TIP-I2V: A Million-Scale Real Text and Image Prompt Dataset for Image-to-Video Generation

Supplementary Material

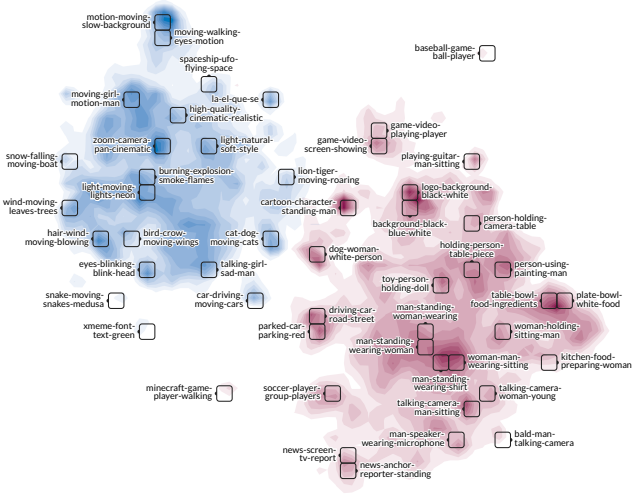

Figure 8. The WIZMAP [65] visualization of our **TIP-I2V** compared to Panda-70M [12]. Please 🔍 zoom in to see the details.

7. Comparing TIP-I2V with Panda-70M

As shown in Fig. 8, we compare the text semantics of our TIP-I2V and Panda-70M [12] using WIZMAP [65] to highlight their differences.

8. Exact Words from Pika's Terms of Service

> "**Your Inputs and Outputs.** You own all Outputs you create with the Service ("Your Outputs"). Notwithstanding the foregoing, nothing herein prevents Mellis or the Service from providing any Outputs to a third party that are the same as, or similar to, Your Outputs, and you hereby agree that such third party is free to use and exploit such Outputs without restriction from or obligation to you. You hereby grant Mellis and other users a license to any of your Inputs and Outputs that you make available to other users on the Service under the Creative Commons Noncommercial 4.0 Attribution International License (as accessible here: https://creativecommons.org/licenses/by-nc/4.0/legalcode)." — *Excerpt from Pika's regulations*

9. Details of Adopted Image-to-Video Models

This section details the image-to-video models utilized in our TIP-I2V and the specifications we choose for each model, as shown in Table 6.

`Pika` Image-to-Video [5] is a commercial AI-driven platform that transforms static images into dynamic video content. Users can sign up using a Discord account to access the platform's services. Currently, the service is free to use; however, generated videos include a Pika Labs watermark and are intended for non-commercial purposes. Additionally, all created clips are publicly shared.

Table 6. The generated video specifications in our TIP-I2V, including frame per second (FPS), duration, and resolution.

Image-to-Video Models	FPS	Duration	Resolution
`Pika` [5]	24	3s	Varied
`Stable Video Diffusion` [8]	7	3.57s	1024×576
`Open-Sora` [73]	24	4.25s	640×360
`I2VGen-XL` [71]	7	2s	1280×704
`CogVideoX-5B` [69]	8	6.13s	720×480

`Stable Video Diffusion` [8] is an open-source generative AI model developed by Stability AI that transforms static images into short video clips (without text guidance). It is available in two versions: one generating 14 frames and another producing 25 frames, both supporting frame rates between 3 and 30 frames per second.

`Open-Sora` [73] is an open-source project developed by HPCAI Tech to democratize efficient video production. In Version 1.2, it supports image-to-video generation for durations from 2s to 15s, resolutions from 144p to 720p, and any aspect ratio, effectively bringing the image to life.

`I2VGen-XL` [71] is an advanced image-to-video synthesis model that generates high-quality videos from static images using a two-stage cascaded diffusion approach. To improve diversity, `I2VGen-XL` was trained on approximately 35 million single-shot text-video pairs and 6 billion text-image pairs. It addresses challenges in video synthesis like semantic accuracy, clarity, and spatio-temporal continuity.

`CogVideoX-5B` Image-to-Video [69] is the latest AI model designed to generate dynamic videos from static images, guided by textual prompts. It is developed by the Knowledge Engineering Group at Tsinghua University and has 5 billion parameters.

10. Details of Calculating User Preference

Calculate the most popular subjects: **(1)** for each data point, embed the subject using `SentenceTransformer` [48] to obtain a 384-dimensional vector; **(2)** cluster the resulting $1,701,935$ vectors using `HDBSCAN` [36], which automatically generates $21,247$ clusters; and **(3)** for each cluster, use the most frequent subject as the representative and then rank the obtained subjects by frequency. **Note that** we adopt this approach because `GPT-4o` [42] may use slightly

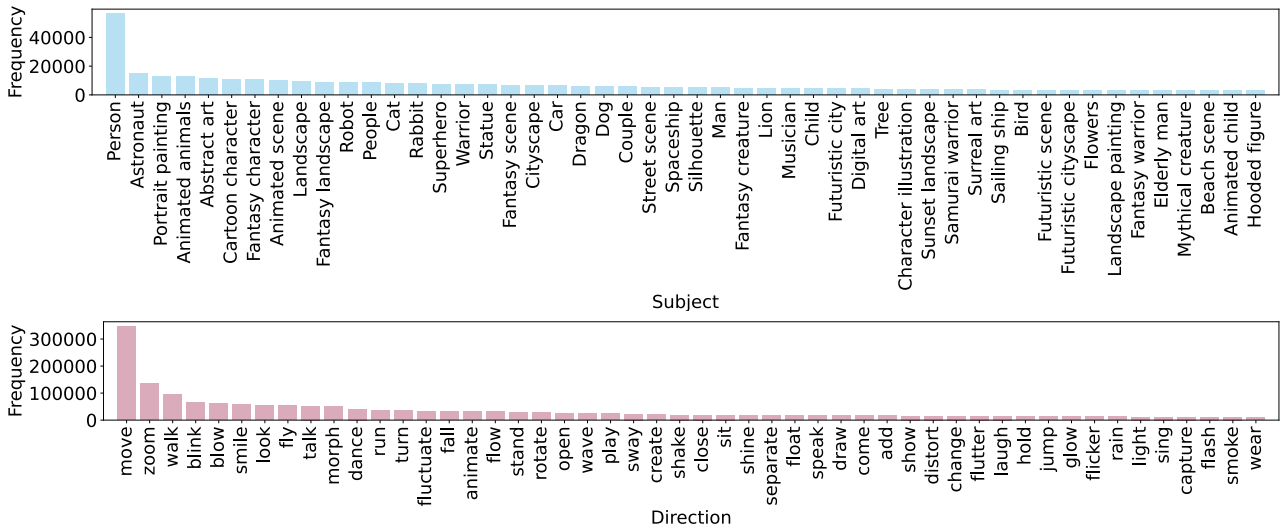

Figure 9. An extension of Fig. 4: the top 50 *subjects* (top) and *directions* (bottom) preferred by users when generating videos from images.

different variations for the same subject. For example, for the subject *'Dragon'*, `GPT-4o` [42] may output *'Dragon'*, *'Dragons'*, *'Dragon, creature'* or *'Dragon creature'*.

Calculate the most popular directions: (1) use `GPT-4o` [42] to extract each verb from the text prompts; **(2)** gather all extracted verbs; and **(3)** rank them by frequency. The used prompt for `GPT-4o` [42] is:

> *"Extract verbs in a given sentence, return their base form, separated by commas, and do not return anything else. If there is no verb, please return ' '. "*

11. Examples for Top Subjects and Directions

As shown in Fig. 10 and Fig. 11, for each of the top 25 most popular subjects and directions, we select one text and one image prompt for illustration. Beyond this, in Fig. 9, we extend Fig. 4 to show the top 50 users' preferred *subjects* (top) and *directions* (bottom).

12. Full Experiments for Benchmarking

Table 7 provides the full experiments for generating the radar chart shown in Fig. 6. For the selected 10 dimensions, '*subject consistency*', '*background consistency*', '*motion smoothness*', '*dynamic degree*', '*aesthetic quality*', and '*imaging quality*' are derived from `VBench-I2V` [25] and `I2V-Bench` [49], while '*temporal consistency*', '*video-text alignment*', '*video-image alignment*', and '*disentangled objective video quality evaluator (DOVER)*' [66] are from `AIGCBench` [18].

13. Details of TIP-ID Dataset

Unlike previous fake image detection datasets, which classify images into **two** classes – real and fake – our TIP-ID dataset emphasizes **three** classes: real videos, videos generated from texts, and videos generated from images.

Table 7. The full experimental results for drawing Fig. 6. Similar to VBench [25], when drawing the radar chart, results are normalized per dimension to a common scale between 0.3 and 0.8 linearly.

Dimension	Pika	SVD	OpS	IXL	Cog
Subject Consistency	0.976	0.950	0.826	0.816	0.949
Background Consistency	0.981	0.959	0.909	0.893	0.962
Motion Smoothness	0.995	0.984	0.992	0.948	0.983
Dynamic Degree	0.058	0.667	0.326	0.775	0.253
Aesthetic Quality	0.659	0.585	0.512	0.555	0.611
Imaging Quality	0.627	0.586	0.514	0.551	0.617
Temporal Consistency	0.997	0.984	0.987	0.953	0.989
Video-text Alignment	0.254	0.252	0.258	0.260	0.255
Video-image Alignment	0.974	0.932	0.767	0.791	0.946
DOVER [66]	0.713	0.607	0.460	0.478	0.652

Sources. (1) Real videos. The real videos are sourced from the VSC22 dataset [46], which comprises approximately $100,000$ videos derived from the YFCC100M dataset [52], ensuring diversity and comprehensiveness. To match the lengths of generated videos, we split the real videos into 3-second segments. This results in $354,486$ real videos totally. **(2) Videos generated from texts.** We randomly select $400,000$ text-generated videos from VidProM [58], with $100,000$ videos from each text-to-video diffusion model: `Pika` [5], `VideoCraft2` [10], `Text2Video-Zero` [28], and `ModelScope` [56]. **(3) Videos generated from images.** We use $500,000$ image-generated videos in our TIP-I2V, with $100,000$ videos from each image-to-video diffusion model. With these sources, the constructed TIP-ID dataset is relatively balanced across each class.

Split. We split the TIP-ID dataset into a 9:1 ratio for training and testing, respectively. It is important to note that: **(1)** When benchmarking existing fake image detection

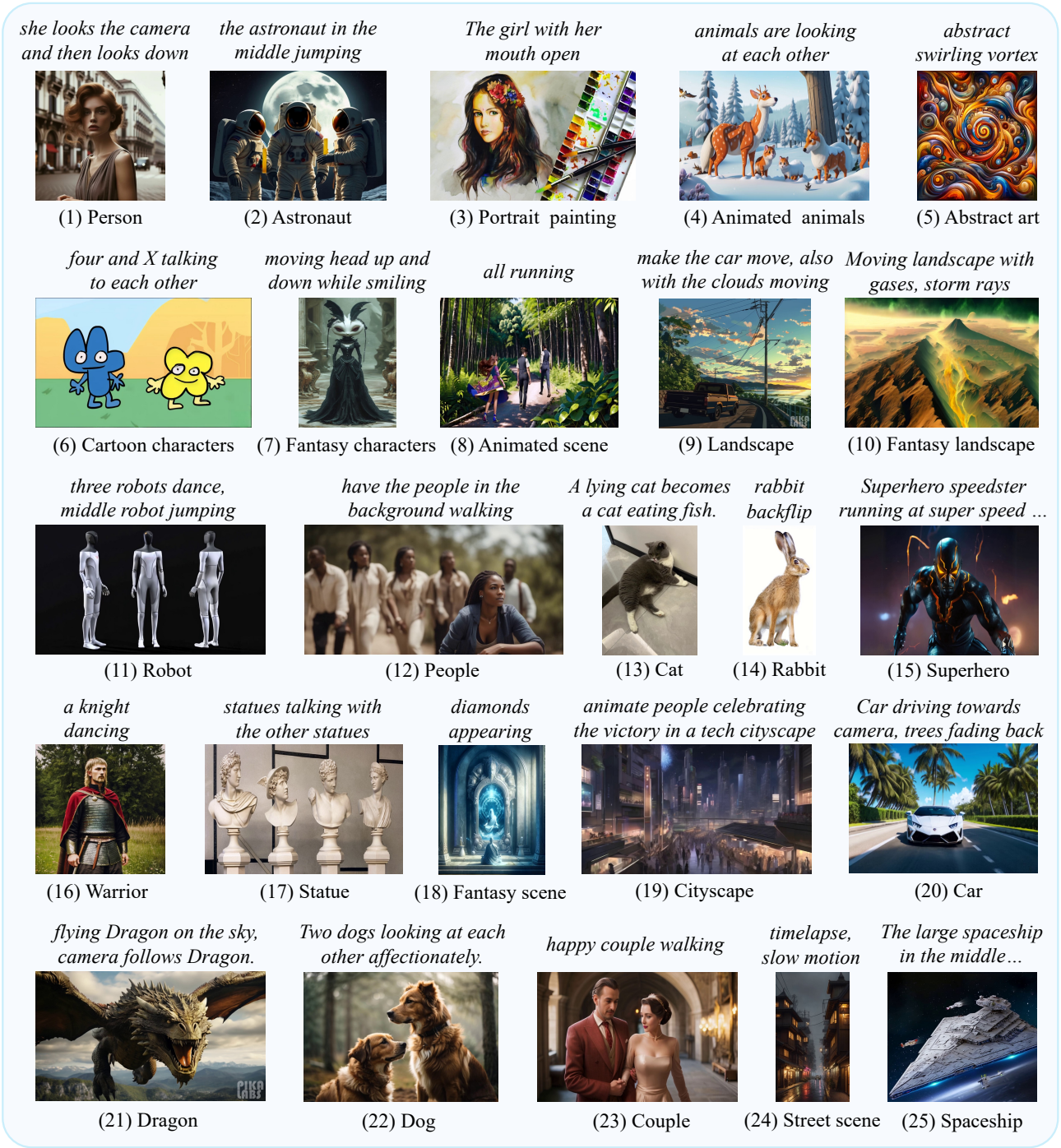

Figure 10. For each top-ranked **subject**, we select one text and one image prompt as examples for illustration.

methods, we exclude the class of text-generated videos, as these methods can only classify videos (images) as real or fake. **(2)** For the training and test sets of image-generated videos, UUIDs do not overlap. This restriction is intended to prevent potential data leakage, as for the same UUID, diffusion models generate videos from the same image. **(3)** Although we split real videos into segments, the segments of any given real video are assigned to either the training set or the test set, but not both. This is also for preventing potential data leakage.

Settings. We consider two settings for evaluating detectors on our TIP-ID dataset. **(1) Same domain.** Both the training and testing image-generated videos are generated by the same diffusion model. For instance, we train and test a detector on videos generated by `Open-Sora`. This setting aims to test whether a detector can achieve high per-

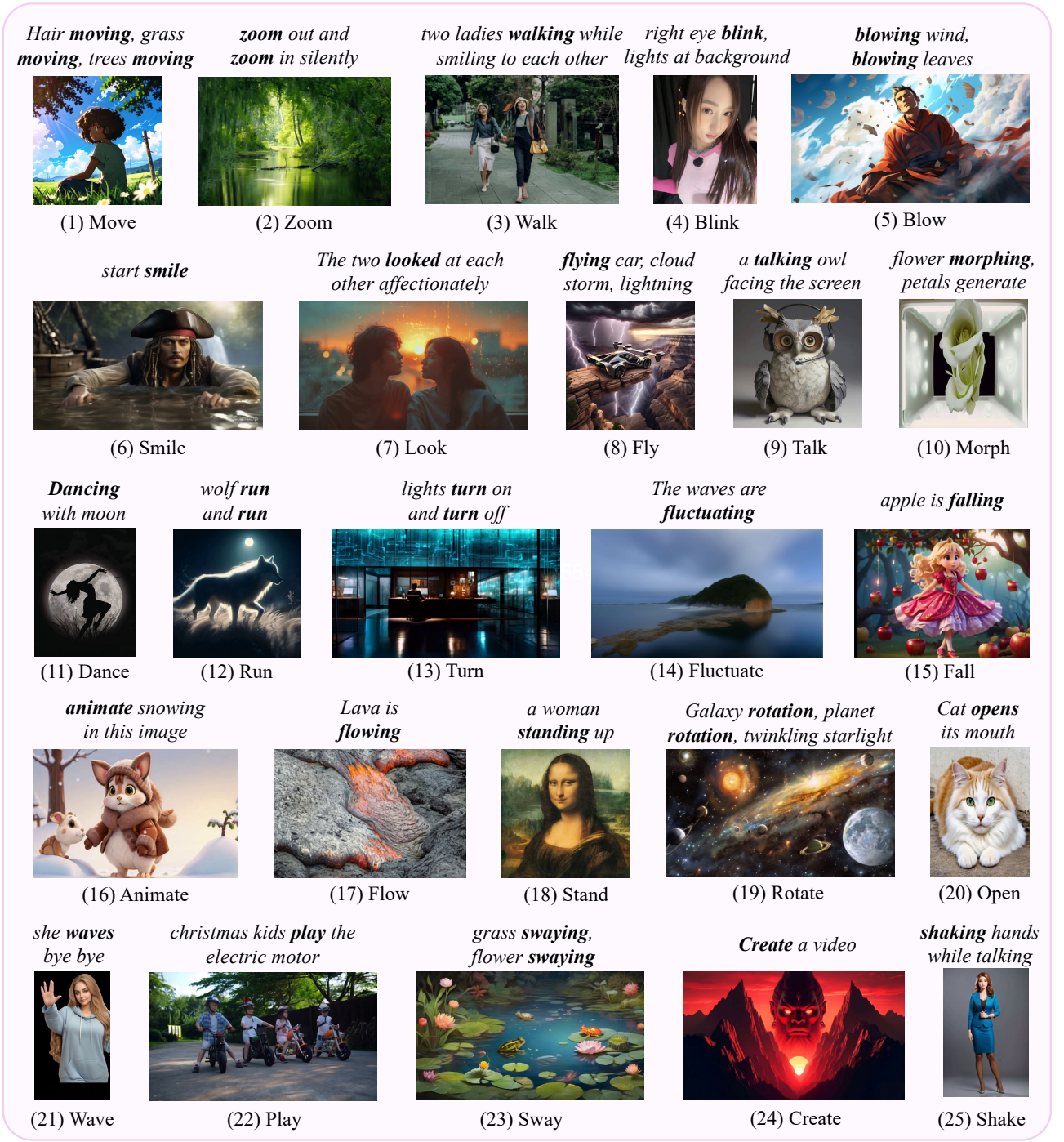

Figure 11. For each top-ranked **direction**, we select one text and one image prompt as examples for illustration.

formance when it has already encountered videos generated by one diffusion model, which can be considered the upper bound for the next setting. **(2) Cross domain.** The training and testing image-generated videos are generated by different diffusion models. For example, we train a detector on videos generated by `Pika`, `Stable Video Diffusion`, `I2VGen-XL`, and `CogVideoX-5B`, but test it on `Open-Sora`. This approach is more practical, as a trained detector will likely encounter newly-developed image-to-video models that it has not previously seen.

Evaluation metrics. Following the fake image detection task, we use Accuracy and Mean Average Precision (mAP) to evaluate the performance of models on the proposed TIP-ID dataset. Specifically, Accuracy measures the proportion of correct predictions among all predictions; whereas mAP evaluates performance for each class, which is useful for

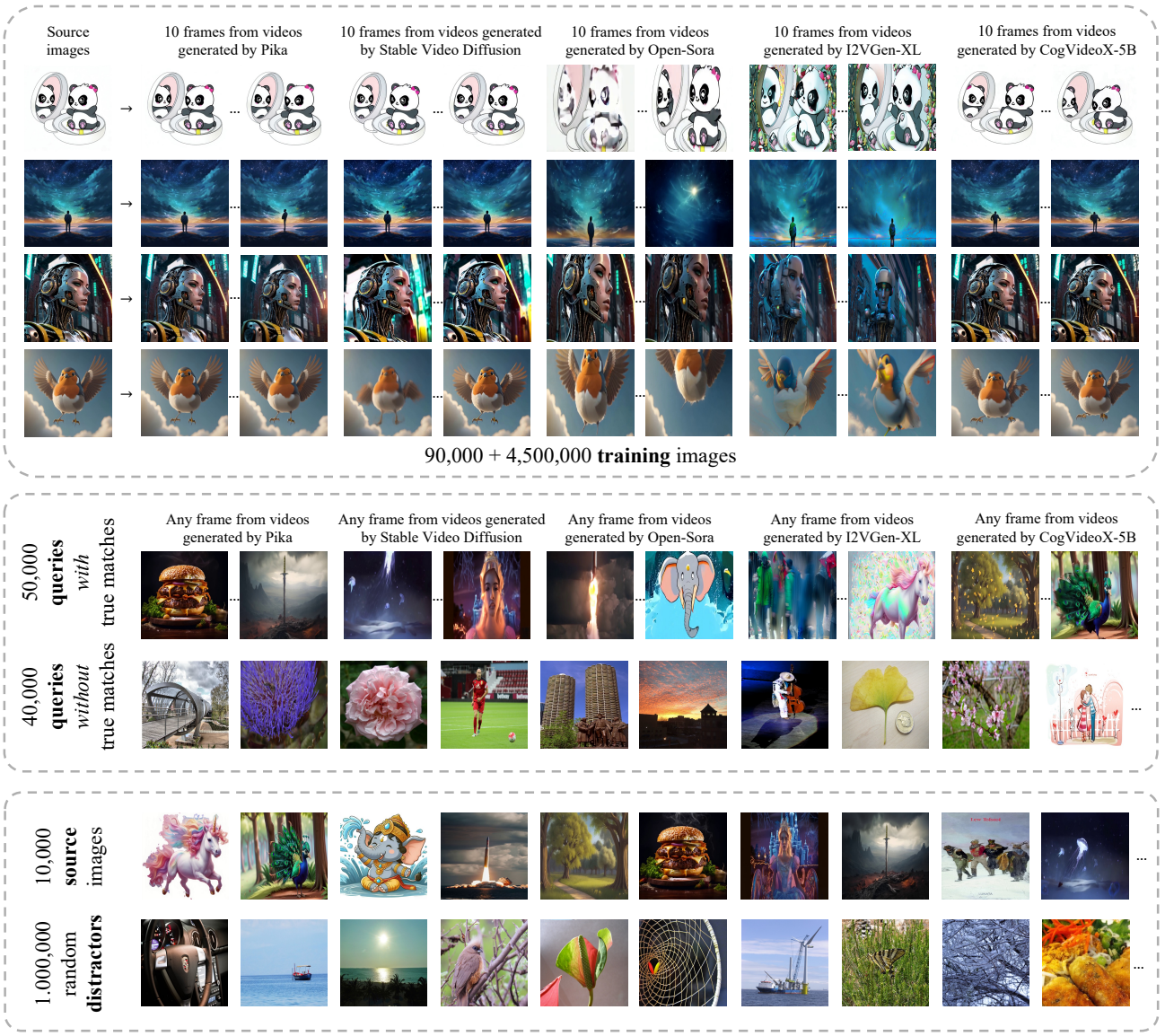

Figure 12. The illustration of the **TIP-Trace**, which is designed to train a model to identify the source image of any given generated frame.

handling class imbalances.

14. Details of Fine-tuning VideoMAE

We fine-tune the Video Masked Autoencoder (VideoMAE) [53] on the TIP-ID dataset for a video classification task. Specifically, our preprocessing pipeline includes (1) temporal subsampling, (2) spatial transformations, and (3) normalization. During training, the spatial transformations consist of random short-side scaling, random cropping to 224×224, and random horizontal flipping; for testing, we only resize frames to 224×224. The pre-trained model is downloaded from the VideoMAE's official Hugging Face repository, and we adjust it to match the number of classes in our dataset, *i.e.*, 3, by updating the classification head. Training is distributed across a server with 8 A100 GPUs. The mini-batch size is 8, the learning rate is 5×10^{-5}, and the number of iterations is $20,000$.

15. Details of TIP-Trace Dataset

As shown in Fig. 12, this section provides a detailed description of the proposed TIP-Trace dataset. It includes *training*, *query*, and *reference* sets:

- **Training set.** Recall that we randomly selected $100,000$ text and image prompts from TIP-I2V to generate videos using state-of-the-art image-to-video models. We use $90,000$ of these text and image prompts to construct the training set. Specifically, for the videos generated by each image-to-video model, we uniformly select 10 frames from each, resulting in a total of $5 \times 90,000 \times 10 = 4,500,000$ training images. Including the image prompts (source im-

ages), we have a total of $90,000+4,500,000 = 4,590,000$ training images, as shown in Fig. 12 (top).

• **Query set.** As shown in Fig. 12 (middle), the query set consists of two parts: **(1) Queries generated from remaining 10,000 prompts.** Instead of uniformly selecting 10 frames from each video, we randomly pick one frame from each video for testing. This results in $50,000$ query images with true matches. **(2) Distractor queries.** The distractor images, *i.e.*, images not extracted from image-to-video generation, serve to replicate real-world scenarios where there is an abundance of authentic images rather than artificially generated ones. We randomly select $40,000$ from Open Images Dataset [30] as the distractor queries.

• **Reference set.** We design the reference set to mimic a "needle-in-a-haystack" scenario in the real world, where the majority of images do not have corresponding queries. Specifically, as shown in Fig. 12 (bottom), we incorporate the $10,000$ source images into a set of $1,000,000$ reference images from DISC21 [44], which is derived from the real-world multimedia dataset YFCC100M [52].

Beyond the split sets, we also introduce two evaluation settings and one evaluation metric to assess model performance on the proposed dataset:

• **Two evaluation settings.** We consider two settings for evaluating model performance on the TIP-Trace dataset. **(1) Same domain.** In this setting, models can be trained and tested on data from all 5 image-to-video models. This setting is used to evaluate whether a model can learn discriminative information after training. **(2) Cross domain.** We observe that, in the real world, new image-to-video models continually emerge. Therefore, in this setting, we assess whether a trained model can generalize to unseen models. Specifically, we exclude one of the five models from the training set and conduct testing on the excluded model. This setting is more challenging and practical than the first.

• **An evaluation metric.** This task uses μAP (micro Average Precision) as the evaluation metric. μAP considers the overall performance across all queries by aggregating true positives, false positives, and false negatives over the entire dataset before calculating precision and recall. This evaluation metric is particularly suitable for this task as it provides a more holistic measure of a model's effectiveness in distinguishing between matching and non-matching images in large-scale datasets.

16. Details of Deep Metric Learning Baseline

We first treat each source image and its generated frames together as a single class, then **train** a CosFace [55] on the resulting $90,000$ classes as a strong deep metric learning baseline. The hyperparameters are set as follows: the model architecture is ViT-Base [16], with a CosFace loss margin of 0.35 and a scale parameter of 64. The training process is with a batch size of 512, using 4 instances per class. We use a cosine learning rate schedule with a maximum learning rate of 0.00035 and a warmup period of 5 epochs. The model is trained for 25 epochs, with $2,000$ iterations per epoch, distributed across 8 A100 GPUs. Input images are resized to a height and width of 224×224. When **testing**, we remove the classification layer and use the trained ViT-Base to extract features from queries and references.

17. Potential Social Impact

The TIP-I2V dataset has potential for **positive** social impact by *enhancing digital creativity* and *fostering responsible AI use*. Specifically, by helping the creation of more user-responsive image-to-video models, TIP-I2V enables content creators to create engaging and customized videos. Additionally, TIP-I2V contributes to the development of detection models that help verify authenticity, trace image sources, and prevent harmful content misuse. Nevertheless, the TIP-I2V dataset may also have potential **negative** social impacts if **misused**. Below, we outline several potential negative social impacts and provide solutions:

• **NSFW content.** Although limited in quantity, our dataset includes some NSFW text and image prompts, which may be sensitive or potentially discomforting for certain individuals. Similar to VidProM [58] and DiffusionDB [64], we choose not to remove these NSFW prompts, as they may provide valuable data for AI safety researchers to analyze and develop content-blocking solutions. Nevertheless, we provide NSFW scores for text and image prompts, allowing regular researchers to easily identify and remove these prompts if they find the content uncomfortable.

• **Privacy.** Although, per Pika's regulations, users agree to make their input and output publicly available, some may still feel uncomfortable with the inclusion of their prompts in TIP-I2V. To enhance user privacy, we implement the following measures: (1) each prompt is assigned a new UUID and an anonymous UserID instead of the identifiable original user name; and (2) users have the right to request that their contributions be removed from TIP-I2V. They can simply email us to make this request.

• **Copyright.** According to Pika Labs' Terms of Service, users are responsible for ensuring that their content does not violate any copyright laws or third-party rights. However, we have noticed that Pika Labs lacks preventive measures, leading some users to upload images, such as "Mickey Mouse", which may be subject to copyright restrictions. Nevertheless, including these images in our TIP-I2V does **not** constitute copyright infringement, as our usage falls under *fair use*. While our dataset is open-sourced under a non-commercial license (CC BY-NC 4.0), some malicious users may still use this dataset for commercial purposes, potentially infringing copyright. Therefore, we strongly recommend that users of TIP-I2V comply with our license to avoid any legal risks.