2}$$

where $\texttt{SA}(\cdot)$ denotes self-attention module.

Self-based fusion. The self-based fusion mechanism, which we employ in the TFA strategy, is similar to the attention-based one (see Fig. 6 (c)). In addition to the self-attention module, it also utilizes an MLP for further information interaction, as follows:

$$\breve{f} = \texttt{Avg}(\texttt{MLP}(\texttt{SA}(\{f\}_{i=1}^{N}))) \tag{3}$$

where $\texttt{MLP}(\cdot)$ is the multi-layer Perceptron module. Experimental results show that this approach achieves the best performance.

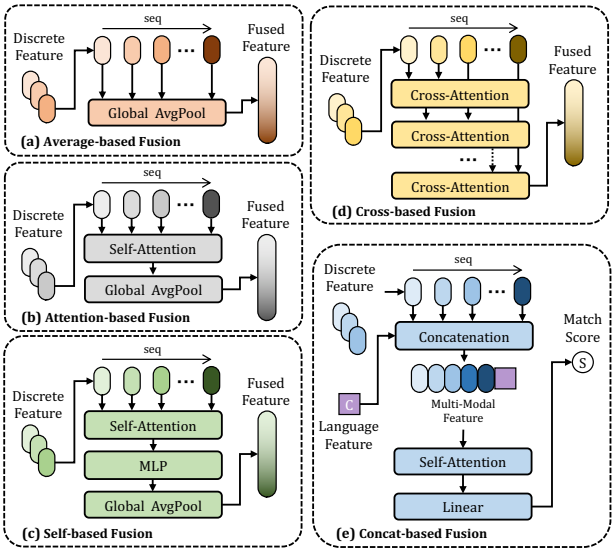


Figure 6. Detailed architectures of the studied 5 fusion mechanisms, *i.e.*, average-based fusion (a), attention-based fusion (b), self-based fusion (c), cross-based fusion (d), and concatenation-based fusion (e).

Cross-based fusion. The fundamental concept of cross-based fusion involves sequentially feed N features as K/V into a cross-attention module, with each step's outcome serving as the subsequent fusion's Q (see Fig. 6 (d)). In specific, the fused trajectory feature $\dot{f}_i$ up to the i^{th} target in the trajectory is formulated as follows:

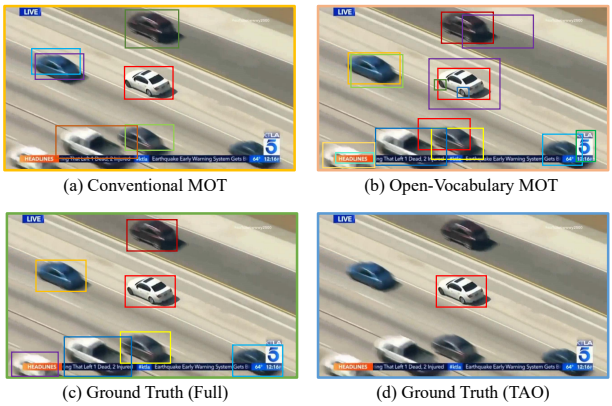
$$\dot{f}_i = \texttt{CA}(\dot{f}_{i-1}, f_i) \qquad i > 1 \tag{4}$$

where $\texttt{CA}(\mathbf{x}, \mathbf{y})$ denotes the cross-attention with $\mathbf{x}$ as query and $\mathbf{y}$ as key and value. For the first feature, we set $\dot{f}_1 = f_1$. Finally, cross-based fusion module outputs the trajectory feature $\breve{f} = \dot{f}_N$.

Concatenation-based fusion. As shown in Fig. 6, the concatenation-based fusion strategy first concatenates N discrete visual features $\{f\}_{i=1}^{N} \in \mathbb{R}^{N \times m}$ with the language feature of a target category $c \in \mathbb{R}^{1 \times m}$ along the sequence dimension, resulting in a unified multi-modal feature tensor $\dot{f} = \texttt{Concat}(f_i, f_2, ..., f_N, c) \in \mathbb{R}^{(N+1) \times m}$. It then uses a self-attention module and two linear projection layers to obtain the similarity score for the corresponding category, as follws:

$$s = \texttt{FC}(\texttt{Pool}(\texttt{SA}(\texttt{Concat}(f_i, f_2, ..., f_N, c)))) \tag{5}$$

where $\texttt{Pool}(\cdot)$ is a projection layer used for internal information interaction, while $\texttt{FC}(\cdot)$ is a projection layer used to



(a) Conventional MOT (b) Open-Vocabulary MOT

(c) Ground Truth (Full) (d) Ground Truth (TAO)

Figure 7. A visual comparison between Conventional MOT detections (a), OV-MOT detections (b), the complete GT annotations (c), and the GT annotations provided by TAO (d). It clearly reveals two points: first, the detection density of OV-MOT is significantly higher than that of Conventional MOT, and second, the GT annotations provided by TAO are incomplete.

obtain the score. Please note, unlike previous fusion strategies, the concatenation-based fusion strategy directly produces classification scores for each category.

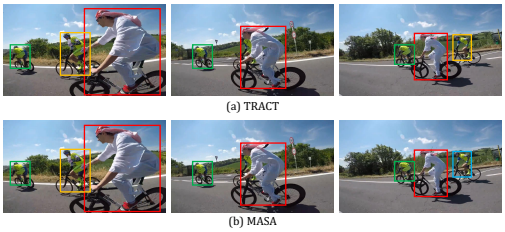
S2 Detailed Training Strategy

The training of TRACT is divided into two main parts: training for the trajectory-enhanced association step and training for the trajectory-assisted classification step. For the former, we use the method proposed in MASA [13], leveraging SAM [10] to enable training on unlabeled image data. For the latter, we train using a contrastive learning loss.

In contrastive learning, given a pair of samples (x_i, x_j), we assign a label y_{ij} to indicate whether the two samples belong to the same class. If $y_{ij} = 1$, then x_i and x_j are similar (a positive sample pair), whereas if $y_{ij} = 0$, they are dissimilar (a negative sample pair). The contrastive loss function is defined as follows:

$$\mathcal{L}(x_i, x_j) = \frac{1}{2}\Big(y_{ij} \cdot D(x_i, x_j)^2 + (1 - y_{ij}) \cdot \max\big(0, m - D(x_i, x_j)\big)^2\Big) \tag{6}$$

where $D(x_i, x_j)$ represents the distance between samples x_i and x_j, which is often measured using Euclidean or cosine distance. The term m is a margin parameter that enforces a minimum distance between dissimilar pairs. For similar pairs ($y_{ij} = 1$), the loss $\frac{1}{2}D(x_i, x_j)^2$ encourages a smaller distance, effectively pulling similar pairs closer together. For dissimilar pairs ($y_{ij} = 0$), the loss $\frac{1}{2}\max(0, m - D(x_i, x_j))^2$ promotes a distance of at least m, pushing these pairs farther apart.



(a) TRACT

(b) MASA

Figure 8. Performance under occlusion condition.

(a) TRACT **with TSE strategy** (b) TRACT **without TSE strategy**

Figure 9. Visualization analysis of the effectiveness of the TSE module.

The objective of contrastive loss is to learn an embedding space in which similar samples are clustered closely, while dissimilar samples are separated by a clear margin, thereby creating a well-defined separation in the embedding space.

S3 Experiment Settings

In Sec. 4.1, we introduced some of the parameter settings used in our experiments. Due to space limitations, we provide a more detailed description here. Specially, during inference, we set the matching threshold $\tau_{\text{match}} = 0.4$. Additionally, there are two critical parameters, τ_{high} and τ_{low}, used in the association process. Since they are closely related to the score range provided by the detection results, we set different values for these parameters depending on the detector used. Specifically, when using ViLD [8] and YOLO-World [4] as the detector, we set $\tau_{\text{high}} = 0.3$ and $\tau_{\text{low}} = 0.1$. While using RegionCLIP [30], since the confidence score consists of two components, values greater than 1 can occur. To adapt this for tracking purposes, we divide the confidence score by 10, which in turn requires adjusting τ_{high} and τ_{low} to 0.01 and 0.005, respectively.

6. S4 Visual Analysis of TSE

In our work, TCR and TFA primarily contribute to the ***overall performance*** of the model. Meanwhile, TSE mainly contributes to ***the classification of novel categories*** (see *Table 3* in paper), which we believe is a critical aspect of open-vocabulary tracking. We provide a visualization-based analysis of TSE (see Fig. 9) and plan to include more comprehensive visualizations in revision to further illustrate their effects. Thanks!

S5 Detection Density Visualization

As analyzed in Sec. 4.5, one reason why OV-MOT is challenging, in addition to the primary classification difficulties, lies in the association difficulty arising from the dense detection results. As illustrated in Fig. 7, the detection den-

sity in OV-MOT (typically evaluated with the TETA metric [11]) is significantly higher than in conventional MOT (evaluated with the HOTA metric). We attribute this to incomplete annotation in the TAO dataset, which includes over 800 categories but lacks exhaustive annotations for most of them. This issue intensifies in the OV-MOT setting, where the TETA metric employs a strategy that does *not* penalize predictions unmatchable to ground-truth annotations. While this approach mitigates the issue of incomplete annotations, it also lessens penalties for numerous false positives. Consequently, detectors often lower detection thresholds to improve tracking of rare categories, resulting in dense and potentially low-quality detection outputs that complicate association further. In summary, we argue that the main challenges in OV-MOT stem from data and evaluation conventions. We hope the OV-MOT community will prioritize addressing these fundamental issues over refining surface-level solutions.

S6 Tracking Results Visualization

A key advantage of using trajectory information is the improved tracking performance under occlusion conditions. We demonstrate the visualization of this scenario in Fig. 8. We can observe that, compared to MASA, TRACT better maintains the stability of target ids after detection loss.

S7 Descriptions with Attributes

OV-TAO use the 1203 categories defined in LVIS dataset, including 866 base classes and 337 novel classes. In TraCLIP, we propose the Trajectory Semantic Enrichment (**TSE**) strategy to enhance language features using category descriptions. In order to able understanding of TSE, below we present each category description:

aerosol_can: *cylindrical container with a spray nozzle on top.*
air_conditioner: *boxy unit with vents for cooling air.*
airplane: *streamlined body with wings and tail.*
alarm_clock: *small rectangular device with a face and hands.*
alcohol: *clear or amber liquid in a bottle.*
alligator: *thick, scaly body with a long snout.*
almond: *oval-shaped nut with a brown skin.*
ambulance: *white vehicle with a red cross and siren.*
amplifier: *boxy device with knobs and speaker connections.*
anklet: *decorative band worn around the ankle.*
antenna: *long, thin rod extending from a device.*
apple: *round fruit with a smooth skin.*
applesauce: *smooth, chunky puree in a jar.*
apricot: *small, orange fruit with a fuzzy skin.*
apron: *cloth garment worn over the front of clothes.*
aquarium: *glass tank filled with water and fish.*
arctic_(type_of_shoe): *insulated boot with a thick sole.*
armband: *fabric band worn on the arm, often colorful.*
armchair: *soft chair with armrests and a cushioned seat.*
armoire: *tall, freestanding cabinet with doors.*
armor: *protective metal covering for the body.*
artichoke: *green, spiky vegetable shaped like a flower.*
trash_can: *cylindrical container for waste with a lid.*
ashtray: *shallow dish for holding cigarette ashes.*
asparagus: *long, thin green vegetable with a pointed tip.*
atomizer: *small bottle with a spray nozzle for perfumes.*
avocado: *green, pear-shaped fruit with a rough skin.*
award: *decorative object, often in the shape of a trophy.*
awning: *fabric canopy extending over a doorway.*
ax: *heavy tool with a sharp blade for chopping.*
baboon: *large primate with a long face and colorful rear.*
baby_buggy: *wheeled carriage for transporting infants.*
basketball_backboard: *large, flat surface with a hoop attached.*
backpack: *fabric bag with straps worn on the back.*
handbag: *small purse carried by hand or on the shoulder.*
suitcase: *hard or soft case for carrying clothes.*
bagel: *round bread with a hole in the center.*
bagpipe: *musical instrument with pipes and a bag.*
baguet: *long, thin loaf of french bread.*
bait: *lure, often in the form of small fish or worms.*
ball: *round object used in various games.*
ballet_skirt: *flowy, short skirt often made of tulle.*
balloon: *inflatable round object made of rubber or plastic.*
bamboo: *tall, slender grass with hollow stems.*
banana: *curved, yellow fruit with a soft interior.*
band_aid: *small adhesive strip with a sterile pad.*
bandage: *cloth strip used to cover wounds.*
bandanna: *square piece of cloth tied around the head or neck.*
banjo: *string instrument with a circular body and neck.*
banner: *large piece of cloth with words or images displayed.*
barbell: *long metal rod with weights on both ends.*
barge: *flat-bottomed boat used for transporting goods.*
barrel: *round container used for storage, often made of wood.*
barrette: *hair clip used to hold hair in place.*
barrow: *wheelbarrow with a large, open bed.*
baseball_base: *square rubber pad used in baseball fields.*
baseball: *round ball with a leather cover and stitching.*
baseball_bat: *long, cylindrical piece of wood or metal.*
baseball_cap: *soft cap with a curved brim.*
baseball_glove: *leather glove designed for catching baseballs.*
basket: *woven container used for carrying items.*
basketball: *round ball with a textured surface.*
bass_horn: *large brass instrument with a wide bell.*
bat_(animal): *winged mammal with a fur-covered body.*

bath_mat: *soft, absorbent mat placed outside a bathtub.*
bath_towel: *large, absorbent cloth for drying off.*
bathrobe: *loose-fitting garment worn after bathing.*
bathtub: *enclosed basin for bathing, often with a faucet.*
batter_(food): *thick mixture used for baking or frying.*
battery: *rectangular device that stores electrical energy.*
beachball: *large, inflatable ball for playing at the beach.*
bead: *small, round object often used in jewelry making.*
bean_curd: *soft, pale food made from soybeans.*
beanbag: *soft, cushioned bag filled with beans or foam.*
beanie: *close-fitting, knitted hat.*
bear: *large, furry mammal with a broad body.*
bed: *rectangular piece of furniture for sleeping.*
bedpan: *shallow container for bodily waste.*
bedspread: *decorative covering for a bed.*
cow: *large farm animal with a sturdy body and udder.*
beef_(food): *red meat cut from a cow.*
beeper: *small device that emits sound or alerts.*
beer_bottle: *tall glass container for beer.*
beer_can: *metal container for beer, often cylindrical.*
beetle: *small, hard-bodied insect with antennae.*
bell: *round object that makes a ringing sound.*
bell_pepper: *colorful vegetable with a bell shape.*
belt: *flexible band worn around the waist.*
belt_buckle: *decorative fastener for a belt.*
bench: *long seat with no backrest.*
beret: *soft, round hat often worn tilted.*
bib: *cloth or plastic garment worn to protect clothing.*
bible: *bound book with a leather or cloth cover.*
bicycle: *two-wheeled vehicle powered by pedaling.*
visor: *brimmed headwear that shields the eyes.*
billboard: *large outdoor advertising structure.*
binder: *cover with rings for holding loose papers.*
binoculars: *two-tube device for magnifying distant objects.*
bird: *small, feathered animal with wings.*
birdfeeder: *container designed to hold birdseed.*
birdbath: *shallow basin for birds to bathe in.*
birdcage: *enclosed space with bars for housing birds.*
birdhouse: *small wooden structure for nesting birds.*
birthday_cake: *decorated cake typically with candles on top.*
birthday_card: *greeting card designed for birthday wishes.*
pirate_flag: *black flag with a skull and crossbones design.*
black_sheep: *sheep with black wool, often symbolizing uniqueness.*
blackberry: *dark purple fruit with small, edible seeds.*
blackboard: *dark surface for writing with chalk.*
blanket: *soft covering for warmth, often made of fabric.*
blazer: *tailored jacket, typically with a single-breasted front.*
blender: *kitchen appliance with a jar and blades for mixing.*
blimp: *large, inflatable airship with a rounded shape.*
blinker: *light signal on a vehicle indicating turns.*
blouse: *women's garment that is loose-fitting and stylish.*
blueberry: *small, round fruit that is blue or purple.*
gameboard: *flat surface marked for playing games.*
boat: *watercraft designed for navigation on water.*
bob: *short haircut that is typically straight around the head.*
bobbin: *small spool for holding thread or yarn.*
bobby_pin: *hairpin used to hold hair in place.*
boiled_egg: *egg that has been cooked in boiling water.*
bolo_tie: *neckwear consisting of a cord with decorative tips.*
deadbolt: *locking mechanism with a solid metal bolt.*
bolt: *metal fastener used for joining objects together.*
bonnet: *soft hat that ties under the chin, often worn by women.*
book: *bound collection of written or printed pages.*
bookcase: *shelved furniture for storing books.*
booklet: *small book with a few pages, often informative.*
bookmark: *strip of paper or fabric used to mark a page.*
boom_microphone: *long microphone mounted on a boom pole.*
boot: *sturdy footwear that covers the ankle.*
bottle: *container, typically made of glass or plastic, for liquids.*
bottle_opener: *tool for removing caps from bottles.*
bouquet: *arrangement of flowers, often for gifting.*
bow_(weapon): *curved weapon used for shooting arrows.*
bow_(decorative_ribbons): *fabric ribbon tied in a loop for decoration.*
bow-tie: *necktie shaped like a bow, often worn with formal wear.*
bowl: *round dish used for serving food.*
pipe_bowl: *part of a pipe where tobacco is placed.*
bowler_hat: *rounded, hard-brimmed hat often worn by men.*
bowling_ball: *heavy, spherical ball used in bowling.*
box: *container with a flat base and sides.*
boxing_glove: *padded glove worn in boxing matches.*
suspenders: *straps worn over the shoulders to hold up trousers.*
bracelet: *decorative band worn around the wrist.*
brass_plaque: *flat, engraved metal plate for display.*
brassiere: *garment worn to support women's breasts.*
bread-bin: *container for storing bread to keep it fresh.*
bread: *staple food made from flour and water, often baked.*
breechcloth: *simple garment worn around the waist.*
bridal_gown: *formal dress worn by a bride during a wedding.*
briefcase: *portable case for carrying documents or laptops.*
broccoli: *green vegetable with a tree-like structure.*
broach: *decorative pin worn on clothing.*
broom: *tool with bristles for sweeping floors.*

brownie: *small, chocolate dessert, often fudgy.*
brussels_sprouts: *small, round green vegetable resembling cabbage.*
bubble_gum: *chewy candy that can be blown into bubbles.*
bucket: *container with a handle, typically for carrying liquids.*
horse_buggy: *small, lightweight carriage pulled by a horse.*
bull: *large, male bovine animal with a sturdy body.*
bulldog: *stocky dog breed with loose skin and a flat face.*
bulldozer: *heavy machinery with a broad blade for pushing earth.*
bullet_train: *high-speed train designed for rapid travel.*
bulletin_board: *board for posting notices or announcements.*
bulletproof_vest: *protective garment designed to resist bullets.*
bullhorn: *loudspeaker used to amplify the voice.*
bun: *round bread roll or a hairstyle with hair pulled back.*
bunk_bed: *bed with one bed stacked on top of another.*
buoy: *floating marker used to indicate water depths or hazards.*
burrito: *rolled tortilla filled with various ingredients.*
bus_(vehicle): *large vehicle designed for transporting passengers.*
business_card: *small card with contact information for professionals.*
butter: *yellow dairy product used for cooking or spreading.*
butterfly: *colorful insect with large, patterned wings.*
button: *small disk used for fastening clothing.*
cab_(taxi): *vehicle for hire, typically painted in bright colors.*
cabana: *sheltered area often found by a pool or beach.*
cabin_car: *passenger car used in trains for seating.*
cabinet: *storage unit with shelves and doors, often for kitchen use.*
locker: *enclosed space with a door for storing personal items.*
cake: *sweet baked dessert, often layered and decorated.*
calculator: *electronic device for performing mathematical calculations.*
calendar: *chart displaying days, weeks, and months.*
calf: *young bovine animal.*
camcorder: *portable device for recording video.*
camel: *long-necked animal with a hump, often found in deserts.*
camera: *device for capturing images or video.*
camera_lens: *optical component used to focus light in cameras.*
camper_(vehicle): *vehicle designed for living and camping.*
can: *metal container for storing food or beverages.*
can_opener: *tool for opening metal cans.*
candle: *wax stick with a wick that burns for light.*
candle_holder: *stand or holder for holding a candle upright.*
candy_bar: *chocolate-covered confection filled with sweet ingredients.*
candy_cane: *hard candy shaped like a cane, often striped.*
walking_cane: *stick used for support while walking.*
canister: *container, typically with a lid, for storing items.*
canoe: *narrow boat propelled by paddling.*
cantaloup: *round, orange-fleshed melon with a netted skin.*
canteen: *container for carrying water or other liquids.*
cap_(headwear): *soft hat that fits snugly on the head.*
bottle_cap: *circular cover for sealing a bottle.*
cape: *flowing garment worn over the shoulders.*
cappuccino: *coffee drink made with steamed milk and foam.*
car_(automobile): *motor vehicle with four wheels for transportation.*
railcar_(part_of_a_train): *car designed for traveling on railway tracks.*
elevator_car: *compartment that moves within a building to transport people.*
car_battery: *rechargeable battery used to power a vehicle.*
identity_card: *official document that verifies a person's identity.*
card: *rectangular piece of paper or plastic used for various purposes.*
cardigan: *knitted sweater with an open front and buttons.*
cargo_ship: *large vessel designed for transporting goods.*
carnation: *flower with frilled petals, often used in bouquets.*
horse_carriage: *wheeled vehicle pulled by horses for transportation.*
carrot: *long, orange root vegetable with a crunchy texture.*
tote_bag: *large bag with handles for carrying items.*
cart: *small wheeled vehicle for transporting goods.*
carton: *container made of cardboard for packaging items.*
cash_register: *machine for recording sales transactions.*
casserole: *dish used for cooking food in the oven.*
cassette: *plastic case containing magnetic tape for audio recording.*
cast: *solid form made to support a broken limb.*
cat: *small domesticated feline known for its agility.*
cauliflower: *white, dense flower vegetable, often used in cooking.*
cayenne_(spice): *hot, red pepper used as a spice.*
cd_player: *device for playing compact discs.*
celery: *crisp, green stalk vegetable, often used in salads.*
cellular_telephone: *portable device for communication over cellular networks.*
chain_mail: *armor made of interlinked metal rings.*
chair: *seat with a backrest for one person.*
chaise_longue: *long chair designed for reclining.*

chalice: *decorative cup used for drinking, often in ceremonies.*
chandelier: *ornate lighting fixture hung from the ceiling.*
chap: *informal term for a young man or boy.*
checkbook: *book containing checks for making payments.*
checkerboard: *board with alternating colored squares for playing checkers.*
cherry: *small, round, red or yellow fruit with a pit.*
chessboard: *board with alternating squares for playing chess.*
chicken_(animal): *domesticated bird raised for meat or eggs.*
chickpea: *round, beige legume often used in cooking.*
chili_(vegetable): *spicy pepper used in various cuisines.*
chime: *musical instrument that produces a ringing sound.*
chinaware: *decorative ceramic dishware, often used for dining.*
crisp_(potato_chip): *thin, fried slice of potato, often crunchy.*
poker_chip: *small disc used in gambling games like poker.*
chocolate_bar: *solid confection made primarily of chocolate.*
chocolate_cake: *cake made with chocolate, often layered and frosted.*
chocolate_milk: *sweet beverage made by mixing chocolate with milk.*
chocolate_mousse: *light, creamy dessert made with chocolate.*
choker: *close-fitting necklace worn around the neck.*
chopping_board: *flat surface used for cutting food.*
chopstick: *pair of slender sticks used for eating asian cuisine.*
christmas_tree: *evergreen tree decorated with ornaments for the holiday.*
slide: *smooth surface for sliding down, often found in playgrounds.*
cider: *beverage made from pressed apples, often fermented.*
cigar_box: *container for storing cigars.*
cigarette: *thin roll of tobacco for smoking.*
cigarette_case: *holder for storing cigarettes, often decorative.*
cistern: *large tank for storing water.*
clarinet: *woodwind instrument with a single-reed mouthpiece.*
clasp: *fastening device for securing items together.*
cleansing_agent: *substance used for cleaning surfaces.*
cleat_(for_securing_rope): *device used to secure a rope on a boat.*
clementine: *small, sweet citrus fruit, often easy to peel.*
clip: *device for holding objects together, often metal or plastic.*
clipboard: *board with a clip for holding papers in place.*
clippers_(for_plants): *tool for cutting plants or flowers.*
cloak: *loose outer garment worn for warmth or protection.*
clock: *device for measuring and displaying time.*
clock_tower: *tall structure housing a clock, often in public spaces.*
clothes_hamper: *container for storing dirty laundry.*
clothespin: *wooden or plastic clip for hanging clothes to dry.*
clutch_bag: *small handbag designed to be held in hand.*
coaster: *small mat for placing drinks to protect surfaces.*
coat: *outer garment worn for warmth or protection.*
coat_hanger: *device for hanging coats to prevent wrinkling.*
coatrack: *furniture piece for hanging coats and jackets.*
cock: *male bird, often used in reference to roosters.*
cockroach: *brown, flattened insect known for its resilience.*
cocoa_(beverage): *hot drink made from cocoa powder and milk.*
coconut: *hard-shelled fruit with a brown outer husk and white flesh.*
coffee_maker: *appliance for brewing coffee.*
coffee_table: *low table for holding drinks and snacks in living rooms.*
coffeepot: *pot used for brewing and serving coffee.*
coil: *spiral or circular formation of wire or material.*
coin: *small, round piece of metal used as currency.*
colander: *bowl with holes for draining liquids from food.*
coleslaw: *salad made from shredded cabbage and dressing.*
coloring_material: *substances used for coloring, such as crayons or markers.*
combination_lock: *lock that requires a specific sequence of numbers to open.*
pacifier: *rubber or plastic device given to infants for sucking.*
comic_book: *illustrated publication that tells a story in comic form.*
compass: *tool used for navigation, indicating direction.*
computer_keyboard: *input device with keys for typing.*
condiment: *substance added to food for flavor, such as ketchup.*
cone: *geometric shape tapering to a point, often used for ice cream.*
control: *device or mechanism used to regulate something.*
convertible_(automobile): *car with a roof that can be folded down.*
sofa_bed: *couch that can be converted into a bed.*
cooker: *appliance used for cooking food.*
cookie: *sweet baked treat, often round and flat.*
cooking_utensil: *tool used in the kitchen for preparing food.*
cooler_(for_food): *insulated container for keeping food and drinks cold.*
cork_(bottle_plug): *plug used to seal a bottle, usually made of cork.*

corkboard: *board covered with cork for pinning notes and messages.*
corkscrew: *tool for removing corks from bottles.*
edible_corn: *yellow grain that can be eaten, often used as food.*
cornbread: *bread made from cornmeal, typically sweet and dense.*
cornet: *musical instrument resembling a trumpet.*
cornice: *decorative molding at the top of a wall or building.*
cornmeal: *ground corn used in cooking and baking.*
corset: *garment worn to shape the waist.*
costume: *outfit worn for a particular occasion or role.*
cougar: *large wild cat known for its agility and stealth.*
coverall: *one-piece garment worn for protection, often in work environments.*
cowbell: *bell hung around a cow's neck to locate it.*
cowboy_hat: *wide-brimmed hat associated with cowboy culture.*
crab_(animal): *shellfish with a hard shell and pincers.*
crabmeat: *meat from crabs, often used in cooking.*
cracker: *crisp, dry biscuit often eaten with cheese or dips.*
crape: *thin, textured fabric used in clothing and decorations.*
crate: *wooden or plastic container for transporting goods.*
crayon: *colored wax stick used for drawing and coloring.*
cream_pitcher: *small container for serving cream.*
crescent_roll: *soft, flaky pastry shaped like a crescent.*
crib: *bed for a baby or small child.*
crock_pot: *slow cooker for preparing meals over time.*
crossbar: *bar used for support or as a boundary.*
crouton: *small, crispy bread piece often used in salads.*
crow: *large, black bird known for its intelligence.*
crowbar: *tool used for prying and lifting objects.*
crown: *ornamental headpiece worn as a symbol of authority.*
crucifix: *cross with a representation of jesus, used in christianity.*
cruise_ship: *large vessel designed for leisure travel on the water.*
police_cruiser: *vehicle used by police for patrol and response.*
crumb: *small piece or fragment of baked goods.*
crutch: *supportive device used to aid walking.*
cub_(animal): *young animal, especially of bears or big cats.*
cube: *solid, three-dimensional shape with six equal square faces.*
cucumber: *long, green vegetable, often used in salads.*
cufflink: *decorative fastener used to secure the cuffs of a shirt.*
cup: *small, typically cylindrical container for drinking.*
trophy_cup: *award given for achievement, usually in sports.*
cupboard: *storage cabinet with shelves for dishes and food.*
cupcake: *small cake baked in a cup, often frosted.*
hair_curler: *device used to curl hair into waves or spirals.*
curling_iron: *heated tool for curling hair.*
curtain: *fabric panel hung to cover windows or divide rooms.*
cushion: *soft pad used for comfort, often on furniture.*
cylinder: *three-dimensional shape with circular bases.*
cymbal: *musical instrument that produces a crashing sound.*
dagger: *short, pointed knife used as a weapon.*
dalmatian: *breed of dog known for its distinctive black spots.*
dartboard: *board used for playing darts, marked with scores.*
date_(fruit): *sweet, chewy fruit from the date palm.*
deck_chair: *folding chair designed for outdoor use.*
deer: *hoofed animal known for its graceful movements.*
dental_floss: *thin string used for cleaning between teeth.*
desk: *surface for working, often with drawers for storage.*
detergent: *cleaning agent used for laundry and dishes.*
diaper: *absorbent garment worn by infants.*
diary: *book for recording daily thoughts and events.*
die: *cube used in games for generating random numbers.*
dinghy: *small boat, often used for short trips or as a tender.*
dining_table: *table used for eating meals.*
tux: *formal suit typically worn for black-tie events.*
dish: *shallow container for serving or cooking food.*
dish_antenna: *satellite dish used for receiving signals.*
dishrag: *cloth used for cleaning dishes.*
dishtowel: *towel used for drying dishes and utensils.*
dishwasher: *appliance for automatically washing dishes.*
dishwasher_detergent: *cleaning agent used in dishwashers.*
dispenser: *device for distributing a product, like soap or food.*
diving_board: *springboard for jumping into a swimming pool.*
dixie_cup: *small paper cup often used for serving drinks.*
dog: *domesticated animal known for companionship and loyalty.*
dog_collar: *strap worn around a dog's neck, often for identification.*
doll: *toy figure representing a human, often used for play.*
dollar: *unit of currency used in the united states and other countries.*
dollhouse: *miniature house designed for dolls and toys.*
dolphin: *intelligent marine mammal known for its playful nature.*
domestic_ass: *donkey used for work and companionship.*
doorknob: *handle used for opening and closing doors.*

doormat: *mat placed at the entrance to a home for wiping feet.*
doughnut: *fried sweet pastry, often ring-shaped.*
dove: *bird symbolizing peace, often white in color.*
dragonfly: *insect known for its large wings and agile flight.*
drawer: *compartment in furniture used for storage.*
underdrawers: *underwear or undergarments.*
dress: *garment worn by women, typically flowing and elegant.*
dress_hat: *formal hat worn with elegant attire.*
dress_suit: *formal outfit worn for special occasions.*
dresser: *furniture piece with drawers for storing clothes.*
drill: *tool used for making holes in various materials.*
drone: *unmanned aerial vehicle often used for surveillance or photography.*
dropper: *tool used for dispensing small amounts of liquid.*
drum_(musical_instrument): *percussion instrument with a circular body.*
drumstick: *stick used for playing a drum.*
duck: *medium-sized bird with a round body, webbed feet, and a flat beak.*
duckling: *small, fluffy bird with a yellow downy coat.*
duct_tape: *wide, silver or black adhesive tape with a textured surface.*
duffel_bag: *cylindrical, soft-sided bag with handles and a zipper.*
dumbbell: *short bar with round weights on each end.*
dumpster: *large, rectangular metal container with hinged lids.*
dustpan: *flat, scoop-shaped tray with a handle.*
eagle: *large bird with a hooked beak and broad wings.*
earphone: *small, round device with a cord, worn in the ear for listening.*
earplug: *small, cylindrical or conical foam piece designed to fit in the ear.*
earring: *small piece of jewelry worn on the earlobe, often shiny or decorative.*
easel: *wooden or metal stand with angled legs, used for holding canvases.*
eclair: *long, puffed pastry with a glossy, chocolate-glazed top.*
eel: *long, slender, snake-like fish with smooth skin.*
egg: *oval-shaped, smooth, and hard-shelled object, usually white or brown.*
egg_roll: *cylindrical roll wrapped in a golden, crispy shell.*
egg_yolk: *yellow, round center of an egg.*
eggbeater: *handheld whisk with rotating blades and a handle.*
eggplant: *large, glossy purple vegetable with an oblong shape.*
electric_chair: *chair with metal components and straps attached.*
refrigerator: *large, boxy appliance with a door and shelves inside.*
elephant: *large, gray animal with a long trunk and large ears.*
elk: *large deer with long antlers and a shaggy coat.*
envelope: *thin, rectangular paper sleeve with a flap for sealing.*
eraser: *soft, rubber block, usually pink or white.*
escargot: *small snail, often served with a spiral shell.*
eyepatch: *small, dark fabric piece worn over one eye.*
falcon: *sleek bird with sharp talons and a pointed beak.*
fan: *round or rectangular device with spinning blades inside a protective grill.*
faucet: *metal fixture with a curved spout and a handle for controlling water.*
fedora: *soft hat with a creased crown and a brim.*
ferret: *small, slender animal with soft fur and a pointed snout.*
ferris_wheel: *large, circular structure with passenger cars attached.*
ferry: *large, flat-bottomed boat with a ramp for vehicles and passengers.*
fig_(fruit): *small, round or oval fruit with a smooth purple or green skin.*
fighter_jet: *sleek aircraft with sharp wings and a pointed nose.*
figurine: *small, decorative model of a person or animal.*
file_cabinet: *tall, rectangular metal cabinet with pull-out drawers.*
file_(tool): *thin, metal tool with a rough surface for smoothing.*
fire_alarm: *small, red box with a glass panel and a lever.*
fire_engine: *large, red truck with hoses and ladders attached.*
fire_extinguisher: *red, cylindrical container with a nozzle and a handle.*
fire_hose: *long, thick, flexible tube coiled around a reel.*
fireplace: *brick or stone opening in a wall with a hearth for burning wood.*
fireplug: *short, round metal hydrant with caps and nozzles.*
first-aid_kit: *small box with a red cross symbol, containing medical supplies.*
fish: *sleek, scaly creature with fins and gills.*
fish_(food): *cooked or raw pieces of fish, often flaky or filleted.*
fishbowl: *round, transparent glass container filled with water.*
fishing_rod: *long, slender pole with a reel and fishing line attached.*
flag: *rectangular piece of fabric with patterns or symbols, attached to a pole.*
flagpole: *tall, straight pole with a mechanism for raising a flag.*

flamingo: *tall bird with pink feathers, long legs, and a curved neck.*
flannel: *soft, plaid-patterned fabric, often used in shirts.*
flap: *thin, flexible piece that covers an opening or edge.*
flash: *bright, quick burst of light.*
flashlight: *cylindrical handheld device with a lens for projecting light.*
fleece: *soft, fluffy fabric, often used for warmth.*
flip-flop_(sandal): *open-toe sandal with a y-shaped strap between the toes.*
flipper_(footwear): *large, flat, flexible fin worn on the feet for swimming.*
flower_arrangement: *assorted flowers grouped together, often in a vase.*
flute_glass: *tall, slender glass with a narrow bowl for holding champagne.*
foal: *young horse with a small, delicate frame and thin legs.*
folding_chair: *collapsible chair with metal or plastic legs and a fabric seat.*
food_processor: *boxy appliance with a clear bowl and rotating blades inside.*
football_(american): *oval-shaped, brown leather ball with white laces.*
football_helmet: *rounded helmet with a faceguard and chin strap.*
footstool: *small, padded stool for resting feet, often square or round.*
fork: *metal utensil with prongs, used for eating.*
forklift: *small, heavy-duty vehicle with prongs for lifting objects.*
freight_car: *long, boxy container attached to a train, used for carrying goods.*
french_toast: *golden-brown slices of bread with a soft texture.*
freshener: *small, scented item with a shape or design, used to freshen air.*
frisbee: *round, flat plastic disc with a slightly concave surface.*
frog: *small amphibian with smooth skin and long, strong legs.*
fruit_juice: *clear or pulpy liquid, often colorful, extracted from fruit.*
frying_pan: *flat, round metal pan with a long handle and low sides.*
fudge: *dense, smooth, chocolate confection, often cut into squares.*
funnel: *cone-shaped tool with a narrow spout for directing liquids.*
futon: *thin, foldable mattress, often placed on a low frame.*
gag: *small, restrictive device, often a strip of cloth or ball.*
garbage: *unwanted or discarded waste, often a mix of materials.*
garbage_truck: *large vehicle with a compactor and a bin-lifting mechanism.*
garden_hose: *long, flexible tube with a nozzle for spraying water.*
gargle: *liquid, often bubbly, swirled in the mouth.*
gargoyle: *stone or metal statue of a creature, often with wings and a grotesque face.*
garlic: *bulb with papery skin and multiple cloves, usually white or purple.*
gasmask: *protective mask with filters and eye covers.*
gazelle: *slender, graceful antelope with long legs and curved horns.*
gelatin: *clear, jiggly substance, often molded into shapes.*
gemstone: *small, faceted stone, usually shiny and colorful.*
generator: *boxy machine with vents, used for producing electricity.*
giant_panda: *large, black-and-white bear with a round body and fluffy fur.*
gift_wrap: *brightly colored paper used to wrap presents.*
ginger: *knobby, beige root with thin skin.*
giraffe: *tall animal with long neck, spotted coat, and small horns.*
cincture: *thin belt or cord worn around the waist, often made of fabric.*
glass_(drink_container): *transparent container with a round body and an open top.*
globe: *round model of the earth with geographical features displayed.*
glove: *fitted covering for the hand, often with separate sections for fingers.*
goat: *small animal with curved horns and a beard.*
goggles: *protective eyewear with round lenses and a strap.*
goldfish: *small, orange fish with a round body and flowing fins.*
golf_club: *long stick with a rounded end for hitting golf balls.*
golfcart: *small, open vehicle with a roof and seats.*
gondola_(boat): *long, narrow boat with a flat bottom.*
goose: *large bird with a long neck and webbed feet.*
gorilla: *large ape with a muscular build and dark fur.*
gourd: *hard-shelled fruit with a round or oblong shape.*
grape: *small, round fruit, usually green or purple.*
grater: *flat or boxy tool with sharp holes for shredding food.*
gravestone: *flat stone slab with engraved text.*
gravy_boat: *small, curved dish with a spout.*
green_bean: *long, thin vegetable with a smooth surface.*
green_onion: *tall, thin vegetable with white and green stalks.*
griddle: *flat, heated cooking surface.*
grill: *metal cooking grate with open flames beneath.*
grits: *coarse, creamy grain dish.*
grizzly: *large bear with thick brown fur.*

grocery_bag: *paper or plastic bag for carrying food.*
guitar: *stringed musical instrument with a hollow body.*
gull: *medium-sized bird with white feathers and long wings.*
gun: *metal weapon with a barrel and trigger.*
hairbrush: *handheld tool with bristles for brushing hair.*
hairnet: *thin mesh worn over hair.*
hairpin: *small metal pin used to hold hair in place.*
halter_top: *sleeveless shirt tied around the neck.*
ham: *thick slice of cured pork.*
hamburger: *round beef patty in a bun.*
hammer: *tool with a metal head and a handle.*
hammock: *hanging fabric sling for resting.*
hamper: *tall basket for holding laundry.*
hamster: *small, fluffy rodent with short legs.*
hair_dryer: *handheld device that blows hot air.*
hand_glass: *small mirror with a handle.*
hand_towel: *small, rectangular cloth for drying hands.*
handcart: *two-wheeled cart with handles.*
handcuff: *metal shackles linked by a chain.*
handkerchief: *small square cloth, often carried in a pocket.*
handle: *part of an object held by the hand.*
handsaw: *saw with a straight blade and a handle.*
hardback_book: *book with a stiff, protective cover.*
harmonium: *small, box-shaped musical instrument with keys.*
hat: *headwear with a brim or visor.*
hatbox: *round, rigid container for storing hats.*
veil: *thin fabric covering the face or head.*
headband: *stretchy band worn around the forehead.*
headboard: *upright panel at the head of a bed.*
headlight: *bright front light on a vehicle.*
headscarf: *cloth worn over the head or hair.*
headset: *headphones with a microphone attached.*
headstall_(for_horses): *leather straps placed over a horse's head.*
heart: *symbol shaped like a curved, pointed oval.*
heater: *boxy device for generating heat.*
helicopter: *aircraft with rotating blades on top.*
helmet: *hard, protective headgear.*
heron: *tall bird with long legs and a sharp beak.*
highchair: *chair with a raised seat for children.*
hinge: *metal joint that allows doors to swing open.*
hippopotamus: *large, bulky animal with thick gray skin.*
hockey_stick: *long, curved stick for playing hockey.*
hog: *large, domesticated pig.*
home_plate_(baseball): *five-sided, flat base in baseball.*
honey: *thick, golden liquid made by bees.*
fume_hood: *enclosed workstation with a vent.*
hook: *curved metal piece for hanging or catching objects.*
hookah: *tall, water-filled pipe with a long tube.*
hornet: *large, yellow and black stinging insect.*
horse: *large, four-legged animal with a mane.*
hose: *long, flexible tube for directing water.*
hot-air_balloon: *large, fabric balloon with a basket beneath.*
hotplate: *small, portable electric stove.*
hot_sauce: *spicy liquid condiment in a bottle.*
hourglass: *glass container with sand that flows through a narrow neck.*
houseboat: *boat designed for living on water.*
hummingbird: *tiny bird with a long beak and rapid wing beats.*
hummus: *smooth, creamy chickpea dip.*
polar_bear: *large, white bear with thick fur.*
icecream: *cold, creamy dessert in various flavors.*
popsicle: *frozen fruit-flavored treat on a stick.*
ice_maker: *machine that produces ice cubes.*
ice_pack: *flexible, cold pouch for pain relief.*
ice_skate: *shoe with a metal blade attached for skating.*
igniter: *small device used to ignite a flame.*
inhaler: *handheld device for delivering medication to the lungs.*
ipod: *small portable music player.*
iron_(for_clothing): *flat, heated tool for smoothing clothes.*
ironing_board: *long, narrow board with padding for ironing clothes.*
jacket: *short, outer garment with sleeves and a zipper.*
jam: *thick, sweet spread made from fruit.*
jar: *cylindrical glass container with a lid.*
jean: *denim pants with pockets and rivets.*
jeep: *boxy, rugged vehicle with large tires.*
jelly_bean: *small, bean-shaped candy with a glossy shell.*
jersey: *lightweight shirt worn for sports.*
jet_plane: *fast aircraft with a sleek body and jet engines.*
jewel: *small, shiny gemstone set in jewelry.*
jewelry: *decorative items like rings and necklaces, often made of metal.*
joystick: *handheld control stick for video games or machines.*
jumpsuit: *one-piece garment with sleeves and pants.*
kayak: *narrow, lightweight boat with pointed ends.*
keg: *large metal barrel with a rounded shape.*
kennel: *small, enclosed shelter for dogs.*
kettle: *round metal container with a spout and handle.*
key: *small, flat metal piece with a jagged edge.*
keycard: *flat plastic card with electronic data.*
kilt: *knee-length skirt with pleats, often plaid.*
kimono: *loose-fitting robe with wide sleeves.*
kitchen_sink: *shallow basin with a faucet for washing.*
kitchen_table: *flat surface with legs, used for dining.*
kite: *lightweight frame with fabric stretched over it, flown in the wind.*
kitten: *small, fluffy baby cat.*
kiwi_fruit: *brown, fuzzy fruit with green flesh inside.*
knee_pad: *cushioned pad for protecting the knee.*

knife: *sharp metal blade with a handle.*
knitting_needle: *long, pointed stick used for knitting.*
knob: *round handle or switch for turning.*
knocker_(on_a_door): *metal ring or bar for knocking on doors.*
koala: *small, gray furry animal with large ears and a round nose.*
lab_coat: *long white coat worn by scientists.*
ladder: *vertical structure with steps for climbing.*
ladle: *long-handled spoon with a deep bowl.*
ladybug: *small red insect with black spots.*
lamb_(animal): *young sheep with soft wool.*
lamb-chop: *slice of lamb meat with a bone.*
lamp: *light source with a bulb and a shade.*
lamppost: *tall post with a light on top.*
lampshade: *cover for a light bulb, usually conical or cylindrical.*
lantern: *portable light with a handle and enclosed flame or bulb.*
lanyard: *cord worn around the neck for holding items.*
laptop_computer: *portable computer with a foldable screen and keyboard.*
lasagna: *layered pasta dish with sauce and cheese.*
latch: *metal fastener for securing doors or gates.*
lawn_mower: *machine with rotating blades for cutting grass.*
leather: *smooth, flexible material made from animal hide.*
legging_(clothing): *tight-fitting pants, often stretchy.*
lego: *small, colorful plastic building blocks.*
legume: *edible seed or pod, like beans or peas.*
lemon: *small, yellow citrus fruit with a bumpy skin.*
lemonade: *sweet, yellow drink made from lemons.*
lettuce: *leafy green vegetable with a crisp texture.*
license_plate: *metal plate with letters and numbers on vehicles.*
life_buoy: *ring-shaped flotation device used for rescue.*
life_jacket: *padded vest worn for flotation in water.*
lightbulb: *glass bulb that produces light when powered.*
lightning_rod: *metal rod placed on buildings to prevent lightning damage.*
lime: *small, green citrus fruit with a sour taste.*
limousine: *long, luxurious car with multiple doors.*
lion: *large, muscular animal with a mane and a tawny coat.*
lip_balm: *small stick or tube of moisturizing substance for lips.*
liquor: *clear or amber-colored alcoholic beverage in a bottle.*
lizard: *small reptile with scaly skin and a long tail.*
log: *thick, solid piece of wood from a tree trunk.*
lollipop: *hard candy on a stick.*
speaker_(stereo_equipment): *boxy device that emits sound.*
loveseat: *small, cushioned sofa for two people.*
machine_gun: *long-barreled gun with rapid-fire capability.*
magazine: *thin booklet filled with articles and images.*
magnet: *small, metallic object that attracts or repels metal.*
mail_slot: *narrow opening in a door for delivering mail.*
mailbox_(at_home): *metal or plastic box for receiving mail.*
mallard: *duck with a green head and a brown body.*
mallet: *hammer with a large, soft head.*
mammoth: *large, woolly prehistoric elephant-like animal.*
manatee: *large, gray aquatic mammal with flippers.*
mandarin_orange: *small, round orange citrus fruit with loose skin.*
manger: *trough for feeding animals, typically found in stables.*
manhole: *round opening in the ground with a metal cover.*
map: *flat representation of geographic areas.*
marker: *thick pen with colorful ink.*
martini: *clear alcoholic drink served in a wide glass.*
mascot: *character or symbol representing a team or organization.*
mashed_potato: *soft, smooth dish made from cooked potatoes.*
masher: *utensil with a flat, grid-like head for mashing food.*
mask: *covering for the face, often with eye holes.*
mast: *tall vertical pole on a ship, supporting sails.*
mat_(gym_equipment): *flat, cushioned surface for exercise.*
matchbox: *small, rectangular box for holding matches.*
mattress: *thick, padded cushion for sleeping.*
measuring_cup: *cup with markings for measuring ingredients.*
measuring_stick: *long stick with markings for measuring length.*
meatball: *round ball of ground meat, usually cooked.*
medicine: *small, solid pills or liquid in a bottle for treatment.*
melon: *large, round fruit with a hard rind and sweet flesh.*
microphone: *small, handheld device for amplifying sound.*
microscope: *optical device with lenses for viewing tiny objects.*
microwave_oven: *box-shaped appliance for quickly heating food.*
milestone: *small, cylindrical stone marker for measuring distance.*
milk: *white, liquid dairy product in a container.*
milk_can: *tall, round metal container for holding milk.*
milkshake: *cold, creamy drink made from milk and ice cream.*
minivan: *boxy, spacious vehicle with sliding doors.*
mint_candy: *small, round candy with a minty flavor.*
mirror: *flat, reflective surface, usually framed.*
mitten: *handwear with a single compartment for the fingers.*

mixer_(kitchen_tool): *handheld or stand appliance with rotating beaters.*
money: *paper bills or metal coins used for transactions.*
monitor_(computer_equipment) computer_monitor: *flat screen for displaying computer output.*
monkey: *small, agile primate with a long tail.*
motor: *box-shaped device that produces mechanical power.*
motor_scooter: *compact, two-wheeled vehicle with a small frame and wide seat.*
motor_vehicle: *large, enclosed vehicle with wheels, typically rectangular or streamlined.*
motorcycle: *two-wheeled, streamlined vehicle with exposed engine and handlebars.*
mound_(baseball): *slightly elevated, round, grassy surface.*
mouse_(computer_equipment): *small, curved device with a smooth surface and buttons.*
mousepad: *thin, flat rectangular surface, often with a soft texture.*
muffin: *small, round, domed baked good with a crinkled surface.*
mug: *cylindrical container with a handle and thick sides.*
mushroom: *rounded cap on a thin, cylindrical stem.*
music_stool: *simple, round or square seat with three or four legs.*
musical_instrument: *various shapes; generally wooden or metal with buttons, strings, or keys.*
nailfile: *long, thin, flat tool with a rough surface.*
napkin: *soft, square piece of cloth or paper, often folded.*
neckerchief: *triangular or square cloth tied around the neck.*
necklace: *thin, looped chain or string with small decorative items.*
necktie: *long, narrow fabric strip, often pointed at one end.*
needle: *thin, pointed metal rod.*
nest: *woven or layered structure, often circular.*
newspaper: *thin, rectangular, folded sheets of printed paper.*
newsstand: *tall, square or rectangular display rack with shelves.*
nightshirt: *loose-fitting, long-sleeved shirt.*
nosebag_(for_animals): *round, soft fabric container with straps.*
noseband_(for_animals): *thin strap placed around the muzzle.*
notebook: *flat, rectangular object with pages bound together.*
notepad: *small, square or rectangular pad with blank pages.*
nut: *small, hard-shelled object, typically round.*
nutcracker: *metal or wooden device with a handle and hinged mechanism.*
oar: *long, wooden pole with a wide, flat end.*
octopus_(food): *curled tentacles with suction cups and a glossy surface.*
octopus_(animal): *round body with long, smooth tentacles.*
oil_lamp: *glass or metal container with a wick and curved body.*
olive_oil: *smooth, greenish-yellow liquid in a clear bottle.*
omelet: *flat, folded, fluffy dish with a golden surface.*
onion: *round, layered vegetable with a papery skin.*
orange_(fruit): *round, bright orange fruit with a textured skin.*
orange_juice: *yellowish liquid, often in a clear glass.*
ostrich: *tall bird with long neck, fluffy body, and strong legs.*
ottoman: *soft, rectangular or square padded footrest.*
oven: *large, rectangular metal appliance with a front door and knobs.*
overalls_(clothing): *one-piece garment with straps and multiple pockets.*
owl: *small, round bird with large eyes and feathery tufts.*
packet: *small, flat, rectangular or square pouch.*
inkpad: *small, flat container with a spongy ink surface.*
pad: *soft, square or rectangular cushion.*
paddle: *long, flat wooden or plastic tool with a wide, flat end.*
padlock: *small, metallic, rectangular lock with a u-shaped shackle.*
paintbrush: *thin, cylindrical handle with soft bristles at one end.*
painting: *flat, rectangular surface with colorful images or designs.*
pajamas: *loose-fitting, soft two-piece clothing set.*
palette: *flat, oval surface with paint dabs.*
pan_(for_cooking): *round, shallow metal container with a handle.*
pan_(metal_container): *square or rectangular metal box for holding or cooking items.*
pancake: *flat, round, golden-brown cooked batter.*
pantyhose: *thin, stretchy fabric in long, leg-shaped tubes.*
papaya: *oval fruit with greenish-yellow skin and orange flesh inside.*
paper_plate: *flat, round, lightweight disposable plate.*
paper_towel: *long, rectangular piece of absorbent paper.*
paperback_book: *thin, rectangular book with a flexible cover.*
paperweight: *small, heavy, decorative object used to hold papers.*
parachute: *large, round fabric canopy with thin ropes attached.*
parakeet: *small, colorful bird with long tail feathers.*
parasail_(sports): *large, colorful parachute with ropes attached to a harness.*
parasol: *small, round umbrella with a thin handle.*

parchment: *thin, flat, beige or white paper-like sheet.*
parka: *long, insulated jacket with a hood.*
parking_meter: *tall, thin post with a round, metal head.*
parrot: *brightly colored bird with a curved beak.*
passenger_car_(part_of_a_train): *long, rectangular vehicle with rows of windows.*
passenger_ship: *large, multi-decked boat with windows and cabins.*
passport: *small, rectangular booklet with a hard cover.*
pastry: *small, flaky, golden-brown baked item.*
patty_(food): *round, flat, cooked food item.*
pea_(food): *small, round, green vegetable.*
peach: *round, fuzzy fruit with a reddish-yellow color.*
peanut_butter: *thick, smooth or chunky brown paste.*
pear: *teardrop-shaped, smooth-skinned fruit.*
peeler_(tool_for_fruit_and_vegetables): *small tool with a handle and a curved blade.*
wooden_leg: *long, cylindrical piece of wood shaped like a leg.*
pegboard: *large, flat board with regularly spaced holes.*
pelican: *large bird with a long bill and a pouch under the beak.*
pen: *thin, cylindrical writing instrument with a pointed tip.*
pencil: *slim, cylindrical stick with a graphite core and wooden body.*
pencil_box: *small, rectangular container for holding pencils.*
pencil_sharpener: *small, rectangular or cylindrical device with a blade for sharpening pencils.*
pendulum: *long, thin rod or string with a weight at the end.*
penguin: *short, stocky bird with black and white feathers.*
pennant: *triangular fabric flag attached to a stick or string.*
penny_(coin): *small, round, copper-colored coin.*
pepper: *small, round or long vegetable, smooth or wrinkled skin.*
pepper_mill: *tall, cylindrical container with a rotating top for grinding pepper.*
perfume: *small, decorative glass bottle with a sprayer.*
persimmon: *round, smooth-skinned orange fruit.*
person: *human figure with limbs and distinct features.*
pet: *small, domesticated animal with varied shapes and sizes.*
pew_(church_bench): *long, narrow wooden bench with a high back.*
phonebook: *thick, rectangular book with printed names and numbers.*
phonograph_record: *flat, round disc with grooves.*
piano: *large, rectangular instrument with black and white keys.*
pickle: *long, green, bumpy vegetable preserved in liquid.*
pickup_truck: *large, boxy vehicle with an open cargo bed.*
pie: *round, flat pastry with a golden crust, often with a textured or lattice top.*
pigeon: *small bird with a round body, smooth feathers, and a short beak.*
piggy_bank: *small, hollow, ceramic or plastic figure, often shaped like a pig, with a coin slot on top.*
pillow: *soft, rectangular or square cushion with a smooth fabric covering.*
pin_(non_jewelry): *small, thin, straight metal object with a round head.*
pineapple: *large, oval-shaped fruit with rough, spiky skin and a tuft of spiky green leaves.*
pinecone: *brown, scaly, conical structure with a rough texture.*
ping-pong_ball: *small, round, lightweight white or orange ball.*
pinwheel: *colorful, flat, spinning paper or plastic shape on a thin stick.*
tobacco_pipe: *small, curved object with a bowl and a long stem.*
pipe: *long, cylindrical tube, typically metal or plastic.*
pistol: *compact, angular, metal firearm with a grip and barrel.*
pita_(bread): *round, flat bread with a slightly puffed texture.*
pitcher_(vessel_for_liquid): *tall, cylindrical container with a handle and spout.*
pitchfork: *long handle with a metal head featuring several sharp, straight prongs.*
pizza: *flat, round base topped with various colorful ingredients and a golden crust.*
place_mat: *flat, rectangular or circular mat, usually made of fabric or plastic.*
plate: *flat, round dish with a slightly raised edge.*
platter: *large, flat, oval or rectangular dish with raised edges.*
playpen: *square or rectangular, soft-sided enclosure with mesh or plastic walls.*
pliers: *handheld tool with two metal, pivoting arms ending in gripping jaws.*
plow_(farm_equipment): *large metal blade or series of blades attached to a long frame.*
plume: *long, colorful feather with a soft, fluffy texture.*
pocket_watch: *small, round, metal watch with a hinged cover and chain.*
pocketknife: *small, foldable knife with a smooth, rectangular handle.*
poker_(fire_stirring_tool): *long, thin metal rod with a pointed tip and handle.*
pole: *tall, thin, cylindrical object, often made of wood or metal.*
polo_shirt: *short-sleeved shirt with a collar and buttons near the neck.*
poncho: *large, rectangular or circular piece of fabric with an opening for the head.*

pony: *small, sturdy horse with a stocky build and a short mane.*
pool_table: *large, rectangular, green felt-covered table with pockets at the corners.*
pop_(soda): *clear or colored liquid in a tall, cylindrical can or bottle.*
postbox_(public): *tall, rectangular or cylindrical box with a small mail slot and a rounded top.*
postcard: *small, rectangular card with a printed image on one side and space for writing on the other.*
poster: *large, flat sheet of paper with images or text, often rectangular.*
pot: *round, deep container with a flat base and straight or sloped sides.*
flowerpot: *small, round, usually terracotta or plastic container with a flared top.*
potato: *oval, rough-textured tuber with a brown or yellowish skin.*
potholder: *small, flat, square piece of padded fabric with a quilted texture.*
pottery: *rounded, smooth or textured objects, typically made of clay, with various shapes.*
pouch: *small, soft, flat container with a drawstring or zipper.*
power_shovel: *large, heavy machine with a long arm and a large bucket at the end.*
prawn: *small, elongated sea creature with a curved, segmented body and thin legs.*
pretzel: *twisted, baked bread snack with a golden-brown, shiny surface.*
printer: *rectangular device with buttons and paper slots, often with a flat top.*
projectile_(weapon): *small, sleek object with a pointed tip and a cylindrical or round body.*
projector: *boxy device with a lens in the front, typically mounted on a stand.*
propeller: *round, central hub with multiple flat blades extending outward.*
prune: *small, wrinkled, dark-colored dried fruit.*
pudding: *smooth, thick, creamy dessert often served in small bowls.*
puffer_(fish): *round, spiny fish with a bloated appearance when inflated.*
puffin: *small bird with a black and white body and a brightly colored beak.*
pug-dog: *small dog with a wrinkled face, short snout, and curled tail.*
pumpkin: *large, round, orange fruit with a smooth, ridged surface.*
puncher: *small, metal device with a handle and circular punching mechanism.*
puppet: *small, soft figure with strings or hand controls for movement.*
puppy: *small, fluffy young dog with large eyes and short legs.*
quesadilla: *flat, round tortilla filled with melted cheese and other ingredients.*
quiche: *round, golden-brown pastry filled with egg and other ingredients.*
quilt: *large, rectangular blanket made of small, colorful fabric patches.*
rabbit: *small, furry animal with long ears and a short, fluffy tail.*
race_car: *low, sleek vehicle with a streamlined body and large tires.*
racket: *oval frame with a tight web of strings attached to a long handle.*
radar: *large, flat or round dish mounted on a stand or structure.*
radiator: *rectangular metal panel with a series of vertical or horizontal fins.*
radio_receiver: *rectangular or square device with dials and a small antenna.*
radish: *small, round or oval root vegetable with a smooth red or white skin.*
raft: *flat, rectangular inflatable or wooden platform for floating.*
rag_doll: *soft, fabric doll with stitched features and loose limbs.*
raincoat: *long, waterproof jacket with a smooth, shiny surface.*
ram_(animal): *sturdy, woolly animal with large, curved horns.*
raspberry: *small, round, red fruit made up of tiny clusters.*
rat: *small rodent with a long tail, pointed nose, and short fur.*
razorblade: *thin, flat piece of metal with a sharp edge.*
reamer_(juicer): *small, pointed, ridged tool for extracting juice.*
rearview_mirror: *small, rectangular mirror mounted inside a vehicle.*
receipt: *long, narrow piece of paper with printed text.*
recliner: *large, padded chair with a footrest and adjustable back.*
record_player: *large, flat device with a spinning turntable and a needle arm.*
reflector: *small, round or rectangular object with a shiny or reflective surface.*
remote_control: *small, rectangular device with buttons and a smooth surface.*
rhinoceros: *large, thick-skinned animal with one or two horns on its snout.*
rib_(food): *curved bone with a layer of cooked meat attached.*
rifle: *long, slender firearm with a smooth barrel and stock.*
ring: *small, circular band of metal or other material.*

river_boat: *long, flat vessel with a shallow draft and a wide deck.*
road_map: *large, flat, folded paper with printed lines and symbols.*
robe: *long, loose-fitting garment with a belt and wide sleeves.*
rocking_chair: *wooden or padded chair with curved, arched rockers at the base.*
rodent: *small, furry animal with a pointed nose and long tail.*
roller_skate: *shoe with four wheels attached to the bottom in two rows.*
rollerblade: *boot with a single row of wheels attached to the sole.*
rolling_pin: *long, cylindrical tool with handles on each end.*
root_beer: *dark, foamy beverage in a glass or bottle.*
router_(computer_equipment): *small, boxy device with antennas and lights on the front.*
rubber_band: *thin, flat, stretchy loop of rubber.*
runner_(carpet): *long, narrow strip of carpet with bound edges.*
plastic_bag: *thin, transparent or colored, flexible bag with handles.*
saddle_(on_an_animal): *curved leather seat with stirrups attached.*
saddle_blanket: *rectangular, thick fabric pad placed under a saddle.*
saddlebag: *soft, rectangular bag that hangs from a saddle, often with buckles.*
safety_pin: *small, thin metal pin with a coil and a clasp at one end.*
sail: *large, flat, triangular or rectangular fabric stretched between poles.*
salad: *colorful mixture of vegetables, usually with green leaves.*
salad_plate: *small, round plate, often with a slight rim.*
salami: *cylindrical sausage with a textured, dark red exterior.*
salmon_(fish): *streamlined fish with a silvery body and pinkish scales.*
salmon_(food): *pink, flaky fish meat often served in slices or fillets.*
salsa: *bright, chunky mixture of chopped vegetables or fruit in a bowl.*
saltshaker: *small, round container with holes on top for dispensing salt.*
sandal_(type_of_shoe): *open shoe with straps across the foot and around the heel.*
sandwich: *two pieces of bread with various fillings inside.*
satchel: *soft, rectangular bag with a long strap, usually worn over the shoulder.*
saucepan: *deep, round metal pan with a long handle.*
saucer: *small, round plate with a slight indentation for a cup.*
sausage: *cylindrical meat with a smooth, shiny skin.*
sawhorse: *a-frame stand with a flat top used to support wood.*
saxophone: *shiny, curved metal instrument with a series of buttons and a mouthpiece.*
scale_(measuring_instrument): *flat platform with a display screen and buttons.*
scarecrow: *human-like figure with clothes and straw limbs, often mounted on a stick.*
scarf: *long, soft piece of fabric worn around the neck.*
school_bus: *long, rectangular vehicle, typically yellow, with rows of windows.*
scissors: *two metal blades connected by a pivot, with handles for cutting.*
scoreboard: *large, flat panel with numbers or lights for displaying scores.*
scraper: *flat, thin tool with a sharp edge for scraping surfaces.*
screwdriver: *tool with a long, narrow shaft and a handle, often with a flat or cross tip.*
scrubbing_brush: *small brush with stiff bristles and a short handle.*
sculpture: *solid, three-dimensional artwork made from stone, metal, or other materials.*
seabird: *medium-sized bird with sleek feathers and long wings.*
seahorse: *small, curved-bodied sea creature with a long tail and a horse-like head.*
seaplane: *small airplane with floats or pontoons for landing on water.*
seashell: *small, smooth or ridged shell with a spiral or fan shape.*
sewing_machine: *boxy device with a needle, thread spool, and a flat working surface.*
shaker: *small, round container used to shake and sprinkle spices or condiments.*
shampoo: *tall, cylindrical bottle with a flip cap, containing liquid soap.*
shark: *large, sleek fish with a pointed fin and sharp teeth.*
sharpener: *small device with a hole for inserting pencils, often with a sharp blade inside.*
sharpie: *thin, cylindrical marker with a pointed tip.*
shaver_(electric): *handheld device with rotating or oscillating blades for trimming hair.*
shaving_cream: *soft, foamy substance in a pressurized canister.*
shawl: *large, rectangular or triangular piece of fabric draped over the shoulders.*
shears: *large scissors with long blades, used for cutting thick materials.*
sheep: *woolly animal with a round body and a gentle face.*

shepherd_dog: *medium-sized, furry dog with pointed ears and a long snout.*
sherbert: *soft, smooth, brightly colored frozen dessert.*
shield: *large, round or rectangular piece of metal or wood, often with a handle.*
shirt: *soft, collared or uncollared garment with buttons and sleeves.*
shoe: *footwear with a solid sole and an enclosed upper.*
shopping_bag: *large, rectangular bag with handles, made of paper or plastic.*
shopping_cart: *metal or plastic basket on wheels with a handle.*
short_pants: *knee-length pants with a waistband and pockets.*
shot_glass: *small, cylindrical glass with thick sides, used for serving drinks.*
shoulder_bag: *soft, rectangular bag with a long strap, worn over the shoulder.*
shovel: *long-handled tool with a wide, flat metal blade for digging.*
shower_head: *circular or square device with holes, attached to a water pipe.*
shower_cap: *waterproof, elastic banded cover for the head.*
shower_curtain: *large, rectangular piece of fabric or plastic hung to block water.*
shredder_(for_paper): *rectangular box with slots for inserting paper, often with blades inside.*
signboard: *flat, rectangular or square board with writing or symbols.*
silo: *tall, cylindrical structure, often with a domed top, for storing grain.*
sink: *shallow basin with a faucet and drain, typically made of metal or porcelain.*
skateboard: *flat, narrow board with four wheels and upward-curved ends.*
skewer: *thin, sharp metal or wooden stick for grilling food.*
ski: *long, narrow board with a slightly curved tip, worn on the feet for sliding on snow.*
ski_boot: *stiff, high-ankle boot with clips or straps for attaching to skis.*
ski_parka: *thick, padded jacket with a hood, designed for cold weather.*
ski_pole: *thin, metal rod with a handle and a small disc near the base.*
skirt: *soft, flowy garment that wraps around the waist and falls to the legs.*
skullcap: *small, tight-fitting cap worn on the head.*
sled: *flat, narrow platform with runners for sliding over snow.*
sleeping_bag: *long, padded fabric bag with a zipper, used for sleeping.*
sling_(bandage): *soft, triangular or rectangular fabric strap for supporting an arm.*
slipper_(footwear): *soft, low-cut shoe with an open or closed heel.*
smoothie: *thick, colorful drink in a glass or bottle, often with fruit on top.*
snake: *long, slender, scaly reptile with a pointed head and a flicking tongue.*
snowboard: *flat, wide board with a slightly upturned front, used for sliding on snow.*
snowman: *rounded figure made of stacked snowballs, often with a scarf and hat.*
snowmobile: *low, motorized vehicle with tracks or skis for traveling on snow.*
soap: *small, rectangular or oval block of solid cleaning substance.*
soccer_ball: *round, black and white paneled ball.*
sock: *soft, tubular fabric worn on the foot.*
sofa: *large, cushioned seat with a backrest and armrests.*
softball: *round ball, slightly larger than a baseball, with visible stitching.*
solar_array: *flat, rectangular panels with a shiny, grid-like surface.*
sombrero: *large, round hat with a wide, floppy brim and a high crown.*
soup: *liquid meal in a bowl, often with visible ingredients like vegetables or noodles.*
soup_bowl: *deep, round bowl with a wide rim, often ceramic.*
soupspoon: *round, shallow spoon with a wide bowl.*
sour_cream: *soft, thick, white dairy product in a small bowl or tub.*
soya_milk: *pale, creamy liquid in a glass or carton.*
space_shuttle: *large, sleek, winged spacecraft with a pointed nose.*
sparkler_(fireworks): *thin metal stick that produces bright, sparkling light when lit.*
spatula: *flat, wide tool with a handle, often used for flipping food.*
spear: *long, straight pole with a pointed tip, often metal.*
spectacles: *thin, wire or plastic frame with two round or rectangular lenses.*
spice_rack: *small, rectangular shelf with jars or bottles of spices.*
spider: *small, eight-legged creature with a round body and segmented legs.*
crawfish: *small, lobster-like crustacean with a hard shell and long claws.*
sponge: *soft, porous material, often rectangular or oval, used for cleaning.*
spoon: *small, curved utensil with a long handle and a round, shallow bowl.*
sportswear: *colorful athletic clothing for exercise.*
spotlight: *round, bright light used for illumination.*
squid_(food): *soft, often translucent, sea creature.*

squirrel: *small, furry rodent with a bushy tail.*
stagecoach: *wooden carriage on wheels, often vintage.*
stapler_(stapling_machine): *metal and plastic device for fastening papers.*
starfish: *star-shaped marine animal with five arms.*
statue_(sculpture): *three-dimensional sculpture made of stone or metal.*
steak_(food): *thick cut of meat, usually red or brown.*
steak_knife: *sharp, serrated knife with a smooth handle.*
steering_wheel: *round wheel for controlling a vehicle.*
stepladder: *foldable ladder with steps for reaching heights.*
step_stool: *small stool for standing on to reach higher places.*
stereo_(sound_system): *boxy audio system with speakers.*
stew: *thick, hearty mixture of ingredients in a bowl.*
stirrer: *long utensil used for mixing liquids.*
stirrup: *curved support for a rider's foot on a saddle.*
stool: *short, simple seat without a backrest.*
stop_sign: *red, octagonal sign indicating to halt.*
brake_light: *red light on vehicles that indicates stopping.*
stove: *flat surface appliance for cooking food.*
strainer: *bowl-shaped tool with holes for draining liquids.*
strap: *long, narrow band for securing or holding.*
straw_(for_drinking): *thin tube for sipping liquids.*
strawberry: *red, heart-shaped fruit with tiny seeds.*
street_sign: *rectangular sign displaying road information.*
streetlight: *tall pole with a light for illuminating roads.*
string_cheese: *soft cheese in long, thin strands.*
stylus: *thin pen-like tool for touchscreen devices.*
subwoofer: *large speaker for deep bass sounds.*
sugar_bowl: *small container for holding sugar.*
sugarcane_(plant): *tall plant with thick, jointed stems.*
suit_(clothing): *formal clothing consisting of a jacket and trousers.*
sunflower: *tall plant with a large yellow flower head.*
sunglasses: *dark glasses for protecting eyes from sunlight.*
sunhat: *wide-brimmed hat for shading the face.*
surfboard: *long, flat board for riding ocean waves.*
sushi: *small rolls of rice and seafood, often colorful.*
mop: *long-handled tool with a cloth head for cleaning.*
sweat_pants: *loose-fitting pants for comfort and exercise.*
sweatband: *cloth band worn on the forehead during sports.*
sweater: *knitted garment worn on the upper body.*
sweatshirt: *pullover top made of thick, soft fabric.*
sweet_potato: *orange, starchy tuber with a smooth skin.*
swimsuit: *tight-fitting clothing for swimming.*
sword: *long, sharp weapon with a straight blade.*
syringe: *cylindrical tube used for injecting liquids.*
tabasco_sauce: *small bottle containing spicy pepper sauce.*
table-tennis_table: *rectangular surface for playing table tennis.*
table: *flat surface supported by legs.*
table_lamp: *small lamp with a base for tables.*
tablecloth: *fabric cover for protecting a table.*
tachometer: *gauge measuring engine speed, often circular.*
taco: *folded tortilla filled with various ingredients.*
tag: *small label attached to an item for identification.*
taillight: *red light at the back of vehicles.*
tambourine: *circular instrument with jingles around the edge.*
army_tank: *armored vehicle with a rotating turret.*
tank_(storage_vessel): *large container for holding liquids.*
tank_top_(clothing): *sleeveless shirt often worn casually.*
tape_(sticky_cloth_or_paper): *flexible strip for sticking items together.*
tape_measure: *flexible ruler for measuring lengths.*
tapestry: *decorative fabric with intricate designs.*
tarp: *large, waterproof cover, often made of plastic.*
tartan: *pattern of crisscrossed horizontal and vertical bands.*
tassel: *decorative hanging made of threads or strands.*
tea_bag: *small pouch containing tea leaves.*
teacup: *small cup used for drinking tea.*
teakettle: *metal pot with a spout for boiling water.*
teapot: *container for brewing and serving tea.*
teddy_bear: *soft, stuffed toy bear.*
telephone: *device for voice communication.*
telephone_booth: *small enclosure for making phone calls.*
telephone_pole: *tall post for holding telephone wires.*
telephoto_lens: *long camera lens for distant subjects.*
television_camera: *device for capturing moving images.*
television_set: *boxy device for displaying video content.*
tennis_ball: *yellow, fuzzy ball used in tennis.*
tennis_racket: *long handle with a netted head for hitting balls.*
tequila: *clear, alcoholic beverage in a bottle.*
thermometer: *instrument for measuring temperature.*
thermos_bottle: *insulated container for keeping liquids hot or cold.*
thermostat: *device for regulating temperature.*
thimble: *small protective cap for sewing fingers.*
thread: *thin strand used for sewing.*
thumbtack: *small pin for attaching papers to surfaces.*
tiara: *ornate crown worn on the head.*
tiger: *large striped cat with orange and black fur.*
tights_(clothing): *close-fitting legwear made of stretchy fabric.*
timer: *device for measuring time intervals.*
tinfoil: *thin, shiny metal sheet used for wrapping.*
tinsel: *shiny strands used for decoration.*
tissue_paper: *soft, thin paper often used for wrapping.*
toast_(food): *browned slice of bread.*
toaster: *appliance for browning bread slices.*
toaster_oven: *small oven for toasting and baking.*
toilet: *ceramic fixture for waste disposal.*
toilet_tissue: *soft paper for personal hygiene.*

tomato: *round, red fruit with a smooth skin.*
tongs: *pincers used for gripping food.*
toolbox: *box for storing tools, often metal or plastic.*
toothbrush: *small brush for cleaning teeth.*
toothpaste: *cream used with a toothbrush for cleaning teeth.*
toothpick: *small stick for cleaning between teeth.*
cover: *material used to protect or enclose something.*
tortilla: *flat, round bread made from corn or wheat.*
tow_truck: *vehicle designed for pulling other vehicles.*
towel: *absorbent cloth for drying hands or body.*
towel_rack: *bar for hanging towels.*
toy: *object for play, often colorful and fun.*
tractor_(farm_equipment): *large vehicle for farming tasks.*
traffic_light: *signal with red, yellow, and green lights.*
dirt_bike: *small motorcycle designed for off-road riding.*
trailer_truck: *large vehicle for transporting goods.*
train_(railroad_vehicle): *long vehicle on tracks for transporting people or goods.*
trampoline: *bouncy surface for jumping.*
tray: *flat surface for carrying items.*
trench_coat: *long, waterproof coat for protection.*
triangle_(musical_instrument): *three-sided musical instrument.*
tricycle: *three-wheeled bicycle for children.*
tripod: *stand with three legs for stabilizing cameras.*
trousers: *long pants for clothing.*
truck: *large vehicle for transporting goods.*
truffle_(chocolate): *rich chocolate confection, often round.*
trunk: *large, sturdy container for storage.*
vat: *large container for holding liquids.*
turban: *cloth wrapped around the head.*
turkey_(food): *large bird, often cooked for holidays.*
turnip: *round, edible root vegetable.*
turtle: *reptile with a hard shell.*
turtleneck_(clothing): *close-fitting sweater with a high collar.*
typewriter: *mechanical device for typing letters.*
umbrella: *portable cover for protection from rain.*
underwear: *garments worn beneath outer clothing.*
unicycle: *one-wheeled bicycle for balance.*
urinal: *fixture for urinating, usually in restrooms.*
urn: *decorative container for holding ashes or flowers.*
vacuum_cleaner: *device for cleaning floors using suction.*
vase: *container for holding flowers.*
vending_machine: *automated machine for selling snacks or drinks.*
vent: *opening for air circulation.*
vest: *sleeveless garment worn over a shirt.*
videotape: *magnetic tape for recording video.*
vinegar: *sour liquid used in cooking and dressing.*
violin: *string instrument played with a bow.*
vodka: *clear, distilled alcoholic beverage.*
volleyball: *light ball used in the sport of volleyball.*
vulture: *large bird often seen scavenging.*
waffle: *light, crispy batter cooked in a patterned iron.*
waffle_iron: *device for cooking waffles.*
wagon: *four-wheeled vehicle for transporting items.*
wagon_wheel: *round wooden or metal wheel for wagons.*
walking_stick: *long stick used for support while walking.*
wall_clock: *clock mounted on a wall.*
wall_socket: *electrical outlet for plugging in devices.*
wallet: *small, foldable case for holding money and cards.*
walrus: *large marine mammal with tusks.*
wardrobe: *tall cabinet for storing clothes.*
washbasin: *bowl for washing hands or face.*
automatic_washer: *appliance for cleaning clothes.*
watch: *small timekeeping device worn on the wrist.*
water_bottle: *container for holding drinking water.*
water_cooler: *appliance for dispensing chilled water.*
water_faucet: *tap for controlling water flow.*
water_heater: *device for heating water.*
water_jug: *large container for holding water.*
water_gun: *toy for spraying water.*
water_scooter: *small vehicle for riding on water.*
water_ski: *equipment for skiing on water.*
water_tower: *tall structure for storing water.*
watering_can: *container with a spout for watering plants.*
watermelon: *large, green fruit with sweet, red flesh.*
weathervane: *device showing wind direction, often decorative.*
webcam: *camera used for video streaming.*
wedding_cake: *decorative cake served at weddings.*
wedding_ring: *circular band worn during marriage.*
wet_suit: *tight-fitting suit for water activities.*
wheel: *circular object for movement.*
wheelchair: *chair with wheels for mobility.*
whipped_cream: *light, fluffy cream used as a topping.*
whistle: *small device that makes a high-pitched sound.*
wig: *artificial hairpiece worn on the head.*
wind_chime: *decorative item that makes sound in the wind.*
windmill: *structure for harnessing wind energy.*
window_box_(for_plants): *container for growing plants outside windows.*
windshield_wiper: *device for clearing rain from windshields.*
windsock: *fabric cone indicating wind direction.*
wine_bottle: *bottle for storing wine.*
wine_bucket: *container for chilling wine bottles.*
wineglass: *stemmed glass for drinking wine.*
blinder_(for_horses): *covers for horses' eyes to reduce distractions.*
wok: *deep, round pan used for stir-frying.*
wolf: *wild canine known for its pack behavior.*
wooden_spoon: *kitchen tool for stirring.*

wreath: *circular decoration, often made of leaves or flowers.*
wrench: *tool for gripping and turning nuts and bolts.*
wristband: *band worn around the wrist, often for events.*
wristlet: *small pouch with a strap for carrying.*
yacht: *luxury boat for recreation.*
yogurt: *creamy dairy product, often flavored.*
yoke_(animal_equipment): *wooden beam for joining animals together.*
zebra: *striped black and white animal.*
zucchini: *long, green vegetable, often used in cooking.*