TRANSPORTER : Transferring Visual Semantics from VLM Manifolds

Alexandros Stergiou
University of Twente, NL
https://alexandrosstergiou.github.io/TRANSPORTER

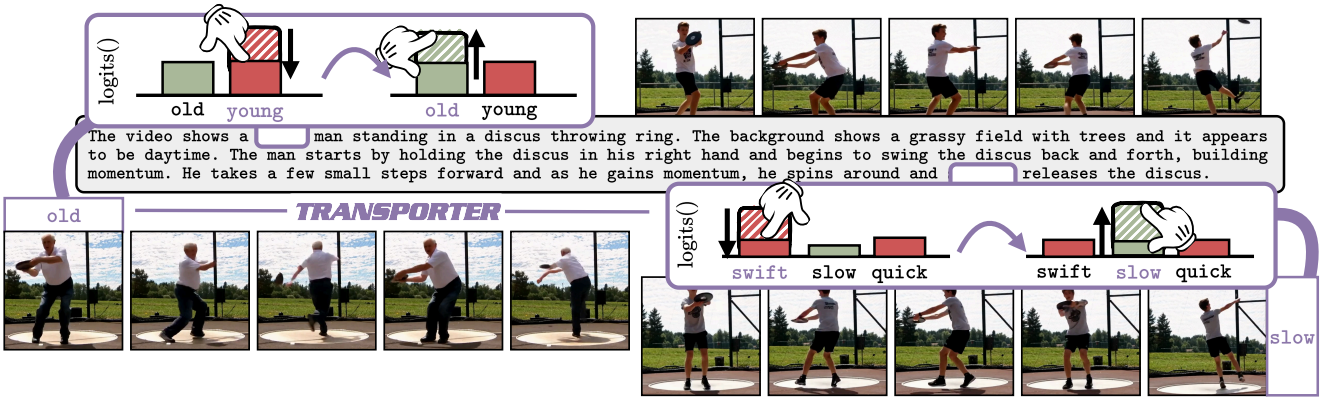


Figure 1. **Generated videos representing VLM logit modulations with TRANSPORTER** . Videos that capture different logit predictions are obtained by coupling VLM embeddings to generative representations. Given the generated VLM caption and a target modulation of objects, actions, or scene attributes, TRANSPORTER guides the video generation process to reflect token logit score changes. Embeddings are decoded into logit-score-aligned videos, as shown by the bottom videos that shift `young` for `old` and `swift` for `slow`.

Abstract

How do video understanding models acquire their answers? Although current Vision Language Models (VLMs) reason over complex scenes with diverse objects, action performances, and scene dynamics, understanding and controlling their internal processes remains an open challenge. Motivated by recent advancements in text-to-video (T2V) generative models, this paper introduces a logits-to-video (L2V) task alongside a model-independent approach, TRANSPORTER , to generate videos that capture the underlying rules behind VLMs' predictions. Given the high-visual-fidelity produced by T2V models, TRANSPORTER learns an optimal transport coupling to VLM's high-semantic embedding spaces. In turn, logit scores define embedding directions for conditional video generation. TRANSPORTER generates videos that reflect caption changes over diverse object attributes, action adverbs, and scene context. Quantitative and qualitative evaluations across VLMs demonstrate that L2V can provide a fidelity-rich, novel direction for model interpretability that has not been previously explored.

1. Introduction

The world is full of rich visual signals. For example, a discus throw, as in Fig. 1, can include variations in the scene dynamics, people or object attributes, and action performances. Video understanding has experienced drastic growth with detecting, predicting, captioning, grounding, and reasoning actions from encodings [83]. This convergence from visual attributes to context-rich semantics has led to Vision-Language Models (VLMs) [1, 17, 87, 109] that address tasks in tandem. Despite significant progress, uncovering explanations behind model decisions has been a longstanding challenge. Existing approaches prompt [100], linearly probe [42], or decode hidden embeddings [15, 72] to gain text-based descriptions that are often sensitive to changes [4], of limited length, or misrepresent internal processes [93]. To address this gap, this paper revisits visual explanations and presents a logits-to-video (L2V) generative task for synthesizing videos corresponding to logit distributions. As VLMs provide answers as probability distributions over a token dictionary, transporting logits to videos can directly and causally represent relevant insights.

To model L2V, a VLM-independent approach is introduced. TRANSPORTER creates probabilistic paths

between high-semantic and visually-fidelitous representations, allowing for fine-grained generation controlled by VLM predictions (logits). Unlike existing methods that rely on text-based explanations or visual-saliency attributions, *TRANSPORTER* visually represents modulations across VLM tokens. During training, video attribute pair modulations are encoded to latent vectors. In inference, their logit divergence is used as a condition for video generation. *TRANSPORTER* provides a novel interactive explainability paradigm that not only verifies the alignment between learned semantics and relevant visual representations, but also explores learned object-attribute correlations.

This work's contributions are: (i) L2V, a novel controlled generation task for VLM visual explanations. (ii) *TRANSPORTER* , a model that learns optimal transport paths between local geometric relationships and global representations across embedding spaces. (iii) Learnable latents to modulate videos by the target logit divergence. (iv) Empirical evaluations and ablations across settings.

2. Related works

Approaches for interpreting deep models can be divided into three groups, detailed below.

Attribution visualizations relate input contributions to model predictions. For image-based models, approaches have focused on the regional saliency of categories [12, 29, 74], object parts [7], or gradients [77, 86]. Extensions also propagated local relevance through attention [9], or layer-averaged activations [13], while other efforts visualized attributions in video classification [59, 84]. As recent models represent vision and language in shared embedding spaces [76, 108], approaches have explored text-based attributions with point-wise vision-language mutual information maximization [10, 41], score-based pairwise similarities [6, 36, 71], and learned ensemble models [106]. Prompt-to-image relevance [63] has also been visualized through image-level object highlights [65, 78], `CLS` token decomposition [107], and cluster graphs over latent representations [27, 54]. In contrast, *TRANSPORTER* focuses on visualizing VLM semantics within the challenging video domain by generating attribute-controlled videos.

Representation decomposition methods visualize parts of objects and category relations [28], or decompose activations to vector directions [38]. Vision-language approaches related object semantics to feature [31, 49, 64], or learned latent [16] descriptions. Works have studied text-to-image correspondence based on embedding distances [103, 111], ablated outputs [67], and modality-specific differences [46, 47, 57]. Prototype-based approaches [21, 85, 96, 101] have gained traction with contrastive-learned embeddings [21], semantics-caching-instance-shifting representations [85], and semantic boundary centering [96, 101]. Optimal Transport (OT) has also been used to align visual features to text semantics [14, 113], and cluster features based on activations [98, 112]. This paper explores a different direction. It uses OT to couple high-semantic embeddings with detailed visual representations for video explanations.

Concept-based methods aim to invert model inputs [39, 69, 79, 105]. Most objectives update inputs with Activation Maximization (AM) [79] to increase activations of neurons [68], class scores [69], gradients [39], features [26, 70], or language semantics [48]. AM has also been adapted for convolution [25] and attention [82] video models. Due to parameter sensitivity [34, 82, 105], more recent works instead trained [53] or fine-tuned [56] copies of models for visually interpretable features [53], semantic-alignment [55, 80], or joint vision-language representations [19, 30]. Concept editing has been explored for T2I models [8, 32, 43, 73]. Approaches employed low-ranking adapters [32, 45], prompt-editing [43], and prompt-pair embedding differences [8, 73] for control over the image generation. Drawing inspiration from concept editing in T2I, the proposed method uses a control-based objective to generate videos that visually represent VLM tokens (L2V).

3. Method

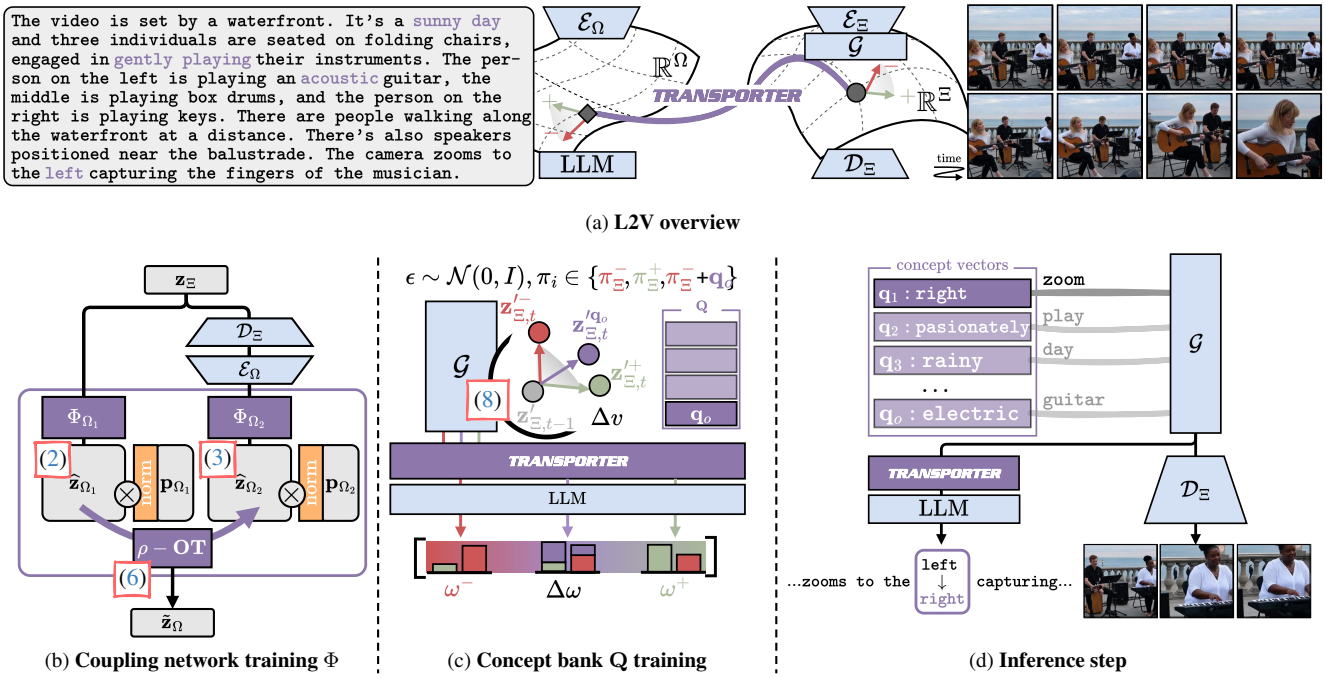
This section provides preliminary definitions for T2V and formulates the L2V objective for explanations (Sec. 3.1); introduces embedding manifold coupling (Sec. 3.2); and the learnable concept bank (Sec. 3.3), which together form *TRANSPORTER* ; and concludes with an overview of a inference pass over all modules (Sec. 3.4).

3.1. Definitions

Preliminaries. Recent T2V models [50, 95] are trained with Conditional Flow Matching (CFM) [3, 60, 61, 91] to model velocity fields between standard Gaussian priors $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and target distributions of (latent, condition) pairs $(\mathbf{z}_\Xi, \pi_\Xi)$, where $\mathbf{z}_\Xi \in \mathbb{R}^\Xi$ are $\mathbf{N}$-token embeddings of video $\mathbf{x}$ and $\pi_\Xi = \mathcal{T}_\Xi(\pi)$ is a tokenized text condition π. Probability paths $\mathbf{z}'_{\Xi,t}$ for time $t \in [0, 1]$ are commonly constructed with a linear interpolation $\mathbf{z}'_{\Xi,t} = t\mathbf{z}_\Xi + (1-t)\epsilon$ and conditional velocity field $v(\mathbf{z}'_{\Xi,t}|\pi_\Xi) \triangleq \mathbf{z}_\Xi - \epsilon$. Network $\mathcal{G}$ is trained to regress the conditional velocity across sampled paths with Mean Squared Error (MSE):

$$\mathcal{L}_{CFM} = \mathbb{E}_{\epsilon,t,(\mathbf{z}_\Xi,\pi_\Xi)} \| v(\mathbf{z}'_{\Xi,t}|\pi_\Xi) - \mathcal{G}(\mathbf{z}'_{\Xi,t}, \pi_\Xi, t) \|^2 \quad (1)$$

L2V, shown in Fig. 2a, considers VLM predicted logits $\omega = \texttt{logit}(\pi_\Omega)$ for caption $\pi_\Omega = \text{LLM}(\beta, \mathbf{z}_\Omega)$ from text query β, and video encodings $\mathbf{z}_\Omega = \mathcal{E}_\Omega(\mathbf{x}) \in \mathbb{R}^\Omega$ with $\mathbf{N}$ tokens. Captions include an arbitrary number of subjects and attributes. L2V aims to generate video $\mathbf{x}'$ with $\Delta\omega$ for the change between token logits ω^- from π^-_Ω to ω^+ from π^+_Ω. For example, given ω^- logit for `happy`, the goal is to generate video $\mathbf{x}'$ to instead match ω^+ logit for `sad`.



(a) **L2V overview**

(b) **Coupling network training Φ**

(c) **Concept bank Q training**

(d) **Inference step**

Figure 2. (a) **L2V with *TRANSPORTER*** : Embeddings $\mathbf{z}_\Xi \in \mathbb{R}^\Xi$ are **coupled with network** Φ and **concept bank** $\mathbf{Q}$. (b) **Coupling network** Φ initially projects $\mathbf{z}_\Xi$ with condition π_Ξ to $\widehat{\mathbf{z}}_{\Omega_1} = \Phi_{\Omega_1}(\mathbf{z}_\Xi, \pi_\Xi)$. Latents $\widehat{\mathbf{z}}_{\Omega_2} \in \mathbb{R}^\Omega$ are obtained with Φ_{Ω_2} over decoder $\mathcal{D}_\Xi$ and encoder $\mathcal{E}_\Omega$ latents. The Learnable Optimal Transport (ρ-OT) module uses updatable projection vectors $\mathbf{p}_{\Omega_1}, \mathbf{p}_{\Omega_2}$ to transport embeddings to $\tilde{\mathbf{z}}_\Omega$. The divergence between pairs π^-, π^+ is used to train the (c) **Concept bank** $\mathbf{Q} = \{\mathbf{q}_o : o \in \mathcal{O}\}$ given the probability path difference Δv between conditions, weighted by the LLM logit distributions change $\Delta\omega$ for ω^- to ω^+. For each (d) **Inference step**, latents $\mathbf{q}_o$ are added to conditions to transport noise latent $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and generate videos and captions.

3.2. Coupling visual fidelity to semantics

Core to the paper's proposal is the coupling of latent representations between the generator's $\mathbb{R}^\Xi$ and VLM's $\mathbb{R}^\Omega$ spaces. As embeddings $\mathbf{z}_\Xi$ can be decoded back to video by the generator's (variational) decoder $\mathcal{D}_\Xi$, a natural choice for L2V would be to decode-and-re-encode $\mathbf{z}_{\Xi\to\Omega} = \mathcal{E}_\Omega(\mathcal{D}_\Xi(\mathbf{z}_\Xi))$. Although the reconstructed embeddings are within $\mathbb{R}^\Omega$'s manifold, the VAE decoder introduces stochastic variation [51]. The resulting token dynamics differ from deterministically encoded encodings $\mathbf{z}_\Xi, \mathbf{z}_\Omega$. Thus, as overviewed in Fig. 2b, a coupling network Φ is used to project and transfer $\mathbf{z}_\Xi$ to $\tilde{\mathbf{z}}_\Omega \approx \mathbf{z}_\Omega$ approximating the representations within $\mathbb{R}^\Omega$.

Projecting to target space. Embeddings $\mathbf{z}_\Xi$ are first projected to $\widehat{\mathbf{z}}_{\Omega_1} = \Phi_{\Omega_1}(\mathbf{z}_\Xi, \pi_\Xi) \in \mathbb{R}^\Omega$ by Φ_{Ω_1}. The module optimizes a MSE objective (2) between obtained projections $\widehat{\mathbf{z}}_{\Omega_1}$ and target encodings $\mathbf{z}_\Omega$.

Attending token structure. As the local geometric relationships between tokens are not directly addressed by (2), $\mathbf{z}_{\Xi\to\Omega}$ are used to learn soft targets $\widehat{\mathbf{z}}_{\Omega_2} = \Phi_{\Omega_2}(\mathbf{z}_{\Xi\to\Omega}) \in \mathbb{R}^\Omega$. Module Φ_{Ω_2} optimizes a Gram-matrix loss [33, 58] to match $\mathbf{z}_\Omega$'s relational token structure (3).

$$\mathcal{L}_{\Phi_{\Omega_1}} = \|\mathbf{z}_\Omega - \widehat{\mathbf{z}}_{\Omega_1}\|^2 \quad (2) \qquad \mathcal{L}_{\Phi_{\Omega_2}} = \|\mathcal{H}(\mathbf{z}_\Omega) - \mathcal{H}(\widehat{\mathbf{z}}_{\Omega_2})\|^2 \quad (3)$$

where $\mathcal{H}(\mathbf{z}_\Omega), \mathcal{H}(\widehat{\mathbf{z}}_{\Omega_2})$ are the $\mathbb{R}^{|\mathbf{N}|\times|\mathbf{N}|}$ Gram matrices computed as the inner product between tokens from $\mathbf{z}_\Omega, \widehat{\mathbf{z}}_{\Omega_2}$ respectively to match token structures within embeddings.

Learnable OT. Given globally projected $\widehat{\mathbf{z}}_{\Omega_1}$ and local structure-preserving $\widehat{\mathbf{z}}_{\Omega_2}$, their complementary representation properties are combined to distilled embeddings $\tilde{\mathbf{z}}_\Omega$. Following recent works on distribution matching over varying-size embedding spaces [75, 88], a novel learnable entropic-based OT (ρ-OT) module is defined. Given $\widehat{\mathbf{z}}_{\Omega_1}$ and $\widehat{\mathbf{z}}_{\Omega_2}$, ρ-OT uses $\{\mathbf{p}_{\Omega_1,\rho}\}_{\rho=1}^P$ and $\{\mathbf{p}_{\Omega_2,\rho}\}_{\rho=1}^P$ sets of P learnable projection vectors. Each $\mathbf{p}_{\cdot,\rho} \in \mathbb{R}^\Omega$ projects tokens $i, j \in \mathbf{N}$ onto scalars for each projection $\rho \in \{1, \dots, P\}$:

$$\mathbf{a}_{i,\rho} = \langle \widehat{\mathbf{z}}_{\Omega_1,i}, \mathbf{p}_{\Omega_1,\rho} \rangle \forall i \in \mathbf{N} \quad (4)$$

$$\mathbf{b}_{j,\rho} = \langle \widehat{\mathbf{z}}_{\Omega_2,j}, \mathbf{p}_{\Omega_2,\rho} \rangle \forall j \in \mathbf{N} \quad (5)$$

For each ρ, transport plan $\tilde{\mathbf{T}}$ is found by minimizing transport cost $\mathbf{M}_{i,j,\rho} = \|\mathbf{a}_{i,\rho} - \mathbf{b}_{j,\rho}\|_2$, with entropic regularization controlled by temperature τ. The full continuous and discrete formulations are expanded in §6. As solving the *full* doubly-constrained problem is computationally intensive, an efficient, closed-form approximation with *partial* double-stochastic iterations is used to regress P-averaged optimal transport plans from $\mathbf{T}_{i,j,\rho} \propto \exp(-\mathbf{M}_{i,j,\rho}/\tau)$. The resulting $\tilde{\mathbf{T}} \in \mathbb{R}^{|\mathbf{N}|\times|\mathbf{N}|}$ can be applied to obtain $\tilde{\mathbf{z}}_\Omega = \tilde{\mathbf{T}}\widehat{\mathbf{z}}_{\Omega_1}$.

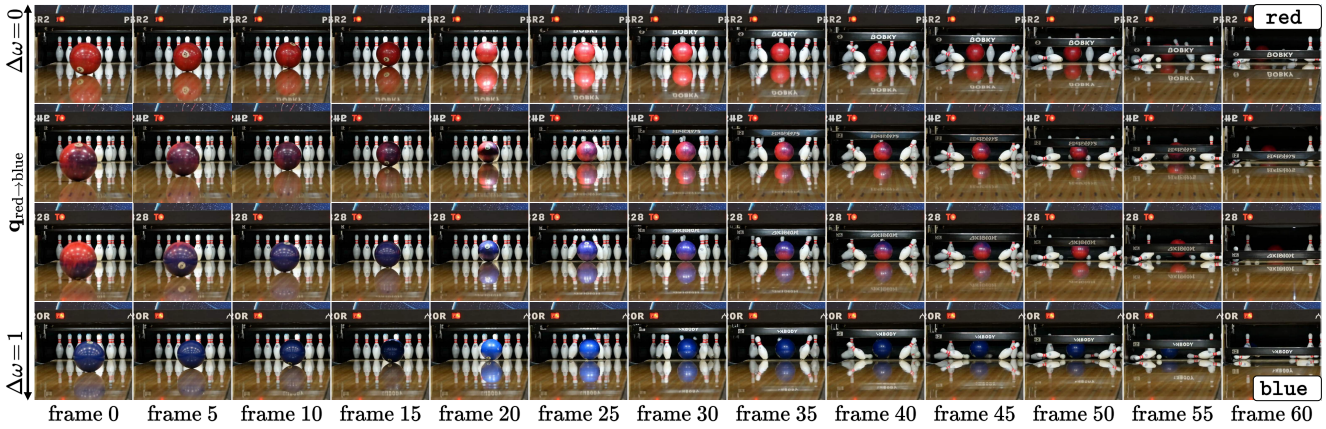


Figure 3. **Concept attribute control with *TRANSPORTER*** given the caption: `A close up shot of a [ ] bowling ball hitting the pins in a bowling alley`. Initially, `[red]` is used to obtain generator/VLM encodings π^-_Ξ, π^-_Ω. Vector $\mathbf{q}_{\text{red}\to\text{blue}}$ is added to π^-_Ξ based on divergence $\Delta\omega$ to π^+=`[blue]`. As shown, the generated videos are of high visual fidelity while also preserving scene dynamics across modulations; *e.g.*, camera view, object motions and their affordances over time, as well as time-relevant interactions. Changes are only seen specifically for the target attribute. Vector $\mathbf{q}_{\text{red}\to\text{blue}}$ used is learned with $\Delta\omega$ from Gemma 3 [87] logits.

The projection vectors $\mathbf{p}_{\Omega_1}, \mathbf{p}_{\Omega_2}$ are updated based on a joint (2) and (3) objective:

$$\mathcal{L}_{\rho\text{-OT}} = \|\mathbf{z}_\Omega - \tilde{\mathbf{z}}_\Omega\|^2 + \|\mathcal{H}(\mathbf{z}_\Omega) - \mathcal{H}(\tilde{\mathbf{z}}_\Omega)\|^2 \quad (6)$$

A detailed algorithmic formulation of the process is available in §7 of the supplementary material.

3.3. Concept bank learning

The goal of L2V is to generate embeddings $\mathbf{z}'_\Xi$ that capture VLM logit score changes. This requires identifying latent directions. Given pairs of semantically contrastive VLM tokens, *e.g.*, π^- =`[baseball hit]`, π^+=`[baseball miss]`, a learnable concept bank $\mathbf{Q} = \{\mathbf{q}_o : o \in \mathcal{O}\}$ consisting of $|\mathcal{O}|$ concepts is created (Fig. 2c). Each concept vector $\mathbf{q}_o \in \mathbb{R}^\Xi$, encodes the latent direction of pair π^-, π^+, in turn tokenized to $\pi^-_\Xi, \pi^+_\Xi \in \mathbb{R}^\Xi$. In addition to video embeddings, the condition pair is passed to coupling network Φ (Fig. 2b) so attribute modulations can be represented across both $\mathbb{R}^\Xi$ and $\mathbb{R}^\Omega$ spaces.

Generator modulations. Initially, a video caption containing concept π^- is tokenized to $\pi^-_\Xi = \mathcal{T}_\Xi(\pi^-)$. Starting from a Gaussian noise latent prior $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, generator $\mathcal{G}$ is used to approximate denoised embedding $\mathbf{z}'^-_\Xi$ given condition π^-_Ξ over $\bar{t}$ steps. At step $t-1$, the probability path is approximated as:

$$\mathbf{z}'_{\Xi,t} = \mathbf{z}'_{\Xi,t-1} + (1/\bar{t})\mathcal{G}(\mathbf{z}'_{\Xi,t-1}, \pi^-_\Xi, t-1) \quad (7)$$

where $\mathbf{z}'_{\Xi,0} = \epsilon$ at timestep $t=0$ and $\mathbf{z}'_{\Xi,\bar{t}} \approx \mathbf{z}'^-_\Xi$ at $t=\bar{t}$.

At next step t, condition π^-_Ξ can be instead adjusted to include one or multiple subject(s)-specific attribute changes π^+_Ξ. The different velocity fields can be predicted for the adjusted conditions as shown in Fig. 4. The divergence between the two predicted probability paths, $\Delta v =$

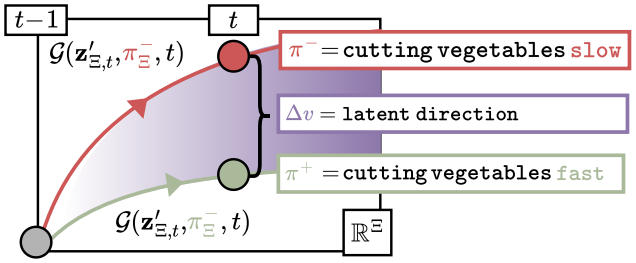


Figure 4. **Flow path modulation**. Given latents $\mathbf{z}'_{\Xi,t}$ two velocity fields are predicted for conditions π^-, π^+ at step t. Their latent divergence Δv corresponds to concept/attribute directions.

$\mathcal{G}(\mathbf{z}'_{\Xi,t}, \pi^+_\Xi, t) - \mathcal{G}(\mathbf{z}'_{\Xi,t}, \pi^-_\Xi, t)$ corresponds to the difference across the target pairs of concept/attribute modulations. Comparably to finding latent directions in diffusion models [8], the CFM loss (1) can be adjusted to include concept vector $\pi^-_\Xi + \delta\mathbf{q}_o$:

$$\mathcal{L}_{\mathbf{q}_o} = \mathbb{E}_{\epsilon,t,(\mathbf{z}'^-_\Xi, \pi^-_\Xi)} \|\text{sg}\big(\mathcal{G}(\mathbf{z}'_{\Xi,t}, \pi^-_\Xi, t) + \delta\Delta v\big) - \mathcal{G}(\mathbf{z}'_{\Xi,t}, \pi^-_\Xi + \delta\mathbf{q}_o, t)\|^2 \quad (8)$$

with δ control parameter and sg stop-gradient. For both π^-_Ξ and $\pi^-_\Xi + \delta\mathbf{q}_o$, generator $\mathcal{G}$ weights remain frozen.

VLM modulations. Path divergence Δv is supervised by VLM logit changes to impose semantic control onto the video generation. Latents $\mathbf{z}'^-_\Xi, \mathbf{z}'^+_\Xi$ obtained from (7), are transported to $\tilde{\mathbf{z}}'^-_\Omega = \Phi(\mathbf{z}'^-_\Xi, \pi^-_\Xi)$ and $\tilde{\mathbf{z}}'^+_\Omega = \Phi(\mathbf{z}'^+_\Xi, \pi^+_\Xi)$ within VLM's $\mathbb{R}^\Omega$ space. Respectively, logits $\omega^- = \texttt{logit}(\pi^-_\Omega)$ and $\omega^+ = \texttt{logit}(\pi^+_\Omega))$ are obtained from $\pi^-_\Omega = \text{LLM}(\beta, \tilde{\mathbf{z}}'^-_\Omega)$ and $\pi^+_\Omega = \text{LLM}(\beta, \tilde{\mathbf{z}}'^+_\Omega)$ respectively. The logit difference $\Delta\omega$ is computed from their Hellinger distance $\Delta\omega = \frac{1}{\sqrt{2}}\left|\sqrt{\omega^-} - \sqrt{\omega^+}\right|$ where $\Delta\omega \in [0, 1]$. The

Table 1. **Quantitative comparisons on VideoLLaMA 3, Gemma 3, Phi 4 MM**. FVD and LPIPS^{v} evaluate the visual quality of generated videos in comparison to videos from VidChapters7M [102]. CLIP^{v} score, aes, and Δ evaluate condition alignment. AM-optimized approaches are adjusted to both video data and VLMs by maximizing either ω^{+} or ω^{-}+$\Delta\omega$. Top performances are in **bold**.

(a) VideoLLaMA 3

Method	Opt	FVD↓	LPIPS^{v}↓	CLIP^{v} score↑	CLIP^{v} aes.↑	CLIP^{v} Δ↑
— Feature vis. maximizing ω^{+} —						
AM^{v} [79]	AM	5.18e^{3}	6.83	9.48	1.36	1.06
GradViT^{v} [39]	AM	2.61e^{3}	5.46	10.11	1.55	1.73
MACO^{v} [26]	AM	2.50e^{3}	4.52	14.23	2.27	1.56
LEAPS [82]	AM	1.85e^{3}	4.37	16.74	2.56	2.28
— Feature vis. maximizing varying $\omega^{+}+\Delta\omega$ —						
MACO^{v} [26]	AM	3.88e^{3}	5.85	11.54	1.91	1.78
LEAPS [82]	AM	2.39e^{3}	5.12	11.95	2.13	2.30
TRANSPORTER	L2V	**1.25e^{2}**	**1.67**	**35.44**	**4.28**	**12.62**

(b) Gemma 3

Method	FVD↓	LPIPS^{v}↓	CLIP^{v} score↑	CLIP^{v} aes.↑	CLIP^{v} Δ↑
— Feature vis. maximizing ω^{+} —					
Baseline [82]	2.18e^{3}	4.55	17.41	2.67	1.82
— Feature vis. maximizing varying $\omega^{+}+\Delta\omega$ —					
Baseline [82]	2.42e^{3}	4.72	14.13	2.05	1.89
TRANSPORTER	**1.05e^{2}**	**1.43**	**36.18**	**4.21**	**11.56**

(c) Phi 4 MM

Method	FVD↓	LPIPS^{v}↓	CLIP^{v}↑ score↑	CLIP^{v}↑ aes.↑	CLIP^{v}↑ Δ↑
— Feature vis. maximizing ω^{+} —					
Baseline [82]	2.34e^{3}	5.12	15.06	2.24	1.73
— Feature vis. maximizing varying $\omega^{+}+\Delta\omega$ —					
Baseline [82]	2.58e^{3}	5.43	10.45	1.87	1.54
TRANSPORTER	**1.42e^{2}**	**1.54**	**35.71**	**4.18**	**11.35**

obtained divergence is used in (8) as the control parameter $\delta = \Delta\omega$. In training, $\Delta\omega$ is averaged over multiple noise initializations to reduce variance across representations of the same concept/attribute pairs π^{-}, π^{+}.

3.4. From logits to videos

Inference (Fig. 2d) allows δ to be manually defined to generate videos that reflect specific changes in VLM logit distributions $\Delta\omega$. Using $\mathbf{q}_o$, the approximated probability path (7) is adjusted with condition $\pi^{-}_{\Xi} + \delta\mathbf{q}_o$. Generated latents $\mathbf{z}'^{\mathbf{q}_o}_{\Xi,t}$ are decoded to video $\mathbf{x}'^{\mathbf{q}_o} = \mathcal{D}(\mathbf{z}'^{\mathbf{q}_o}_{\Xi,t})$ as in Fig. 3. Additionally, the coupling also enables generating captions for $\mathbf{x}'^{\mathbf{q}_o}$ by transporting $\mathbf{z}'^{\mathbf{q}_o}_{\Xi,t}$ to $\mathbb{R}^{\Omega}$ via Φ.

4. Experiments

VLMs, training datasets, and *TRANSPORTER* details are described in Sec. 4.1. L2V is quantitatively compared to adjacent feature visualization tasks over video quality and semantic alignment metrics in Sec. 4.2. Qualitative results in Sec. 4.3 are followed by ablations in Sec. 4.4.

4.1. Implementation details

Models and Datasets. Modulation videos are generated for VideoLLaMA 3 (7B) [109], Gemma 3 (12B) [87], and Phi 4 MM (5B) [1] logits. The VLM selection is based on ViT/LLM diversity. Wan2.2 [95] is used as the base generator. The imported models only run inference with frozen parameters. Coupling is trained on an ensemble of semantically-rich, high-resolution, and egocentric videos from VATEX [99], LAVIB [81], and Ego4D [37]. The concept bank is trained on different seeds of target concepts.

[v] In house conversion of image method/metric to video.

***TRANSPORTER* settings**. Training is done in two stages. The coupling network Φ is trained first. Φ_{Ω_1} consists of a 24-layer MLPMixer [90] projector and Φ_{Ω_2} is a 12-layer Transformer [20]. Module ρ-OT uses 100 projections. Coupling is trained with AdamW [62] for 100K iterations with a learning rate of $1e^{-3}$, batch size of 8, and gradients accumulated every 8 steps. Concept bank $\mathbf{Q}$ is trained at a second stage for 1K iterations per vector with $1e^{-4}$ learning rate. Per generative step, ϵ is of $16{\times}90^2$ size then decoded to $61{\times}720^2$ video resolution. Δv and $\Delta\omega$ are averaged over multiple seeds. Further details are available in §8.

4.2. Results

Tab. 1 reports alignment scores between encoded condition text and generated video embeddings, alongside visual quality metrics between generated and real videos. Text-to-video alignment is evaluated with frame-averaged CLIP^{v} scores [44], aesthetic correspondence [40], and cos-sim divergence score [8] between generated $\mathbf{x}'^{\mathbf{q}_o}$ and tokenized target captions π^{+}. For video-to-video visual quality, Fréchet Video Distance (FVD) [94] and frame-averaged Learned Perceptual Image Patch Similarity [110] (LPIPS^{v}) are reported between generated and 1K VidChapters7M [102] video embeddings. As L2V focuses on visual explanations, baselines include prominent methods from image- [26, 39, 79] and video- [82] classification, adjusted to maximize a target logit ω^{+} or alternatively, logits with varying divergence $\omega^{-} + \Delta\omega$. Only one previous method has directly addressed visual explanations for video models [82]. Based on its relevance, it is selected as the baseline method for comparisons across settings.

Baselines. Compared to VLM-adjusted prior works, *TRANSPORTER* yields consistently better semantic align-

Table 2. **Embedding similarity between generated and real videos**. Cosine similarity (cos), $l1$/$l2$ distance, and Kullback–Leibler divergence (KL) metrics are compared between mean encodings from VidChapters7M and generated videos. The respective video encoder per VLM is used for each metric.

Method	Metric			
	$cos \uparrow$	$l1 \downarrow$	$l2 \downarrow$	KL $\downarrow$
VideoLLaMA 3				
Baseline [82]	$6.45e^{-8}$	$2.93e^{+2}$	$2.75e^{+2}$	56.31
TRANSPORTER	$\mathbf{3.28e^{-2}}$	$\mathbf{5.63e^{0}}$	$\mathbf{5.88e^{0}}$	**1.67**
Gemma 3				
Baseline [82]	$2.31e^{-8}$	$4.29e^{+2}$	$5.65e^{+2}$	73.00
TRANSPORTER	$\mathbf{1.58e^{-2}}$	$\mathbf{4.79e^{0}}$	$\mathbf{7.85e^{0}}$	**4.00**
Phi 4 MM				
Baseline [82]	$1.31e^{-8}$	$2.47e^{+2}$	$6.09e^{+2}$	45.39
TRANSPORTER	$\mathbf{1.25e^{-2}}$	$\mathbf{8.37e^{0}}$	$\mathbf{1.21e^{0}}$	**3.66**

ment and visual quality scores. Notably, as shown in Tab. 1a, CLIP$^{\text{v}}$ scores on VideoLLaMA 3 embeddings are improved two-fold with L2V to 35.44, compared to the best AM optimization setting with 16.74. Similarly, significant CLIP$^{\text{v}}$ aesthetic score improvements are observed. The largest gains reported are for generations conditioned on Phi 4 MM logits, as in Tab. 1c, alongside better embedding cossim divergence score Δ. The suitability of L2V for video model explanations is further evident across visual quality metrics. Videos corresponding to Gemma 3 logits, in Tab. 1b, are significantly closer to real videos with a $-2.31e^{3}$ decrease in FVD. Similar trends are also observed for VideoLLaMA 3 logits with LPIPS$^{\text{v}}$ consistently reduced by -5.29 compared to ω^{+} and $\omega^{-} + \Delta\omega$ AM.

Multi-metric results. Tab. 2 compares embeddings of generated and real videos across VLM encoders. As shown, the baseline struggles to approximate the distribution of real videos. AM does not suffice to capture the high embedding complexity, nor can it effectively disentangle the multimodal representations of VLMs. With the refined L2V task for visual explanations, *TRANSPORTER* can generate detailed videos that best correspond to expected token distributions, as shown by the improvements in the cosine similarity and KL divergence scores. In tandem, the videos generated by *TRANSPORTER* maintain high semantic and visual quality, as evidenced by decreases in distance metrics.

4.3. Examples

Qualitative comparisons. Fig. 5 illustrates videos generated with different methods over target token `walk`. The AM baseline cannot generate visually distinct frames, resulting in a video containing only an abstract shape. Instead, L2V produces substantially detailed videos of full scenes that capture target attributes. *TRANSPORTER* is also not limited to representations of individual concepts.

Figure 5. **Preferred input generation with AM (top) and proposed L2V (bottom)** based on VideoLLaMA 3 logits corresponding to `walk`. Beyond visualizing single logits, *TRANSPORTER* further enables generating videos to explore intermediate modulations of the logit distribution when shifting towards `run`.

The inferred videos can visualize details across attributes, such as expected modulation in pace for tokens corresponding to `walk` or `run`. Effectively, *TRANSPORTER* is a generative tool that enables users to visualize *what* VLMs' token predictions correspond to.

Generalizability. Fig. 6 demonstrates *TRANSPORTER*'s applicability to a range of VLMs and its capability to generate videos that maintain scene aesthetics. As shown in the obtained videos, VLMs reason about distinct attributes, such as the types of objects used in actions, *e.g.* juggling, or what a sign reads, throughout entire scenes. However, changes in action performance, *e.g.*, gymnastics spin, seem to be more influential at specific times, without necessarily having long-term effects. This shows that current models learn action sequentiality over different granularities.

Logit divergence. Fig. 7a shows videos generated over multiple attribute modulations across logit divergences. Interestingly, general attributes such as the number of objects, `two` or `one`, exhibit greater variance in the logit divergence, as shown by multiple $\Delta\omega$ values. In comparison, finer details such as object interactions with `thin` or `thick` cuts are only present for a fraction of $\Delta\omega$. This shows that learned semantic correspondences are, in turn, reflected onto token divergences.

Flow timestep selection. The effects of modulations over different timesteps are shown in Fig. 7b. Modulations across large distributions are generated first, as evidenced by the immediate change in the number of peppers. Instead, later generative steps focus on finer details, such as the slicing thickness, which become progressively less visible as modulations are added in subsequent generation steps.

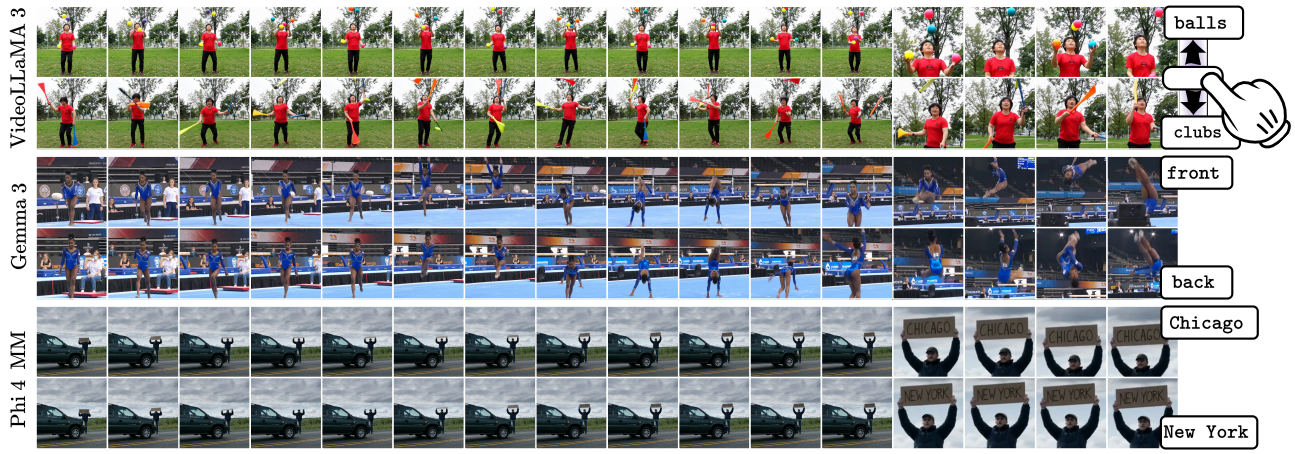
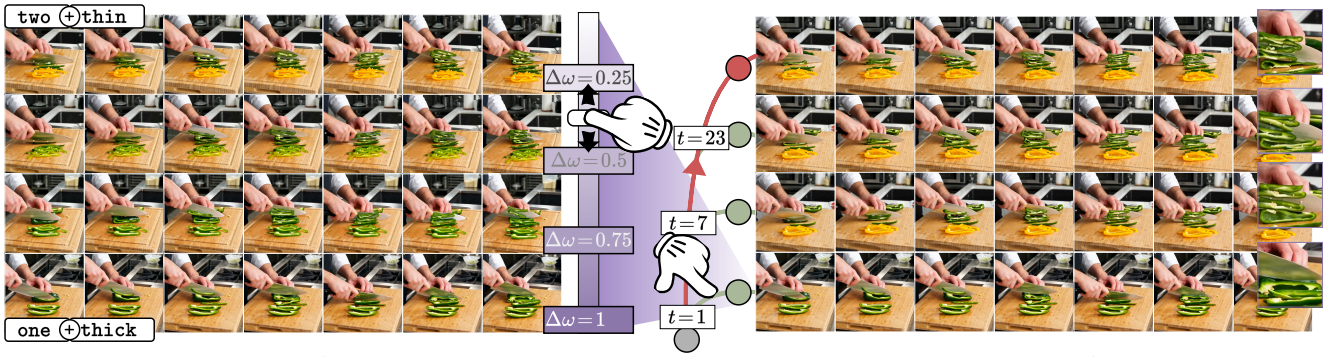
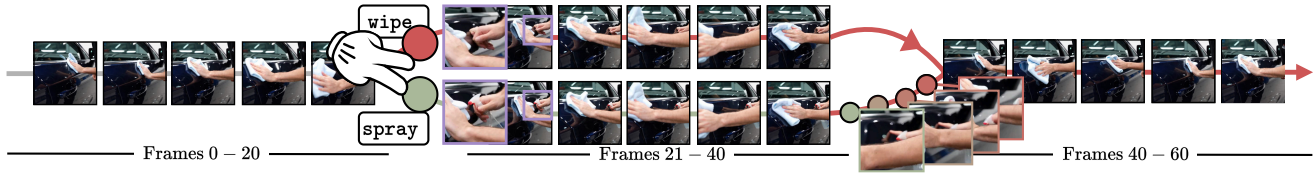


Figure 6. **Generated video modulations with *TRANSPORTER* across VLMs**. Concept vectors can visualize videos corresponding to logit distributions over a variety of video attributes which can relate to (**top**) active objects and affordances, such as juggling `balls` or `clubs`, (**middle**) changes or details in the performance of actions, with `front` and `back` handspring, and (**bottom**) fine-grained scene details, such as holding a sign that reads `Chicago` or `New York`.



(a) Logit divergence $\Delta\omega$ modulation for $t = 1$

(b) Step t modulations for for $\Delta\omega = 1$.



(c) Partial latent modulation.

Figure 7. **Generated videos with Phi 4 MM logits over alternative settings**. (a) **Divergence modulations** can be done over combined attributes, such as cutting `two` peppers over `thin` strips ($\Delta\omega = 0$) to cutting `one` pepper over `thick` strips ($\Delta\omega = 1$). (b) ***TRANSPORTER* modulations can be introduced at different generation steps** to highlight differences of attribute modulations ($\Delta\omega$) with larger divergence, as that of `thin` and `thick`, being better visible within the first few steps of video generation, compared to smaller concept divergences, such as `two` and `one`. (c) **Modulations can also be applied partially to latents**. The resulting videos include both original attributes, e.g. `wipe` a car door, and target changes; e.g. `spray` for their respective frames.

Partial conditioning. Fig. 7c presents videos with only a fraction of their frames modulated during generation. The visualizations demonstrate how and at what intensity target changes are expected to occur within time segments as actions unfold. As shown, a spray bottle blends into the `wipe` to `spray` transition, showing how VLMs bind objects to actions and to adjacent actions.

4.4. Ablations

Ablations are performed for both optimization steps and method settings.

Coupling network settings. To investigate the impact of each coupling network component, results between condition captions and generated LLM captions from *TRANSPORTER* videos are reported at the top of Tab. 3. Notably,

Table 3. **Semantic scores over *TRANSPORTER* ablations** between target caption π^{+} and generated video caption $\pi^{\mathbf{q}_o}$. For each architectural and optimization setting, cosine similarity (∇cos), BLEU (B@1, B@2, B@3, B@4), CIDEr (C), METEOR (M), and SPICE (S) scores are reported between VideoLLaMA 3/Gemma 3 captions from *TRANSPORTER* videos and target captions. Settings are grouped in relation to changes in either the coupling network or the concept bank. Best results are **bold** and second best are underlined.

Method	VideoLLaMA 3								Gemma 3							
	∇cos	B@1	B@2	B@3	B@4	C	M	S	∇cos	B@1	B@2	B@3	B@4	C	M	S
Baseline [82]	0.11	20.14	10.57	5.23	2.05	2.74	11.47	9.12	0.05	20.62	11.18	5.73	2.19	1.95	11.60	7.55
Coupling network ablations																
decode-and-re-encode	0.19	28.86	16.49	10.38	5.24	22.72	14.94	23.45	0.11	29.36	17.80	11.07	8.77	15.36	16.23	25.03
Φ_{Ω_1} only	0.16	26.75	15.79	8.67	4.49	18.17	13.28	18.78	0.09	28.02	15.59	10.39	7.46	14.11	15.92	23.58
Φ_{Ω_2} only	0.21	30.48	18.20	11.73	7.75	25.56	15.09	25.21	0.13	31.69	19.43	13.64	10.57	18.36	20.92	26.57
mean$\langle\Phi_{\Omega_1,\Omega_2}\rangle$	0.23	31.25	19.39	11.99	7.79	26.70	17.16	25.46	0.14	33.21	21.57	15.14	11.21	19.88	21.42	27.20
sGW OT	0.14	24.87	13.43	6.71	4.09	7.92	12.46	11.88	0.08	27.47	13.74	7.56	4.41	7.00	13.87	9.24
P 50	0.24	34.04	20.18	12.36	8.29	27.63	20.12	29.27	0.13	35.45	22.16	15.24	11.64	20.87	22.12	28.89
P 200	**0.26**	34.63	20.64	12.69	8.30	**28.14**	20.34	29.62	0.15	35.84	22.65	15.59	**11.80**	21.30	22.57	29.88
P 400	**0.26**	**34.72**	**20.71**	**12.82**	**8.40**	28.13	**20.41**	**29.65**	**0.16**	**35.93**	22.65	**15.70**	**11.80**	21.31	**22.63**	**29.96**
Concept bank ablations																
$\Delta\omega$ KL	0.19	29.83	17.56	10.45	5.29	24.37	15.72	24.80	0.10	30.56	18.25	11.44	9.69	17.09	19.34	26.52
$\Delta\omega$ JS	0.22	32.48	19.07	11.36	6.58	26.26	18.93	26.75	0.13	33.13	19.88	13.04	10.61	19.78	21.33	27.76
TRANSPORTER	**0.26**	34.59	20.59	12.67	8.34	27.99	20.34	29.59	0.15	35.80	**22.70**	15.65	11.78	**21.33**	22.51	29.85

Figure 8. ***TRANSPORTER* comparison to decode-and-re-encode inference** for Gemma 3 logit divergence between `stone` and `leaves`. Learned modulations with *TRANSPORTER* maintain scene aspects unchanged. This robustness does not translate in inference-only settings.

a significant improvement across all semantic metrics is observed with the coupling modules $\Phi_{\Omega_1}/\Phi_{\Omega_2}$ compared to decoding and re-encoding the generated latents, as in the inference setting. For OT, the heuristic approach based on Sliced Gromov-Wasserstein (sGW) [88] shows a noticeable decrease in caption quality. Increasing the number of projections offers some marginal improvements, but introduces additional computational costs as discussed in §9.

Concept bank divergence metric. The second half of Tab. 3 ablates divergence metrics. As shown, normalizing the bounds of the divergence metrics affects performance. Hillinger distance is chosen given its strict bounds and slightly favorable performance to the Jaccard Similarity (JS). Unbounded metrics, such as the KL divergence, resulted in lower performance.

Inference only. Figure 8 presents qualitative comparisons to decode-and-re-encode setting. *TRANSPORTER* reflects VLM token modulations across different logit differences, which is not possible if only using models at inference. The visualizations further demonstrate *TRANSPORTER* 's ability to maintain the motion and dynamics of the original scene while visually reflecting the target changes across relevant VLM tokens.

5. Conclusion

This paper introduces logits-to-video (L2V). A generative task for video model explainability that goes beyond the adaptation of current classifier-based approaches to video VLMs. Alongside L2V, the paper proposes *TRANSPORTER* , a visual representation method for VLM token predictions. During training, *TRANSPORTER* learns the optimal transport between semantic VLM spaces and the visual spaces of generator models. From this coupling, the divergence between tokens is used to define modulations in the embedding space corresponding to attributes. During inference, *TRANSPORTER* generates videos to represent this divergence. The proposed method has shown qualitatively and quantitatively that the explanations generated are of high visual quality and semantically aligned to target modulations. The video fidelity alongside the high visual and semantic scores make L2V and *TRANSPORTER* a first step towards visualizing VLMs' predictions and a novel direction for understanding their reasoning process.

Acknowledgments. Research was funded by the University of Twente's EEMCS NexAI4Health Internal Grant, and used the Dutch national e-infrastructure supported by the SURF Cooperative with grant EINF-15225.

References

[1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv:2503.01743*, 2025. 1, 5

[2] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *NeurIPS*, 2023. 1

[3] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *ICLR*, 2023. 2

[4] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *NeurIPS*, 2024. 1

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv:2502.13923*, 2025. 1

[6] Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. Decomposing and interpreting image representations via text in vits beyond clip. *NeurIPS*, 2024. 2

[7] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. 2

[8] Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. In *CVPR*, 2025. 2, 4, 5

[9] Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. Legrad: An explainability method for vision transformers via feature formation sensitivity. In *ICCV*, 2025. 2

[10] Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina Höhne. Labeling neural representations with inverse recognition. In *NeurIPS*, 2023. 2

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020. 4

[12] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018. 2

[13] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, 2021. 2

[14] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *ICLR*, 2023. 2

[15] Haozhe Chen, Carl Vondrick, and Chengzhi Mao. Selfie: self-interpretation of large language model embeddings. In *ICML*, 2024. 1

[16] Haozhe Chen, Junfeng Yang, Carl Vondrick, and Chengzhi Mao. Interpreting and controlling vision foundation models via text explanations. *ICLR*, 2024. 2

[17] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*, 2025. 1

[18] Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow models. In *CVPR*, 2025. 4

[19] Antonio D'Orazio, Maria Rosaria Briglia, Donato Crisostomi, Dario Loi, Emanuele Rodolà, and Iacopo Masi. Implicit inversion turns clip into a decoder. *ICLR*, 2026. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 5

[21] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *CVPR*, 2022. 2

[22] Kamath et al. What's up w vision-language models? investigating struggles with spatial reasoning. *EMNLP*, 2023. 2

[23] Parcalabescu et al. Valse: A task-independent benchmark for vision and language models. In *ACL*, 2022. 2

[24] Rahmanzadehgervi et al. Vision language models are blind. In *ACCV*, 2024. 2

[25] Christoph Feichtenhofer, Axel Pinz, Richard P Wildes, and Andrew Zisserman. Deep insights into convolutional networks for video recognition. *IJCV*, 2020. 2

[26] Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Lore Goetschalckx, et al. Unlocking feature visualization for deep network with magnitude constrained optimization. *NeurIPS*, 2023. 2, 5

[27] Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *NeurIPS*, 2023. 2

[28] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *CVPR*, 2023. 2

[29] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017. 2

[30] Stanislav Fort and Jonathan Whitaker. Direct ascent synthesis: Revealing hidden generative capabilities in discriminative models. *arXiv:2502.07753*, 2025. 2

[31] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *ICLR*, 2024. 2

[32] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adap-

tors for precise control in diffusion models. In *ECCV*, 2024. 2

[33] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 3

[34] Amin Ghiasi, Hamid Kazemi, Steven Reich, Chen Zhu, Micah Goldblum, and Tom Goldstein. Plug-in inversion: Model-agnostic inversion for vision with data augmentations. In *ICML*, 2022. 2

[35] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's mergekit: A toolkit for merging large language models. In *EMNLP*, 2024. 4

[36] Shizhan Gong, LEI Haoyu, Qi Dou, and Farzan Farnia. Boosting the visual interpretability of clip via adversarial fine-tuning. In *ICLR*, 2025. 2

[37] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 5

[38] Mara Graziani, Laura O' Mahony, An-Phi Nguyen, Henning Müller, and Vincent Andrearczyk. Uncovering unique concept vectors through latent space decomposition. *TMLR*, 2023. 2

[39] Ali Hatamizadeh, Hongxu Yin, Holger R Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *CVPR*, 2022. 2, 5

[40] Simon Hentschel, Konstantin Kobs, and Andreas Hotho. Clip knows image aesthetics. *FAI*, 2022. 5

[41] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *ICLR*, 2021. 2

[42] Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *COLM*, 2024. 1

[43] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ICLR*, 2023. 2

[44] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 5

[45] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 2

[46] Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities? *arXiv:2507.01790*, 2025. 2

[47] Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *EMNLPS*, 2024. 2

[48] Animesh Jain and Alexandros Stergiou. Mimic: Multimodal inversion for model interpretation and conceptualization. *arXiv:2508.07833*, 2025. 2

[49] Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *ICLR*, 2025. 2

[50] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *ICLR*, 2025. 2

[51] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014. 3

[52] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIMAX*, 2008. 1

[53] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020. 2

[54] Matthew Kowal, Richard P Wildes, and Konstantinos G Derpanis. Visual concept connectome (vcc): Open world concept discovery and their interlayer connections in deep models. In *CVPR*, 2024. 2

[55] Akshay Kulkarni, Ge Yan, Chung-En Sun, Tuomas Oikarinen, and Tsui-Wei Weng. Interpretable generative models through post-hoc concept bottlenecks. In *CVPR*, 2025. 2

[56] Sonia Laguna, Ričards Marcinkevičs, Moritz Vandenhirtz, and Julia Vogt. Beyond concept bottleneck models: How to make black boxes intervenable? *NeurIPS*, 2024. 2

[57] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, 2024. 2

[58] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *IJCAI*, 2017. 3

[59] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *WACV*, 2021. 2

[60] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023. 2

[61] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 2

[62] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 5

[63] Chengzhi Mao, Revant Teotia, Amrutha Sundar, Sachit Menon, Junfeng Yang, Xin Wang, and Carl Vondrick. Doubly right object recognition: A why prompt for visual rationales. In *CVPR*, 2023. 2

[64] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023. 2

[65] Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. In *ICML*, 2023. 2

[66] Nao Nakagawa, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Gromov-wasserstein autoencoders. In *ICLR*, 2023. 2

[67] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *ICLR*, 2025. 2

[68] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *NeurIPS*, 2016. 2

[69] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. In *ICMLw*, 2016. 2

[70] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. 2

[71] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *NeurIPS*, 2025. 2

[72] Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. *CoNNL*, 2023. 1

[73] Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In *ECCV*, 2024. 2

[74] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 2

[75] Moritz Piening and Robert Beinert. A novel sliced fused gromov-wasserstein distance. *AAAI*, 2026. 3

[76] 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*, 2021. 2

[77] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017. 2

[78] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, 2023. 2

[79] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *ICLRw*, 2014. 2, 5

[80] Divyansh Srivastava, Ge Yan, and Lily Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *NeurIPS*, 2024. 2

[81] Alexandros Stergiou. Lavib: A large-scale video interpolation benchmark. In *NeurIPS*, 2024. 5

[82] Alexandros Stergiou and Nikos Deligiannis. Leaping into memories: Space-time deep feature synthesis. In *ICCV*, 2023. 2, 5, 6, 8, 4

[83] Alexandros Stergiou and Ronald Poppe. About time: Advances, challenges, and outlooks of action understanding. *IJCV*, 2025. 1

[84] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Remco Veltkamp, and Ronald Poppe. Saliency tubes: Visual explanations for spatio-temporal convolutions. In *ICIP*, 2019. 2

[85] Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *WACV*, 2025. 2

[86] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 2

[87] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv:2503.19786*, 2025. 1, 4, 5

[88] Vayer Titouan, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced gromov-wasserstein. *NeurIPS*, 2019. 3, 8

[89] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *ICLR*, 2018. 2

[90] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 2021. 5

[91] Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *TMLR*, 2024. 2

[92] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786*, 2025. 1, 4

[93] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *NeurIPS*, 2023. 1

[94] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *ICLRw*, 2019. 5

[95] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv:2503.20314*, 2025. 2, 5

[96] Chong Wang, Yuyuan Liu, Yuanhong Chen, Fengbei Liu, Yu Tian, Davis McCarthy, Helen Frazer, and Gustavo Carneiro. Learning support and trivial prototypes for interpretable image classification. In *ICCV*, 2023. 2

[97] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *CVPR*, 2024. 4

[98] Jian Wang, Tianhong Dai, Bingfeng Zhang, Siyue Yu, Eng Gee Lim, and Jimin Xiao. Pot: Prototypical optimal transport for weakly supervised semantic segmentation. In *CVPR*, 2025. 2

[99] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 5

[100] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 1

[101] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *IJCAI*, 2024. 2

[102] Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vidchapters-7m: Video chapters at scale. *NeurIPS*, 2023. 5

[103] Xingyi Yang and Xinchao Wang. Language model as visual explainer. *NeurIPS*, 2024. 2

[104] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *NeurIPS*, 2024. 4

[105] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, 2020. 2

[106] Weiqiu You, Helen Qu, Marco Gatti, Bhuvnesh Jain, and Eric Wong. Sum-of-parts: Self-attributing neural networks with end-to-end learning of feature groups. In *ICML*, 2025. 2

[107] Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. In *ECCV*, 2024. 2

[108] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 2

[109] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv:2501.13106*, 2025. 1, 5

[110] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[111] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. *ICLR*, 2023. 2

[112] Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Towards interpretable deep metric learning with structural matching. In *ICCV*, 2021. 2

[113] Yuhan Zhu, Yuyang Ji, Zhiyu Zhao, Gangshan Wu, and Limin Wang. Awt: Transferring vision-language models via augmentation, weighting, and transportation. *NeurIPS*, 2024. 2