Contrastive Learning: A Mathematical Perspective

Allowing Image And Text Data To Communicate

Ricardo Baptista

Statistical Sciences
University of Toronto

CMU STAMPS Workshop: Trustworthy Statistical Inference for the Physical Sciences
May 14, 2026

Contrastive Learning: An Unconventional Approach to Inference

Allowing Image And Text Data To Communicate

Ricardo Baptista

Statistical Sciences
University of Toronto

CMU STAMPS Workshop: Trustworthy Statistical Inference for the Physical Sciences
May 14, 2026

Collaborators

A Mathematical Perspective on Contrastive Learning

Ricardo Baptista[1], Andrew Stuart[2,3], Son Tran[3]

[1]Statistical Sciences
University of Toronto

[2]Computing+Mathematical Sciences
California Institute of Technology

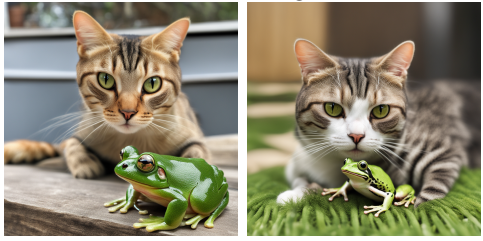
[3]Stores Foundational AI
Amazon

arXiv:2505.24134

Generating Images: Prompt-to-Image Models

- Sample using Stable Diffusion XL: Podell et al. [4] (2023)
- Relies on a model to link text and images

A cat and a frog



A large aircraft taking off

An elephant in the jungle

A skateboarder in California

Interpreting Images: Visual-Language Models

- Vision Question Answering analyzes visual and text data to provide responses: Wang et al. [6] (2024)
- Relies on a model to link text and images

Plant identification

What kind of flower is this?

Model Response

The flowers in the picture are forget-me-nots (Myosotis). These small, delicate flowers are known for their vibrant colors, typically blue, pink, or white, and are often used in bouquets and gardens for their charming appearance.

Multilingual Text Recognition

What is the text?

Model Response

The text on the wall is in Portuguese and reads: “não é normal ter medo de andar sozinha na rua!!!” This translates to: “It is not normal to be afraid to walk alone on the street!!!”

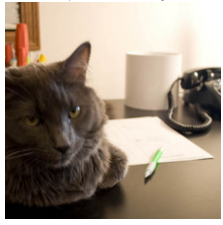
Aligning Text and Images

- Key contrastive learning methodology CLIP: Radford et al. [5] (2021)
- 61028 Google Scholar citations as of May 13th 2026 (44K in October)
- Cosine similarity between $a, b \in \mathbb{S}^{n_e-1}$: $\langle a, b \rangle$
- CLIP represents text as $a \in \mathbb{S}^{n_e-1}$ and image as $b \in \mathbb{S}^{n_e-1}$: then calculate $\langle a, b \rangle$.

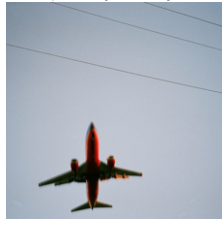
Retrieving Images Given Text MS COCO: Lin et al [3] (2014)

- 71732 Google Scholar citations as of May 13 2026
- CLIP trained on MC COCO $118K$ (image,text) pairs
- Retrieve closest image from $42K$ MS COCO test set given new test prompt

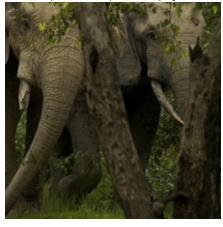
A cat and a frog



A large aircraft taking off



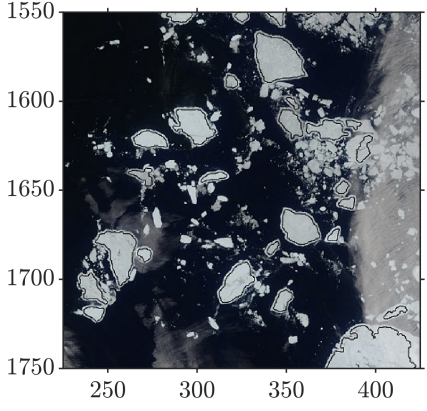
An elephant in the jungle

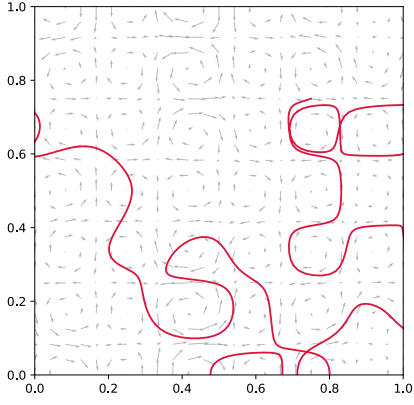


A skateboarder in California

Mapping Tracer Positions to Velocity

- Infer ocean currents conditioned on data from passive tracers (e.g., sea ice floes)
- Relies on a model relating velocity to observations





Chen et al. (2021)

Table of Contents

Table of Contents

Setup for Contrastive Learning

Data is generated from a common reality

Text: $u \in \mathcal{U}$

Images: $v \in \mathcal{V}$

Distribution: $\mu(du, dv)$

Data pairs: $\{(u^i, v^i)\}_{i=1}^N \sim \mu$ i.i.d.

Embed data into a common low-dimensional space

$$g_u : \mathcal{U} \times \Theta \to \mathbb{S}^{n_e - 1}$$
$$g_v : \mathcal{V} \times \Theta \to \mathbb{S}^{n_e - 1}$$

Encoders are represented using data-dependent architectures:

- Text: Byte-pair encoding and transformers
- Images: Convolution layers and transformers
- L^2 normalization to map to the sphere

Contrastive Learning Problem

CLIP Objective Function Radford et al [5] (2021)

Find $\theta = (\theta_u, \theta_v, \tau)$ using only samples from μ

$$\mathsf{L}^N_{\text{clip}}(\theta) := \frac{1}{N}\sum_{i=1}^{N} \langle g_u(u^i;\theta_u), g_v(v^i;\theta_v)\rangle/\tau$$
$$- \frac{1}{2N}\sum_{i=1}^{N} \log\left(\sum_{j=1}^{N} \exp(\langle g_u(u^i;\theta_u), g_v(v^j;\theta_v)\rangle/\tau)\right)$$
$$- \frac{1}{2N}\sum_{j=1}^{N} \log\left(\sum_{i=1}^{N} \exp(\langle g_u(u^i;\theta_u), g_v(v^j;\theta_v)\rangle/\tau)\right)$$
$$\theta^\star = \arg\max_{\theta} \ \mathsf{L}^N_{\text{clip}}(\theta).$$

Objective aligned pair samples and penalizes unaligned pairs

Population-level Picture I B, Stuart and Tran (2025) [1]

Data Notation

Data Measure: $\mu(du, dv)$

Data Marginals: $\mu_u(du), \mu_v(dv)$

Data Conditionals: $\mu_{u|v}(du|v), \mu_{v|u}(dv|u)$

Target Notation

Target Measure: $\nu(du, dv; \theta) = \rho(u, v; \theta)\mu_u(du)\mu_v(dv)$

$$\rho(u, v; \theta) \propto \exp\big(\langle g_u(u; \theta_u), g_v(v; \theta_v)\rangle/\tau\big)$$

Target Conditionals: $\nu_{u|v}(du|v; \theta), \nu_{v|u}(dv|u; \theta)$

Population-level Picture II B, Stuart and Tran (2025) [1]

Objective Function: Population

$$\mathsf{J}(\theta) = \frac{1}{2}\mathbb{E}_{v\sim\mu_v}\left[\mathsf{D}_{\mathsf{kl}}(\mu_{u|v}||\nu_{u|v}(\cdot;\theta))\right] + \frac{1}{2}\mathbb{E}_{u\sim\mu_u}\left[\mathsf{D}_{\mathsf{kl}}(\mu_{v|u}||\nu_{v|u}(\cdot;\theta))\right]$$

$$\theta^* = \arg\min_{\theta} \mathsf{J}(\theta)$$

The non-parametric minimizer is $\nu_{u|v}(\cdot;\theta) = \mu_{u|v}$ and $\nu_{v|u}(\cdot;\theta) = \nu_{v|u}$.

Objective Function: Empirical

$$\begin{aligned}
\mathsf{J}(\theta) &= \text{const} - \mathsf{L}(\theta) \\
\mathsf{L}(\theta) &= \mathbb{E}_{(u,v)\sim\mu}\langle g_u(u;\theta_u), g_v(v;\theta_v)\rangle/\tau \\
&\quad - \frac{1}{2}\mathbb{E}_{v\sim\mu_v}\log\mathbb{E}_{u'\sim\mu_u}\exp\big(\langle g_u(u';\theta_u), g_v(v;\theta_v)\rangle/\tau\big) \\
&\quad - \frac{1}{2}\mathbb{E}_{u\sim\mu_u}\log\mathbb{E}_{v'\sim\mu_v}\exp\big(\langle g_u(u;\theta_u), g_v(v';\theta_v)\rangle/\tau\big). \\
\theta^* &= \arg\max_{\theta}\ \mathsf{L}(\theta)
\end{aligned}$$

Table of Contents

Analysis of CLIP: Gaussian Setting B, Stuart and Tran (2025) [1]

Goal: Analyze closed-form solutions for CLIP optimization

Setting: Gaussian data distribution $\mu = \mathcal{N}(0, \mathcal{C})$ with block covariance matrix

$$\mathcal{C} = \begin{bmatrix} \mathcal{C}_{uu} & \mathcal{C}_{uv} \\ \mathcal{C}_{vu} & \mathcal{C}_{vv} \end{bmatrix}$$

Linear Model: Project $u \in \mathbb{R}^{n_u}$ and $v \in \mathbb{R}^{n_v}$ into $\mathbb{R}^{n_e}$

$$\begin{aligned} g_u(u) &= Gu, \quad G \in \mathbb{R}^{n_u \times n_e}, \\ g_v(v) &= Hv \quad H \in \mathbb{R}^{n_v \times n_e}. \end{aligned}$$

CLIP Model:

$$\begin{aligned} \nu(u, v; \theta) &\propto \exp\left(\langle Gu, Hv \rangle\right) \mu_u(u) \mu_v(v) \\ &= \exp\left(\langle u, Av \rangle\right) \mu_u(u) \mu_v(v), \quad \theta = A = G^\top H. \end{aligned}$$

Analysis of CLIP: Gaussian Setting B, Stuart and Tran [1]

Theorem: No Dimension Reduction

Minimizing J over A with $n_e = \min(n_u, n_v)$ has solution

$$A = \mathcal{C}_{uu}^{-1}\mathcal{C}_{uv}\mathcal{C}_{vv}^{-1}$$

results in the conditional distributions

$$\nu_{u|v}(u|v;\theta^*) = \mathcal{N}(\mathcal{C}_{uv}\mathcal{C}_{vv}^{-1}v;\mathcal{C}_{uu})$$
$$\nu_{v|u}(v|u;\theta^*) = \mathcal{N}(\mathcal{C}_{vu}\mathcal{C}_{uu}^{-1}u;\mathcal{C}_{vv})$$

Corollary: Conditional means of $\mu_{u|v}, \mu_{v|u}$ are recovered, but not variances

Theorem: With Dimension Reduction

Minimizing J over A with $n_e < \min(n_u, n_v)$ has solution

$$A = \mathcal{C}_{uu}^{-\frac{1}{2}}(\mathcal{C}_{uu}^{-\frac{1}{2}}\mathcal{C}_{uv}\mathcal{C}_{vv}^{-\frac{1}{2}})_{n_e}\mathcal{C}_{vv}^{-\frac{1}{2}}$$

Corollary: Small n_e yields low rank approximations of the conditional mean

Analysis of CLIP: Novel Objective Functions

$\mathsf{J}_{\mathsf{one-sided}}$, $\mathsf{J}_{\mathsf{joint}}$ (variants on J) introduced in B, Stuart and Tran (2025) [1]

Original: Match Both Conditionals

$$\mathsf{J}(\theta) = \frac{1}{2}\mathbb{E}_{v\sim\mu_v}\left[\mathsf{D}_{\mathsf{kl}}\left(\mu_{u|v}\|\nu_{u|v}(\cdot;\theta)\right)\right] + \frac{1}{2}\mathbb{E}_{u\sim\mu_u}\left[\mathsf{D}_{\mathsf{kl}}\left(\mu_{v|u}\|\nu_{v|u}(\cdot;\theta)\right)\right].$$

Variant: Match One Conditional

$$\mathsf{J}_{\mathsf{one-sided}}(\theta) = \mathbb{E}^{v\sim\mu_v}\mathsf{D}_{\mathsf{kl}}\left(\mu_{u|v}\|\nu_{u|v}(\cdot;\theta)\right).$$

Variant: Match Joint

$$\mathsf{J}_{\mathsf{joint}}(\theta) = \mathsf{D}_{\mathsf{kl}}\left(\mu\|\nu(\cdot;\theta)\right).$$

Analysis of CLIP: Novel Objective Functions

Theorem 1

When embedding dimension is large enough:

- minimizing J exactly matches conditional means; overestimates conditional covariances
- minimizing $J_{one-sided}$ exactly matches the conditional mean and covariance

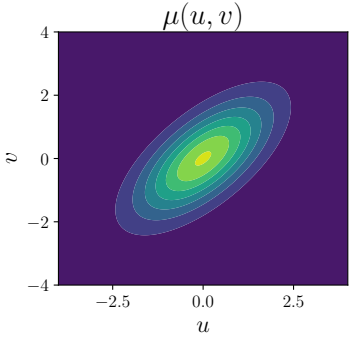
Theorem 2

When embedding dimension is large enough:

- minimizing J underestimates marginal covariances
- minimizing J_{joint} more closely estimates marginal covariances

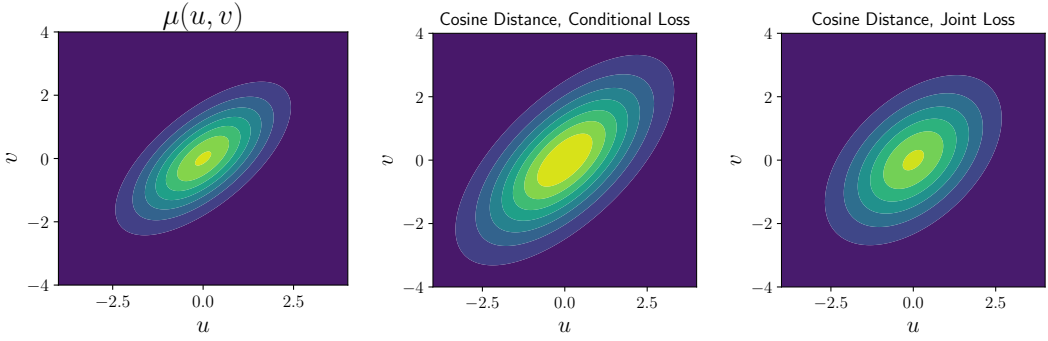
Visualizations of CLIP Generalizations I

- Two-dimensional Gaussian data distribution $\mu = \mathcal{N}(0, \mathcal{C})$
- Used un-normalized linear encoders to preserve Gaussian structure



Visualizations of CLIP Generalizations I

- Two-dimensional Gaussian data distribution $\mu = \mathcal{N}(0, \mathcal{C})$
- Used un-normalized linear encoders to preserve Gaussian structure



Visualizations of CLIP Generalizations II

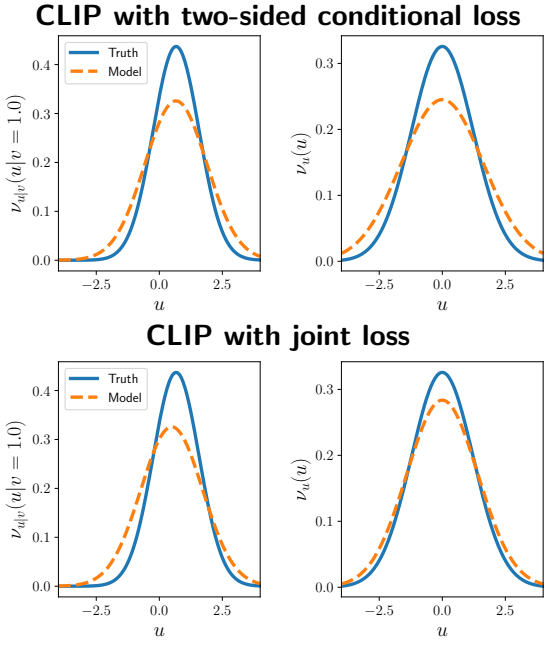
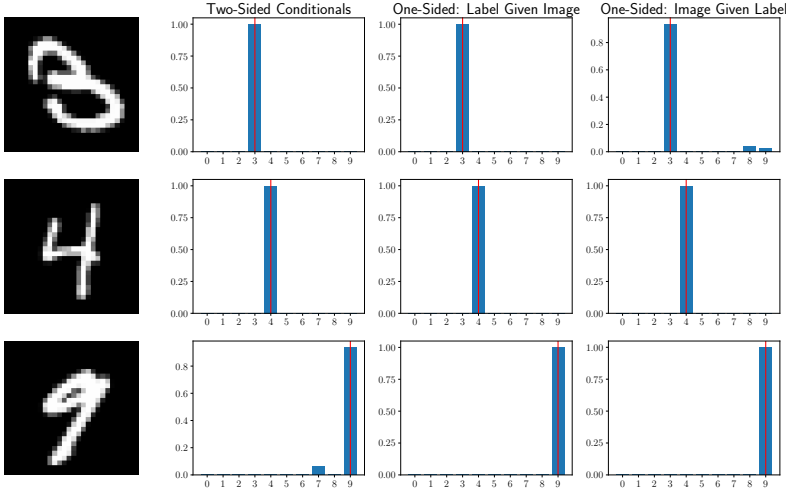


Table of Contents

Application: MNIST

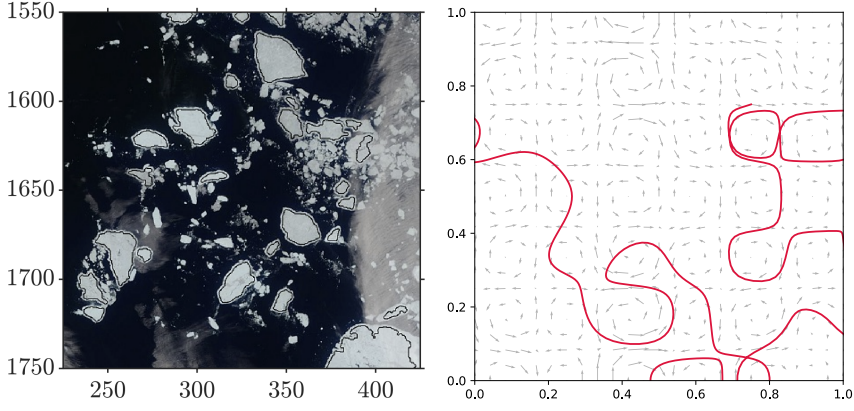
- Data modalities: images $u \in \mathcal{U} = [0,1]^{28\times 28}$ and digits $v \in \mathcal{V} = \{0,\ldots,9\}$
- Performed classifications using models learned with different loss functions



Application: Lagrangian Data Assimilation

Original problem formulation: Kuznetsov, Ide and Jones (2003) [2]

- Infer ocean currents conditioned on data from passive tracers (e.g., sea ice floes)
- Typically relies on a model relating velocity to observations

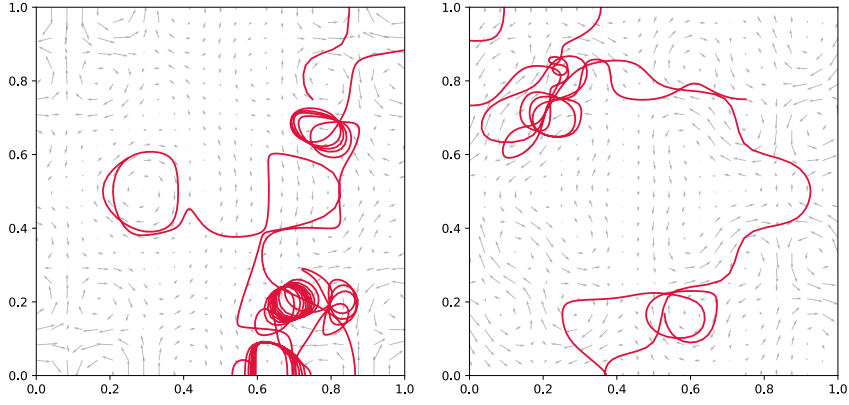


Chen et al. (2021)

Application: Lagrangian Data Assimilation

- Image: Velocity Field is represented using a time-dependent potential
- Time-series: Lagrangian trajectory of a tracer's position
- Velocity is encoded using Fourier representation of the potential

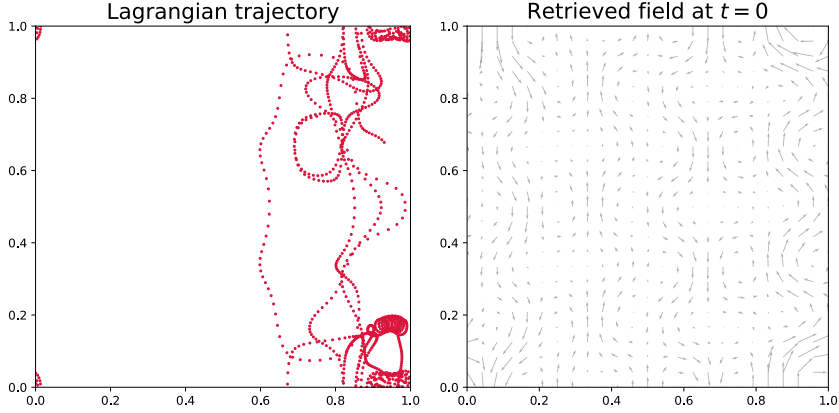
Paired potentials (background) and trajectories (red)



Application: Lagrangian Data Assimilation

- CLIP model aligns trajectory and time-dependent potentials
- Retrieval identifies the most likely velocity from a database given a trajectory

Paired potentials (background) and trajectories (red)



Application: Lagrangian Data Assimilation

- Measured accuracy of retrieval between both modalities

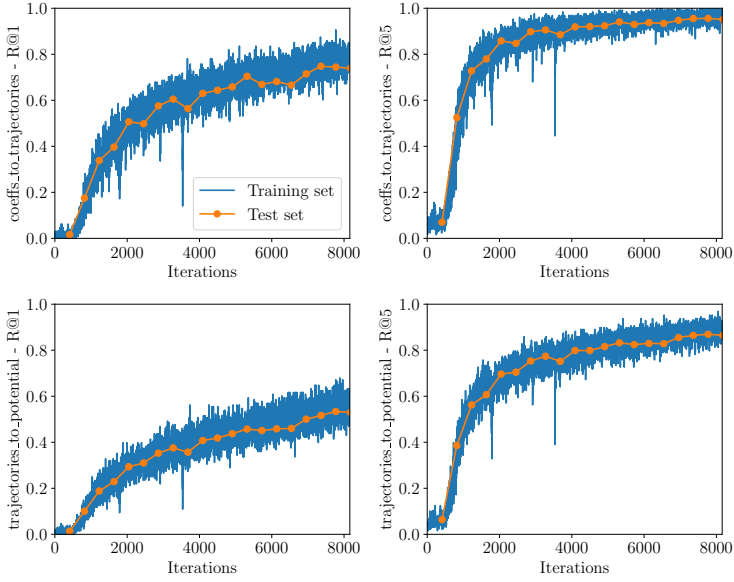


Table of Contents

Conclusions

Main Messages

- Contrastive learning relates two data modalities by tilting a product distribution
- Theory in Gaussian setting:
  - Standard loss matches both conditional means
  - Connections to low rank matrix approximation
- Application to Lagrangian data assimilation

Outlook

- Extending framework to more than two modalities
- Theoretical study of compositional generalization
- Other applications in science and engineering

References I

[1] R. Baptista, A. M. Stuart, and S. Tran.
A mathematical perspective on contrastive learning.
arXiv:2505.24134, 2025.

[2] L. Kuznetsov, K. Ide, and C. K. Jones.
A method for assimilation of Lagrangian data.
Monthly Weather Review, 131(10):2247–2260, 2003.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick.
Microsoft COCO: Common objects in context.
In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[4] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach.
SDXL: Improving latent diffusion models for high-resolution image synthesis.
arXiv:2307.01952, 2023.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al.
Learning transferable visual models from natural language supervision.
In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[6] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al.
Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.
arXiv preprint arXiv:2409.12191, 2024.