Video and Language Alignment in 2D Systems for 3D Multi-object Scenes with Multi-Information Derivative-Free Control

Jason Armitage
University of Zurich
Switzerland

Rico Sennrich
University of Zurich
Switzerland

ABSTRACT

Cross-modal systems trained on 2D visual inputs are presented with a dimensional shift when processing 3D scenes. An in-scene camera bridges the dimensionality gap but requires learning a control module. We introduce a new method that improves multivariate mutual information estimates by regret minimisation with derivative-free optimisation. Our algorithm enables off-the-shelf cross-modal systems trained on 2D visual inputs to adapt online to object occlusions and differentiate features. The pairing of expressive measures and value-based optimisation assists control of an in-scene camera to learn directly from the noisy outputs of Vision-Language Models. The resulting pipeline improves performance in cross-modal tasks on multi-object 3D scenes without resorting to large-scale pretraining or finetuning.

1 INTRODUCTION

Vision-Language Models (VLMs) trained on 2D visual inputs ingest 3D scenes in the form of sequences of discrete viewpoints. Systems rely on this method to perform tasks where test samples include an additional dimension (Wang et al., 2022; Xue et al., 2024; Liu et al., 2024). Consider a visual sequence to be a set of views rendered from a camera moving over a 3D scene. In this case, the problem reduces to predicting in order the location of viewpoints over x-, y-, and z-axes. Given a system and a 3D scene, we posit that an optimal sequence of 2D viewpoints exists, that is the sequence that - when paired with the respective linguistic representations - is most likely to return an accurate prediction from the VLM. But the internal dynamics of systems vary, as do 3D scenes. In tasks where the visual inputs are 2D, generalisation benefits from the existence of large sets of vision and language pairs. Availability of cross-modal training data with 3D visual inputs is by comparison restricted (Poole et al., 2023) resulting in increased likelihood of prediction errors and degraded performance.

If we know the conditions that lead to an error, then the prospect of drawing valid conclusions from a VLM's predictions improves. At a high level, the sources of error in system outputs are the VLM's internal dynamics and the states of the inputs - in this case, 3D scenes and linguistic descriptions. While models to identify anomalies may be learned for each of these sources, generating diagnostics implies offline system identification with a large set of surplus samples (Jin et al., 2020; Coulson et al., 2019).

We propose a straightforward approach of handling predictions as outcomes from a single process and tune an in-scene controller using measures of mutual information (MI) over multiple variables. Assuming no access to the VLM's parameters at inference time, a zeroth-order (ZO) algorithm optimises the expressiveness of the measures on value-based assessments (Chen et al., 2017; Malladi et al., 2023) of the model's predictions. Our multi-information (Studenỳ, 1987) optimised with ZO algorithm (MI-ZO) overcomes the theoretical challenges of estimating MI measures over $n > 2$ (Berrett et al., 2019) mixed discrete and continuous (Gao et al., 2017) variables using active regret normalisation.

Returning to our optimal sequence, viewpoint selections increase in importance with the complexity of the 3D scene. In appraising an object such as a car, we may predict with confidence that the left door will share the same colour as the right. The front view of a building, in contrast, is a less reliable indicator of the surface appearance of the structure's roof or rear exit. Furthermore our understanding of a set of 3D objects benefits from switching perspectives to handle occlusions and resolve specific features of objects.

In our framework, viewpoint control is optimised on a few demonstrations and the relation of a VLM's outputs to each viewpoint and description pair. Accurate appraisals of 3D scenes in relation to linguistic inputs are a key to unlocking the potential of integrating 2D VLMs into processes where automation complements human-initiated generation. Limitations on scaling systems are imposed by data availability. Domain-relevant pairs of 3D assets and language inputs remains restricted to a handful of classes (Poole et al., 2023). A deficit of efficient methods with low latencies to evaluate and perform rapid testing on 3D generations slows development and constrains adoption of systems (Hofmann et al., 2023). An accute challenge is presented by scenes populated with similar objects containing minor differences. Recurring components in large scientific instruments, buildings in 3D virtual environments for embodied AI research, and polygonal shapes in clinical tests (Bandieramonte et al., 2020; Peixoto et al., 2020; Chen et al., 2022; Michalski et al., 2023) rely on discriminating objects drawn from a single class with varying geometric and surface appearances.

We formulate the problem of the dimensional shift presented by 3D scenes as adaptive control of the in-scene camera to return an optimal sequence of actions. Control of the camera is optimised

©2024 Copyright held by the authors. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution.

by measuring the information of the visual scene in relation to the linguistic description. The result is a method for applications with complex 3D scenes that requires no access to the VLM's parameters or costly backpropagation (Malladi et al., 2023). Our research presents four contributions to enable the use of 2D systems for diverse tasks with 3D multi-object scenes:

- A new algorithm denoted as MI-ZO that estimates online multi-information with active regret normalisation on $n > 2$ variables using ZO optimisation. Our method reduces overlaps between a set of continuous and discrete variables representing cross-modal inputs. A set of multivariate measures on multi-object scenes is also provided.
- A novel framework to apply and evaluate adaptive control and multivariate information theoretic measures to predict the actions of an in-scene camera on feedback in the form of noisy VLM outputs. Our controller is tailored to low-data settings with compute-efficient polynomial regression, least-squares approximation, and an interaction matrix.
- A diagnostic comprising 3D scenes of polygons and descriptions with two levels of complexity and a new custom measure of variance to demonstrate the relation of online feedback and active regret normalisation in estimating multivariate mutual information.
- Two new cross-modal benchmarks with 3D multi-object scenes. We introduce sets of 3D scenes with language descriptions to evaluate methods for controlling an in-scene camera and enabling a VLM system to adapt to visual occlusions and handle feature identification on constrained action counts.

2 PROBLEM

Researchers and developers creating 3D assets for deployment in research and real-world applications are challenged by limitations in controlling the generative process and the time required to assess automated generations (Hofmann et al., 2023). Accurate appraisals of 3D scenes in relation to linguistic inputs are a key to unlocking the potential of integrating 2D VLMs into online processes where automation aligns with human-initiated generation and feedback.

We reduce this alignment objective to predicting a sequence of camera actions that returns an accurate assessment of the 3D scene by the VLM system w.r.t. the linguistic input. Controlling an in-scene camera is formulated as estimation of the information capacity of the scene and description as a product of input pairs relative to changes in system correctness. In the low-data settings that are symptomatic of cross-modal tasks with 3D inputs, entropy estimation and value-based optimisation are natural selections for quantifying the information content of inputs. Optimisation with a ZO algorithm is further motivated by running the VLM in inference mode where information on the system is restricted to value-based measures of task performance (Malladi et al., 2023).

Information theoretic metrics that extend entropy estimation to $n > 1$ variables provide properties that are ideal for our objectives:

- Entropy is a fundamental logarithmic measure of information and a basis of the statistical theory of estimation (Shannon and Weaver, 1949; Kullback, 1997).
- Li et al. (2004) demonstrated that distance in the similarity of $n = 2$ sequences provided by algorithmic mutual information is symmetric upto a constant.
- Theoretical guarantees of entropy estimators for $n = 2$ variables are given at different scales - notably when sample sizes are small (Kraskov et al., 2004).

Mutual information provides no such guarantees when the number of variables exceeds two. Short of advance knowledge of all constituents, a measurement for $n > 2$ may be negative when some nonzero quantity of information is redundant (Te Sun, 1980). We begin our study with empirical demonstrations of these theoretical considerations using a set of 3D scenes composed of simple polygons and descriptions with varying levels of complexity. Our analysis is designed to test the case for information-theoretic control with multiple variables to perform indirect closed loop control of a blackbox system using only function values.

2.1 A Numerical Analysis of Entropy and System Performance

The estimation of entropy H based on observations is well-studied in tasks with visual and linguistic samples where information is

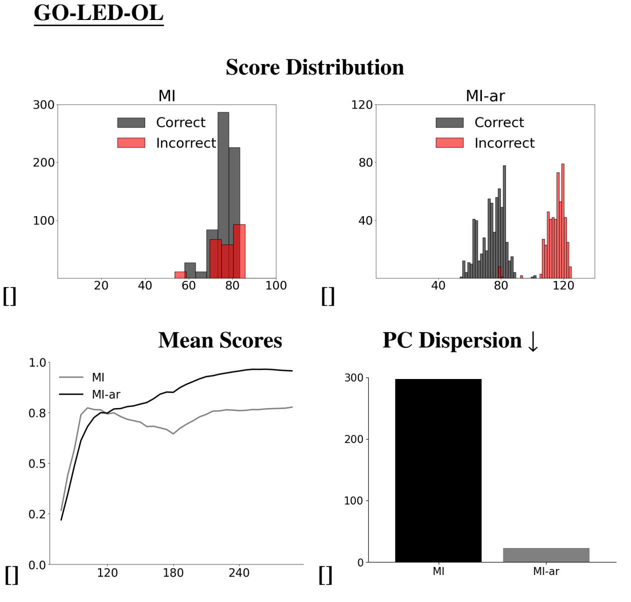


Figure 1: A variant of the *GO-LED-OL* metric with active regret normalisation *MI-ar* is compared against a version calculated with standard methods *MI* on score distribution, normalised means, and PC dispersion for correct and incorrect responses. *MI-ar* groups the respective centres of mass for correctness decisions over separate ranges of the score distribution. PC dispersion is a novel diagnostic derived with Gibbs sampling and designed to measure variance in models trained on entropy measures as sample size increases.

measured on one or two dimensions (Su et al., 2023). Consider the complexity of a visual scene ς to be the sum of distinct visual elements v of greater than or equal to two dimensions:

$$H(\varsigma) = \sum_{i=1}^{n} H(x_{v_i}) \quad (1)$$

A single measurement is the difference in entropy on observing quantities for each dimension Dim:

$$\Delta H = H(Dim_1) - H(Dim_2) \quad (2)$$

We continue with a numerical analysis as an example of calculating multivariate mutual information measurements in relation to the stage of calculating information on the accuracy of VLM responses. Dataset $\mathcal{D}$ is a balanced set of scenes with groups of objects defined by geometry and color (see Technical Appendix). Generating scenes with abstract solids aims to minimise bias originating in class representation in the data used to train the VLM. We compute the means of multivariate measurements MI and correctness labels for each viewpoint to estimate the joint probability densities over both the full balanced set of scenes and binary groups defined by complexity level. Two variants of MI metrics learned by our MI-ZO algorithm that minimises regret with ZO optimisation (MI_{ar}) are assessed against bivariate and multivariate measures estimated by standard methods:

- $GO\text{-}LED\text{-}OL_{ar}$. Regret minimisation over four inputs including global and local visual inputs from the CIELAB colour space.
- $GH\text{-}LED_{ar}$. Regret minimisation over three inputs including a global visual hue variable extracted from the HSV colour model.

To underscore the empirical significance of feedback for minimising regret when combining multiple entropy sources, results are presented for the full information setting unless stated (Bubeck et al., 2012). In this scenario, feedback on labels is provided for all $i - 1$ rounds, where the i-th instance is the current round.

As context for the aims and results of the analysis, we provide a compact summary of the theoretical considerations when calculating mutual information in bivariate and multivariate settings.

- **Shannon information.** For a pair of variables in the set X, the mutual information of any member $H(x_i)$ is in the distribution over x_i in relation to y.
- **Bivariate mutual information.** To limit overlap between members of X, information between multiple input variables and the target y is a combination of single and joint entropies.
- **Multi-information.** Additional joint entropies over all pairs of variables are included for X with more than 2 members.
- **Nonpositivity of multivariate estimates.** An intersection of sources is a positive or negative value when members of X are greater than 1. The theoretical guarantee on positive values for the bivariate does not extend to $n > 2$

	Information setting		
	Full	**50%**	**20%**
Single visual input benchmarks			
- OL (local CIELAB measure)	0.370	0.370	0.370
- LED (local edge density)	0.422	0.424	0.423
- GO (global CIELAB measure)	0.458	0.440	0.452
Multivariate benchmarks			
- GO-LED-OL (no ar)	0.381	0.341	0.383
- GO-OL$_{ar}$	0.761	0.652	0.603
- GO-LED$_{ar}$	0.784	0.677	0.595
Proposed metric			
- GO-LED-OL$_{ar}$	0.832	0.736	0.686

Table 1: Our active regret (ar) measure $GO\text{-}LED\text{-}OL_{ar}$ is assessed against metrics with less inputs and a variant calculated with no ar. Scores for information setting demonstrate the impact when metrics with ar are provided continuous feedback on y correctness labels. Decomposition over variables indicates more inputs contribute to the sensitivity of MI metrics and d the benefits of minimising regret in securing additive effects from new sources when $n > 2$.

Building on these definitions, our analysis compares metric efficacies for samples where the VLM system responses are correct and incorrect. Two variants of multi-information metrics are evaluated: MI estimates information content over sets of variables extending bivariate principles (Studenỳ, 1987) and $MI\text{-}ar$ is calculated using active regret normalisation.

Our first concern is the expressiveness of variants in relation to multi-object scenes presented as a balanced set of uniform and complex configurations (see Figures 2 and 5 (see Technical Appendix)). We also provide scores on the full dataset in relation to the information setting for $MI\text{-}ar$ variants to demonstrate the role of information in minimising regret by managing the contributions of each input on a multi-information measure when $n > 2$ (see Table 1).

A second diagnostic quantifies the stability of a model fitted with each MI variant. Given the low availability of data with 3D scenes paired with language descriptions, a desirable property is enabling models to fit a model with a minimum number of samples from $\mathcal{D}$ that predict the boolean target of system responses. We design the analysis in the form of direct comparison to demonstrate the contribution MI methods in training a model to identify viewpoints commensurate with the level of complexity presented by a scene.

The relation between minimising regret and sensitivity in MI on limited samples demonstrated in Figure 1 motivates a diagnostic on statistical complexity in model optimisation. We study changes in the parameters of a logistic regression model estimated in a framework for Gibbs sampling. Our metric is the difference between maximum and minimum posterior concentrations that starts with samples from the conditional distribution $p(X, y) =$

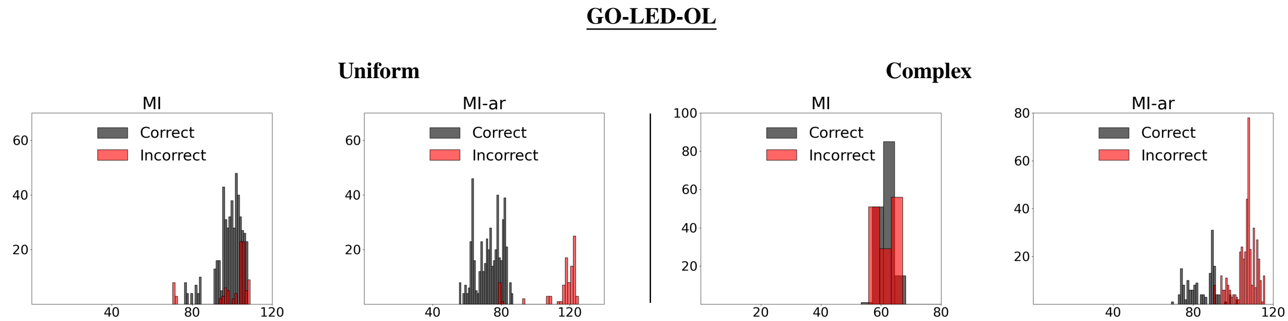


Figure 2: The MI-ar metric groups correct and incorrect scores in distinct regions of the distribution in comparison to the variant with no ar. The advantages of active regret normalisation apply both to scenes with uniform objects and scenes with disparities in object-level geometry and color.

$p(rX \mid y) \cdot p(y)$:

$$\Delta y = \max(y_t) - \min(y_t) \tag{3}$$

where

$$y_t \sim \text{Bernoulli}\left(\frac{1}{1+e^{-X_{t-1}}}\right), \quad p(X_t y_t) = \\ p(y_t \mid X_t \cdot \text{Bernoulli}\left(\frac{1}{1+e^{-r_t}}\right) \tag{4}$$

Model selection here is determined by the low model bias and theoretical guarantees provided by logistic regression when the target for real-based measures is $\{0, 1\}$ (Efron, 1975). Our diagnostic examines variance in the posterior β of the model while a logistic regression is applied to cumulative counts of paired MI estimates in increments of 6 samples. PC dispersion is a point difference over the inverse of the standard deviation σ for β

$$\text{PC Dispersion} = \frac{1}{\sigma(p(\beta|X, y))} \tag{5}$$

where $p(\beta|X, y) \propto p(y|X, \beta)p(\beta)$. Our end measure quantifies absolute dispersion of variance in the model as the number of input samples increases. Posterior concentration is proportional to model stability supplied by the metric over the range of sample sizes from the increment when at least one each of $\{0, 1\}$ labels is recorded.

Results in the above quantities are reported by variant and method in Figures 1 and 6 (see Technical Appendix). Sensitivity in relation to combined visual and linguistic outputs is dependent on the combination of mixture components in computing MI. The advantages of negating overlaps between variables is apparent in the distances between scores for MI-ar. A low value on the change in posterior concentration for MI-ar methods indicate stable updates in the beta parameter of the model over a run. Sensitivity for MI-ar variants to information (see Table 1) supports application in scenarios with limited demonstrations and online feedback. Additional results on information setting and performance of single constituent variables relative to the MI variants are presented in the Technical Appendix.

3 METHOD

We build on the results in the warm-up numerical analysis above to design a controller and algorithm for adaptive camera control in 3D scenes with limited samples. To address the information measurement by analytical methods, we propose combining source entropies on non-differentiable scores with zeroth-order optimisation (ZO). Specifications are motivated by the benefits of low latencies to downstream applications with feedback from human users (Hofmann et al., 2023).

3.1 Controller

Our controller consists of a chain of functions to predict camera actions $\mathfrak{a}$ for $n > 1$ conversation rounds. We combine sample efficient data filters to exploit the information capacity in MI measurements estimated by the MI-ZO algorithm detailed below. A Central Unit f predicts errors $\mathbb{P}^{[x \neq x']}$ and confidence scores CS on axis-level traces and scores from two Component Models (g_1, g_2). View-level data updates a low dimensional representation of the 3D space in the form of an interaction matrix updated with a strong product at each step. A full specification of the controller is presented in the Technical Appendix.

$$\{\mathfrak{a}_1, \mathfrak{a}_2, \ldots, \mathfrak{a}_n\} \leftarrow f\left(\left\{\mathbb{P}_i^{[x \neq x']} \sim g_1(x_1, \hat{y})\right\}_{i=1}^{m}, \\ \left\{CS_j \sim g_2(x_2, \hat{y})\right\}_{j=1}^{n}\right) \tag{6}$$

3.2 Multi-information Estimation with Zeroth-order Optimisation

In this section, we specify an algorithm for estimating multivariate mutual information with active regret normalisation. An entropy source is represented as a component distribution in a weighed mixture distribution. The aim is to add information capacity over the set of distributions by selecting a policy on the mixing weights with active regret normalisation in relation to the optimal mixture. Online updates of a set of policies on mixing weights are performed with predictions on entropy estimates tuned by optimisation on real-value scores.

Sources We first detail the inputs constituting the sources of entropy. A viewpoint render is converted to a color space and a series of intervals is constructed over n axes where $n = 2$ for the CIELAB colour space and the single hue axis is selected for the HSV colour model. Intervals are adjacent and enable multivariate estimation with mixed variable types Hall and Morton (1993). A Sobel operator is passed over grayscale object masks segmented from the image to estimate local edge density (LED). MI variant GH-LED is the product of global hue (GH) and local edge density histograms. GO-LED-OL is composed of global AB axes (GO) with an additional series at object level on LAB axes (OL). Linguistic components are scaling factors in both measures computed over counts of noun phrases and descriptor terms in the viewpoint description.

Weighted Mixture Distribution To avoid integration when combining histograms on viewpoints and real-valued measures of linguistic inputs, we propose a mixture distribution over the sources with a set of policies Π for the component weights $\text{Mixture}(X) = \sum_{i=1}^{k} \pi_i(\theta_i^{mix}) \cdot p_i(x)$ scaled by viewpoint description factors Λ:

$$\text{Mixture}^t(x)' \left(\sum_{i=1}^{k} \pi_i(\theta_i^{mixt}) \cdot p_i(x) \right) \cdot \left(\prod_{j=1}^{m} \Lambda_j \right) \theta_i^{mixt+1} \theta_i^{mixt} \tag{7}$$

Active Regret Normalisation An information setting for regret estimation is a regime with options for feedback on every turn or a subset of turns. Assume the information of a constituent source is composed of units $\mathfrak{b}_u$ and $\mathfrak{b}_d$ in the set $\mathfrak{B}$ specified by the states Up and $Down$. Orientation is defined as the contribution to multi-information where Up is additive and $Down$ is reductive. Means are computed for each state when feedback is provided. A regret minimisation process of this form is equivalent to an optimal mixture of unit states from each source. The goal is to identify a separating margin that minimises the loss on the distance of the closest units $\mathfrak{b}_\mathfrak{d}$ and $\mathfrak{b}_\mathfrak{d}$:

$$\text{Regret} = \min_{\vec{w}, bias} \left\{ \max_{\mathfrak{b}_\mathfrak{u} \in \mathfrak{B}\mathfrak{d}, \mathfrak{b}_\mathfrak{d} \in \mathfrak{B}\mathfrak{d}} |\vec{w}^T \mathfrak{b}_\mathfrak{u} + bias - (\vec{w}^T \mathfrak{b}_\mathfrak{d} + bias)| \right\} \tag{8}$$

Objective The objective is a reduction in the delta on losses for the current set of sources in relation to the mean over all sets accumulated to $i - 1$ rounds. A finite difference approximates a set of parameters for each set of components ϕ in $Mixture(X)$. Our objective is to minimise the finite difference between the two loss terms:

$$f_t(\theta_t) = \frac{1}{T} \sum_{j=1}^{t} \left(\frac{1}{t-1} \sum_{i=1}^{t-1} Loss(\hat{\theta}_{\phi_i}) - Loss(\hat{\theta}_{\phi_t}) \right) \vec{v}_{t_j} \tag{9}$$

Algorithm 1 MI-ZO

```
Input: Sources $H(Dim_n)$, system responses $Y$
initialise: Information setting $\alpha_\tau$, input dimension $\mathfrak{m}$, policies $\Pi$
for $t$ in $\tau$ do
    Mixture$^t(x)' \leftarrow MI(H(Dim_n)_i, y_i)$ for $\pi_i$ in $\Pi$
    for i = 1, . . . , N do
        Compute $\theta_i \leftarrow \arg\min_\Theta Loss(\theta_t - 1, x, y)$
        Minimise $Loss(\hat{\theta}_{\phi_i}) - Loss(\hat{\theta}_{\phi_t})$
        Compute $\frac{1}{T} \sum_{t=1}^{t} MI \equiv MI_t$
    end for
    Maximise $d(\mathfrak{b}_\mathfrak{u}, \mathfrak{b}_\mathfrak{d})$ in $H(Dim_n)$
end for
return $MI$ estimates for $\varsigma$ =0
```

3.3 Outline of the Theoretical Analysis

We start the analysis (see Technical Appendix) by defining multi-information and specify the condition observed in the numerical analysis where - precluding access to full information on all outcomes - the respective densities for uniform and complex pairs tend to equivalent means. If the definition of Shannon entropy is adhered to, there is no guarantee of returning a positive value in all cases (Cover and Joy, 2006) (see also the Proof 1 in the Technical Appendix). As a counter case, assume a set of variables $x(t) = (x_1(t) \ldots x_n(t))$ and outcome y_t specifying a system. A policy π^* returns at time t an operator z to compute $MI(x, y|z)$. Given a set of trials where each X and y are $P(X = 1) = p$ and $P(y = 1) = q$ in a binomial, then the probability mass for the count of X and y are True is

$$P(Z = k) = \binom{n}{k} (pq)^k (1 - pq)^{n-k}. \tag{10}$$

The entropy of z-marginal summed over all k is

$$H(z) = -\sum_{k=0}^{n} \binom{n}{k} (pq)^k (1 - pq)^{n-k} \log \left(\binom{n}{k} (pq)^k (1 - pq)^{n-k} \right), \tag{11}$$

then

$$P(z) = P(X = 1) \times P(y = 1) = p \times q. \tag{12}$$

The multi-information is updated with the term $H(z)$:

$$H(z) = -pq \log(pq) - (1 - pq) \log(1 - pq) \tag{13}$$

assuming $X \perp\!\!\!\perp y|z$.

To demonstrate that such a policy is dependent on observed information, the analysis continues by defining a function that finds a margin γ between additive and reductive elements where the decision variables are a weight vector $\vec{w}$ and $bias$ of the separating hyperplane in the space $\mathbb{R}^D$. [Function to maximise the margin] For any X, an optimisation process $f(x) = \vec{w}^T x + b$

exists to solve the minimax form of deriving the margin using the hyperplane in $\mathbb{R}^D$ that maximises the distance between $\mathfrak{b}_\mathfrak{u}$ and $\mathfrak{b}_\mathfrak{d}$:

$$\max_{\mathfrak{b}_\mathfrak{d} \in \mathfrak{B}\mathfrak{d}} \min_{\mathfrak{b}_\mathfrak{u} \in \mathfrak{B}\mathfrak{u}} f(\mathfrak{b}_\mathfrak{u}, \mathfrak{b}_\mathfrak{d}) = \min_{\mathfrak{b}_\mathfrak{d} \in \mathfrak{B}\mathfrak{d}} \max_{\mathfrak{b}_\mathfrak{u} \in \mathfrak{B}\mathfrak{u}} f(\mathfrak{b}_\mathfrak{d}, \mathfrak{b}_\mathfrak{u}). \tag{14}$$

The proof (see Technical Appendix) assumes that the optimiser algorithm has access to the finite difference of labels on decisions and restates the problem as a dual to lower bound the case that a function exists that guarantees γ with maximum distance between the nearest points in x_1 and x_n. An online setting where information is observed α to a point τ limits the distance between additive and reductive elements at a rate that is linear with minimising regret. We conclude by finding an upper bound on this rate as the bound on the inner product of the vector on B units and the vector of observed information α_τ in T.

	Video-LLaMA-13B		**Chat-UniVi-13B**	
	Acc@8	**Δ on R1**	**Acc@8**	**Δ on R1**
PID	23.8	5.0	22.2	4.4
Kalman-extended	25.0	7.5	20.4	6.6
Linear+SGD	27.4	10.3	22.4	9.8
Poly+ZO+MI (ours)	24.5	6.0	20.4	5.7
+GO-LED-OL	28.2	8.7	22.8	8.0
+GH-LED	32.7	13.4	31.1	12.9
+GO-LED-OL_ar	34.7	14.7	32.7	13.3
+GH-LED_ar	**39.3**	21.0	**35.3**	20.2

Table 2: Assessment on object occlusion of standard control and SGD-optimised neural methods to benchmark our derivative-free controller paired with multi-information measures. A vertical partition in the scene occludes one of the objects depending on camera position. Mean accuracy is reported by camera action count in the correction round and the delta Δ on the measurement round. We report mean values over 10 runs (see Technical Appendix for variance).

4 EXPERIMENTS

4.1 Set Up

We implement a controller framework for empirical testing of multi-information and optimal control methods with open source VLMs. Evaluation on closed systems is precluded to limit the likelihood of data contamination invalidating subsequent benchmarking (Xu et al., 2024). Our benchmarks are designed to benefit methods that predict the optimal sequence of viewpoints for appraising scenes. Experiments present a video-to-text system with a render of a viewpoint of the 3D scene and description. State is preserved over turns in a conversation. At each step, inputs depend on the camera viewpoint in relation to the scene and the system returns a boolean label $\{True, False\}$. Prediction errors in VLM responses in demonstrations are provided as feedback and predicted subsequently.

In the measurement round, the action sequence contains default rotations to traverse a set of cardinal viewpoints and the camera is initialised at the nearest of four distances to the scene. Camera actions predicted by the controller constitute the sequence in the correction round. Data for demonstrations on n scenes run prior to the measurement round are available at the start of the run (see Technical Appendix). In the controller, nodes in the interaction matrix are updated with a strong product operation to predict the next camera action on (X, Y) and $z-$axes dimensions.

4.2 Experiment Settings

Systems: Testing is performed with Video-LLaMA-13B (Zhang et al., 2023) and Chat-UniVi-13B (Jin et al., 2024) as the VLM baselines. Video-LLaMA-13B runs with Llama 2 (Touvron et al., 2023) for decoding and the number of frames defined by camera action count. Conversation hyperparameter settings are as follows: temperature = 1.0, beams = 2, repetition penalty = 1.1. Chat-UniVi-13B uses Vicuna-13B v1.5 (Zheng et al., 2023) and hyperparameter settings: temperature = 0.2, beams = 1. **Benchmarks** Controller performance using polynomial regression is assessed with standard MI and MI-ZO derived MI-ar variants. Tests are conducted with two canonical control methods (Johnson and Moradi, 2005; Ljung, 1979) and a linear layer optimised with stochastic gradient descent (SGD). **Camera:** A single camera is added to the 3D scene at $x-, y-, -z-$ coordinates $(0, 0.5, -40)$ and pointed at the origin in all viewpoints. Initial viewpoint is front and a single rotation about the $x-$axis places the camera equal to ±90° to the left or right from its current position. Rotation about the $y-$axis is ±45° and constrained to the front or back as starting positions. Zoom operations are rotations on the $z-$axis performed in increments of ±5 in the range $[-10, -25]$ forward or backward to the origin. **Metric:** Test performance is measured with mean accuracy Acc on VLM decisions Dec in the scene summary question SQ and calculated over all scenes:

$$\mathrm{Acc}_{\mathrm{SQ}} = \left(100 \times \sum_{i=1}^{N} \frac{\mathrm{Dec}^{\mathrm{SQ}}_{\mathrm{Correct},i}}{\mathrm{Dec}^{\mathrm{SQ}}_{\mathrm{Correct},i} + \mathrm{Dec}^{\mathrm{SQ}}_{\mathrm{Incorrect},i}} \right)$$

4.3 Object Occlusion

Benchmark: Scenes in the PartialViewEf benchmark consist of two objects located on opposite sides of a partition to test the adaptation of control methods when there is full or partial occlusion of one of the objects at every viewpoint. A matching description is selected from a set of five descriptions presented on each turn. At the end of the round, the system uses the information collected to select the correct summary of the scene. **Setting:** Camera action count in both rounds is set to 8. **Results:** A simple controller learning on multi-information metrics with active regret normalisation (see variants with ar in Table 2) adapts camera actions to return views that improve VLM performance. All results display variation over runs related to noise in VLM responses and stochastic operations in method calculations (see Technical Appendix for details).

	Video-LLaMA-13B				Chat-UniVi-13B			
	Acc@5	Δ on R1	Acc@8	Δ on R1	Acc@5	Δ on R1	Acc@8	Δ on R1
PID	21.1	2.1	32.8	7.3	18.1	1.6	28.1	6.5
Kalman-extended	23.6	3.6	34.2	8.5	21.4	2.9	27.1	7.2
Linear + SGD	22.8	2.9	30.1	6.5	17.5	2.4	26.0	5.9
Poly + ZO + MI (ours)	23.6	3.1	35.8	9.2	20.8	2.6	27.7	7.8
+GO-LED-OL	25.9	4.2	37.1	12.4	18.6	3.4	29.1	9.3
+GH-LED	26.2	4.9	39.6	13.9	22.9	4.1	33.3	10.8
+GO-LED-OL_ar	**31.4**	14.3	44.5	19.7	**27.4**	11.4	40.3	15.3
+GH-LED_ar	30.9	10.3	**53.3**	27.2	26.8	7.8	**44.4**	21.5

Table 3: Analysis of methods on prioritising viewpoints in feature identification given a restriction on the number of actions. The objective of the VLM system is to match a summary of the 3D scene by discriminating on descriptions that describe an object feature visible only from selected viewpoints. As in Table 2 mean accuracy is reported by camera action count in the correction round and the delta Δ on the measurement round. We report mean values over 10 runs (see Technical Appendix for variance).

4.4 Feature Identification on Short Runs

Benchmark: In the EffEd benchmark the VLM provides a boolean response on world features (eg "ladder", "doorway") on presentation of a language description. A scene summary containing the features is presented on the final turn. **Setting:** Applications with low latency between feedback and generation rely on efficient estimation (Hofmann et al., 2023). In order to assess control methods on prioritising high-information viewpoints, camera action count is restricted to 5 in the correction round. **Results:** Our $GO\text{-}LED\text{-}OL_{ar}$ metric assists a Poly + ZO + MI model to prioritise viewpoints with fewer actions (see Table 3). Results on variation in results over runs are presented in the Technical Appendix.

5 RELATED WORK

Multivariate Mutual Information Seminal theoretical analysis on challenges in estimating mutual information on multiple variables includes Te Sun (1980) and Berrett et al. (2019). Mohammadi et al. (2018) propose a method to eliminate redundant mutual information in sets of variables ahead of learning. Cabeli et al. (2021) improve conditional mutual information estimation for mixed variable types designed for causal models (Verny et al., 2017) by switching negative results to null values. Steeg (2017) extended the hierarchical decomposition in the work of Watanabe (1960) and earlier publications (Ver Steeg and Galstyan, 2016) for unsupervised representation learning in several domains. **Control Algorithms in Computer Vision** Efficient control methods for tasks with 3D visual inputs form a canonical topic in robotics (Christie et al., 2008; Wiedemann et al., 2023). Gonultas et al. (2023) proposed methods for system identification with gradient-based methods and a controller to steer a vehicle in simulation. System identification methods were adapted by Jaques et al. (2021) to calibrate a camera and estimate 3D poses of objects in scenes using videos. The same authors introduced system identification and control into deep learning models to learn physics from images (Jaques et al., 2020). **3D Visual Inputs in Cross-Modal Tasks** Cross-modal reasoning and generative tasks use 3D inputs to improve textual descriptions of objects (Xue et al., 2024) and scenes Zhao et al. (2023). Reasoning over and generating visual outputs with camera renders of 3D objects includes 2D (Liu et al., 2024) and 3D outputs (Gao et al., 2022; Lin et al., 2023). Voigt et al. (2023) propose search algorithms to retrieve optimal viewpoints of 3D objects to improve a 2D CLIP model. **Zeroth-order Optimisation** Stochastic methods for gradient free optimisation with Gaussian smoothing (Ghadimi and Lan, 2013) formed the basis for the ZO-SVRG optimiser (Liu et al., 2018) that reduces variance in blackbox estimation. Malladi et al. (2023) modified the ZO-SGD process with a memory-efficient that finetunes LLMs with non-differentiable metrics. Efficient optimisation on function values motivated Hoffman et al. (2022) to use the ZO-AdaMM (Chen et al., 2019) algorithm to enhance molecule design.

6 CONCLUSION

In this paper, we propose a novel multi-information estimation method and efficient derivative-free control to predict camera actions that provide optimal sequences of views on 3D scenes. Numerical and theoretical analysis provides the basis for obtaining the first application of multivariate mutual information estimation to enhance the performance of VLM systems on empirical 3D tasks. As part of this research, we design and implement a framework for evaluating control and information theoretic measures, design a set of scenes to illustrate the impact of minimising regret when selecting inputs for calculating multi-information metrics, and present two cross-modal benchmarks on 3D multi-object scenes.

7 ACKNOWLEDGMENTS

This publication is supported by the Digital Visual Studies program at the University of Zurich and funded by the Max Planck Society. RS received funding by the Swiss National Science Foundation (project MUTAMUR; no. 176727).

REFERENCES

M. Bandieramonte, R. M. Bianchi, and J. Boudreau. Fullsimlight: Atlas standalone geant4 simulation. In *EPJ Web of Conferences*, volume 245, page 02029. EDP Sciences, 2020.

T. B. Berrett, R. J. Samworth, and M. Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1): 288–318, 2019.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

V. Cabeli, H. Li, M. da Câmara Ribeiro-Dantas, F. Simon, and H. Isambert. Reliable causal discovery based on mutual information supremum principle for finite datasets. In *Paper presented at WHY21 workshop, 35rd Conference on Neural Information Processing Systems*. NeurIPS, 2021.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9): 2050–2057, 2004.

A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *European Conference on Computer Vision*, pages 638–655. Springer, 2022.

X. Chen, S. Liu, K. Xu, X. Li, X. Lin, M. Hong, and D. Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.

M. Christie, P. Olivier, and J.-M. Normand. Camera control in computer graphics. In *Computer Graphics Forum*, volume 27, pages 2197–2218. Wiley Online Library, 2008.

J. Coulson, J. Lygeros, and F. Dörfler. Data-enabled predictive control: In the shallows of the deepc. In *2019 18th European Control Conference (ECC)*, pages 307–312. IEEE, 2019.

T. M. Cover and A. T. Joy. *Elements of information theory*. John Wiley & Sons, 2006.

B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35: 31841–31854, 2022.

W. Gao, S. Kannan, S. Oh, and P. Viswanath. Estimating mutual information for discrete-continuous mixtures. *Advances in neural information processing systems*, 30, 2017.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

B. M. Gonultas, P. Mukherjee, O. G. Poyrazoglu, and V. Isler. System identification and control of front-steered ackermann vehicles through differentiable physics. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4347–4353. IEEE, 2023.

P. Hall and S. C. Morton. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45:69–88, 1993.

J. H. Halton. Quasi-probability: Why quasi-monte-carlo methods are statistically valid and how their errors can be estimated statistically. *Monte Carlo Methods & Applications*, 11(3), 2005.

Z. Han, C. Chen, Y.-S. Liu, and M. Zwicker. Shapecaptioner: Generative caption network for 3d shapes by learning a mapping from parts detected in multiple views to sentences. In *Proceedings of the 28th ACM International conference on multimedia*, pages 1018–1027, 2020.

HARFANG. Harfang website. `https://www.harfang3d.com/en_US/`, 2024. Accessed: 2024-08-15.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2017.

S. C. Hoffman, V. Chenthamarakshan, K. Wadhawan, P.-Y. Chen, and P. Das. Optimizing molecules using efficient queries from property evaluations. *Nature Machine Intelligence*, 4(1):21–31, 2022.

S. Hofmann, C. Özdemir, and S. von Mammen. Record, review, edit, apply: A motion data pipeline for virtual reality development & design. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, pages 1–4, 2023.

M. Jaques, M. Burke, and T. Hospedales. Physics-as-inverse-graphics: Unsupervised physical parameter estimation from video. In *Proceedings of the International Conference on Learning Representations (ICLR 2020)*, pages 1–16, Apr. 2020. URL `https://iclr.cc/Conferences/2020`. Eighth International Conference on Learning Representations, ICLR 2020 ; Conference date: 26-04-2020 Through 30-04-2020.

M. Jaques, M. Asenov, M. Burke, and T. Hospedales. Vision-based system identification and 3d keypoint discovery using dynamics constraints. *arXiv preprint arXiv:2109.05928*, 2021.

P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024.

W. Jin, Z. Wang, Z. Yang, and S. Mou. Pontryagin differentiable programming: An end-to-end learning and control framework. *Advances in Neural Information Processing Systems*, 33:7979–7992, 2020.

M. A. Johnson and M. H. Moradi. *PID control*. Springer, 2005.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6): 066138, 2004.

A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 324–331, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

R. Krichevsky. A relation between the plausibility of information about a source and encoding redundancy. *Problems Inform. Transmission*, 4(3):48–57, 1968.

S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

M. Li, X. Chen, X. Li, B. Ma, and P. M. Vitányi. The similarity metric. *IEEE transactions on Information Theory*, 50(12):3250–3264, 2004.

C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023.

M. Liu, R. Shi, K. Kuang, Y. Zhu, X. Li, S. Han, H. Cai, F. Porikli, and H. Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

S. Liu, B. Kailkhura, P.-Y. Chen, P. Ting, S. Chang, and L. Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

L. Ljung. Asymptotic behavior of the extended kalman filter as a parameter estimator for linear systems. *IEEE Transactions on Automatic Control*, 24(1): 36–50, 1979.

S. Malladi, T. Gao, E. Nichani, A. Damian, J. D. Lee, D. Chen, and S. Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

D. McAllester and K. Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

S. C. Michalski, N. C. Gallomarino, A. Szpak, K. W. May, G. Lee, C. Ellison, and T. Loetscher. Improving real-world skills in people with intellectual disabilities: an immersive virtual reality intervention. *Virtual Reality*, 27(4):3521–3532, 2023.

S. Mohammadi, V. Desai, and H. Karimipour. Multivariate mutual information-based feature selection for cyber intrusion detection. In *2018 IEEE electrical power and energy Conference (EPEC)*, pages 1–6. IEEE, 2018.

A. Peixoto, R. M. Bianchi, I. Vukotic, C. Potter, C. A. Bourdarios, A. Collaboration, et al. Getting the public closer to the experimental facilities: How virtual reality helps high energy physics experiments engage public interest. In *40th International Conference on High Energy Physics*, page 954, 2020.

B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

R. T. Rockafellar. *Conjugate duality and optimization*. SIAM, 1974.

C. E. Shannon and W. Weaver. *The mathematical theory of communication*. University of illinois Press Champaign, IL, USA, 1949.

G. V. Steeg. Unsupervised learning via total correlation explanation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 5151–5155. AAAI Press, 2017. ISBN 9780999241103.

M. Studenỳ. Asymptotic behaviour of empirical multiinformation. *Kybernetika*, 23(2):124–135, 1987.

W. Su, X. Zhu, C. Tao, L. Lu, B. Li, G. Huang, Y. Qiao, X. Wang, J. Zhou, and J. Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15888–15899, June 2023.

H. Te Sun. Multiple mutual informations and multiple interactions in frequency data. *Inf. Control*, 46:26–45, 1980.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

G. Ver Steeg and A. Galstyan. The information sieve. In *International Conference on Machine Learning*, pages 164–172. PMLR, 2016.

L. Verny, N. Sella, S. Affeldt, P. P. Singh, and H. Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS computational biology*, 13(10):e1005662, 2017.

H. Voigt, J. Hombeck, M. Meuschke, K. Lawonn, and S. Zarrieß. Paparazzi: A deep dive into the capabilities of language and vision models for grounding viewpoint descriptions. *arXiv preprint arXiv:2302.10282*, 2023.

C. Wang, M. Chai, M. He, D. Chen, and J. Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022.

S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

N. Wiedemann, V. Wüest, A. Loquercio, M. Müller, D. Floreano, and D. Scaramuzza. Training efficient controllers via analytic policy gradient. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1349–1356. IEEE, 2023.

C. Xu, S. Guan, D. Greene, M. Kechadi, et al. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

L. Xue, N. Yu, S. Zhang, A. Panagopoulou, J. Li, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27091–27101, 2024.

H. Zhang, X. Li, and L. Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Y. Feng and E. Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL `https://aclanthology.org/2023.emnlp-demo.49`.

Y. Zhao, H. Fei, W. Ji, J. Wei, M. Zhang, M. Zhang, and T.-S. Chua. Generating visual spatial description via holistic 3d scene understanding. *arXiv preprint arXiv:2305.11768*, 2023.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf`.

Appendices

A NOTATION

Notations are defined here for fast reference.

Table 4: Reference List for Standard Notation.

Notation	Usage in this paper
α	Observed information
$\mathbb{A}$	Interaction matrix
$\mathfrak{B}$	Constituent units of entropy sources
$\mathfrak{b}_{\mathfrak{d}}$	Down unit
$\mathfrak{b}_{\mathfrak{u}}$	Up unit
Cap	Information capacity of a channel
$\mathcal{D}$	Dataset
Δ	Difference
Dim	Dimension
η	Step size
γ	Margin
H	Entropy
$\mathbb{M}$	Mutual information
MI	Multi-information
MI_{ar}	Multi-information with active regret normalisation
ω	A computational solution
π	Policy
$\vec{w}$	Weight vector of a hyperplane
$x-$axis, $y-$axis, $z-$axis	Axes
$x-$, $y-$, $z-$	Coordinates

B TECHNICAL ANALYSIS

We present details of the theoretical analysis and proofs for the MI-ZO algorithm. The section begins by proving nonpositivity when estimating multi-information, presents the reformulation of the problem in our approach using a duality, moves the analysis into the online setting, and concludes with a proof to Theorem 1, which states that a function exists that places an upper bound on elements in component variables that reduce the expressivity of multivariate combinations:

Theorem 1 (Function with upper bound on nonpositive contribution)**.** *There exists a function that combines a set of $n > 2$ entropies $H(Dim_n)$ with an upper bound on the nonpositive contribution of reductive units to an output estimate MI that is constant with a bound on the inner product of the vector on total units and the vector of observed information.*

B.1 Multi-information

In this section, we define the conditions and detail a proof for when multi-information is nonpositive. We default to multi-information as a term for multiple mutual information terms (Studenỳ, 1987). We use a semicolon to indicate a function that is not symmetric. In all instances, joint entropy $(x, y) < \infty$.

We begin with the interpretation of mutual information as differences in Shannon entropies (McAllester and Stratos, 2020):

Definition 1. Mutual information for a variable x, $\mathbb{M}(x : x)$ is equal to the entropy of x $(H(x))$. For discrete random variables (x, y), mutual information $\mathbb{M}(x, y)$ is the sum of entropy x and entropy y minus the joint entropy of (x, y):

$$\mathbb{M}(x : y) = H(x) + H(y) - H(x|y). \tag{15}$$

Definition 2. For a pair of variables in the set X, the mutual information of any member $H(x_i)$ is in the distribution over x_i in relation to y. To limit redundancy between members of X, multi-information MI between multiple input variables and the target y is a combination of single and joint entropies (Te Sun, 1980):

$$MI(X : y) = H(x|y) + H(y) - H(X, y) - [H(x_1) + H(x_2, y) - H(x_1) - H(x_2)] \tag{16}$$

Definition 3. Additional joint entropies over all pairs of variables are included for X with more than 2 members. For any n, multi-information follows the chain rule and is summarised as

$$MI(X_n) = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{I \subseteq 1,2,\ldots,n, |I|=k} H(X_I|Y) \right) \tag{17}$$

Lemma 1 (Nonpositivity)**.** Let each member of X be a set A_n, then the intersection of A is a positive or negative value when members of X are greater than 1.

Proof of Lemma 1. First consider that when X contains exactly two elements, then the intersection by the inclusion-exclusion principle for $n = 3$ is:

$$|X_1 \cup X_2 \cup X_3| = |X_1| + |X_2| + |X_3| - |X_1 \cap X_2| - |X_2 \cap X_3| - |X_1 \cap X_3| + |X_1 \cap X_2 \cap X_3| \tag{18}$$

when $X = \{x_1, x_2, \ldots, x_n\}$where $|x_i| < \infty$ for all $i = 1, 2, \ldots, n$ and

$$\text{If } \exists I \subseteq \{1, 2, ..., n\}, |I| > 1 \text{ such that } H(X_I) > \sum_{i \in I} H(x_i) - H(X_I|Y), \text{ then } MI(X;Y) \ngeq 0 \tag{19}$$

Then if the condition is met:

$$H(X_1 \cup X_2 \cup X_3) \leq H(X_1) + H(X_2) + H(X_3) - H(X_1 \cap X_2) - H(X_1 \cap X_3) - H(X_2 \cap X_3) + H(X_1 \cap X_2 \cap X_3) \tag{20}$$

the case in Equation 18 extends by the submodularity of entropy (Krause and Guestrin, 2005) to:

$$H(X_1 \cap X_2) + H(X_1 \cap X_3) + H(X_2 \cap X_3) + H(X_1 \cup X_2 \cup X_3) \leq H(X_1) + H(X_2) + H(X_3) + H(X_1 \cap X_2 \cap X_3) \tag{21}$$

□

B.2 Separation by Hyperplane

In this section, we state assumptions and introduce terms for separable regions in a space over the variables with finite dimensions. Our problem of identifying information by unit type when $n > 2$ is converted to a duality. Definitions and a lower bound are provided for a function that separates constituent units of inputs.

Assumption 1. For the set of elements in X in a multivariate function that results in a single output, note that for $H(X_1), H(X_2), \ldots H(X_n)$, each input is conditionally independent from the other inputs given certain subsets of the variables (Te Sun, 1980). We assume that each member consists of multiple units $\mathfrak{B} \subseteq \mathbb{R}^D$.

Definition 4. The state of any unit $\mathfrak{b} \in \mathfrak{B}$ is additive or reductive in relation to the product of the MI calculation. To specify these states, we use the term $\mathfrak{u}$ to denote Up and $\mathfrak{d}$ to denote $Down$.

Definition 5. A real vector space with all $\mathfrak{B}$ contains affine subspaces with dimension $n - 1$ in $\mathbb{R}^n$ creating regions defined by the inequalities

$$\vec{w}_1\, x_1 + \vec{w}_2\, x_2 + \cdots + \vec{w}_n\, x_n > \mathfrak{c} \tag{22}$$

and

$$\vec{w}_1\, x_1 + \vec{w}_2\, x_2 + \cdots + \vec{w}_n\, x_n < \mathfrak{c} \tag{23}$$

where $\vec{w}$ is a weight parameter of the hyperplane in $\forall 2 + \cdots \vec{w} \in 2 + \vec{W}, \exists \vec{w} \in \vec{W} : \vec{w} \neq 0$ and $\mathfrak{c}$ is a constant term.

Definition 6. The margin γ is the distance between $\mathfrak{b}_{\mathfrak{u}}{}^*$ and $\mathfrak{b}_{\mathfrak{d}}{}^*$ where

$$(\mathfrak{b}_{\mathfrak{u}}{}^*, \mathfrak{b}_{\mathfrak{d}}{}^*) = \underset{\mathfrak{b}_{\mathfrak{u}} \in \mathfrak{B}\mathfrak{u}, \mathfrak{b}_{\mathfrak{d}} \in \mathfrak{B}\mathfrak{d}}{\arg\min}\; d(\mathfrak{a}) \tag{24}$$

and d is the scaling of the vector $\vec{v}$ that is orthogonal to the hyperplane applied as $\mathfrak{b}_{\mathfrak{u}}{}^* = \mathfrak{b}_{\mathfrak{d}}{}^* + d\frac{\vec{v}}{\|\vec{v}\|}$.

Remark 1. Note that γ can also be expressed in terms of the weight vector $\vec{w}$, which is normal to the hyperplane:

$$\gamma = \frac{2}{||\vec{w}||} \tag{25}$$

where $\vec{w}^T x + b = 0$ defines the separating hyperplane, and $\vec{w}^T$ is the transpose of $\vec{w}$.

Theorem 2 (Function to maximise the margin)**.** *For any X, an optimisation process $f(x) = \vec{w}^T x + b$ exists to solve the minimax form of deriving the margin using the hyperplane in $\mathbb{R}^D$ that maximises the distance between $\mathfrak{b}_\mathfrak{u}$ and $\mathfrak{b}_\mathfrak{d}$:*

$$\max_{\mathfrak{b}_\mathfrak{d} \in \mathfrak{Bd}} \min_{\mathfrak{b}_\mathfrak{u} \in \mathfrak{Bu}} f(\mathfrak{b}_\mathfrak{u}, \mathfrak{b}_\mathfrak{d}) = \min_{\mathfrak{b}_\mathfrak{d} \in \mathfrak{Bd}} \max_{\mathfrak{b}_\mathfrak{u} \in \mathfrak{Bu}} f(\mathfrak{b}_\mathfrak{d}, \mathfrak{b}_\mathfrak{u}). \tag{26}$$

Definition 7. A full version of the function $f(x) = \vec{w}^T x + b$ for any X aims to

$$\begin{aligned} &\text{maximize} && t \\ &\text{subject to} && \vec{w}^T x_i - bias \geq t, \quad i = 1, \ldots, N, \\ &&& \vec{w}^T y_i - bias \leq -t, \quad i = 1, \ldots, M, \\ &&& ||\vec{w}||_2 \leq 1. \end{aligned} \tag{27}$$

Proof of Theorem 2. We will convert the primal problem into a dual and derive a lower bound on the latter following classic Lagrangian duality (Rockafellar, 1974). First if we apply a multiplier λ for the constraint $||\vec{w}||_2^2 \leq 1$, a sum of Lagrange multiplications for each group $\sum_{i=1}^{M} u_i y_i$ and $\sum_{i=1}^{M} v_i y_i$ can be derived as follows:

$$\begin{aligned} L(\vec{w}, bias, \gamma, u, v, \lambda) = &- t + \sum_{i=1}^{D} u_i(t + bias - \vec{w}^T x_i) \\ &+ \sum_{i=1}^{M} v_i(t - bias + \vec{w}^T y_i) + \lambda(||\vec{w}||_2^2 - 1) \end{aligned} \tag{28}$$

where $\vec{w}$ and $bias$ are the weight and bias parameters of the hyperplane, M is the number of elements in the group, u_i is the multiplier for the constraints $\vec{w}^T x_i - bias \geq t$, and v_i is the multiplier for the constraints $\vec{w}^T x_i - bias \geq t$.

We have formulated the primal problem with weak duality and can now progress to the second stage of deriving from this new form a lower bound by finding the supremum of the factor $\sup_{\lambda \in \mathbb{R}^m} q(\lambda)$

$$\begin{aligned} &\text{maximize} && \sup_{\lambda \in \mathbb{R}^m} q(\lambda) > q(\bar{\lambda}) = -1 \cdot \bar{\lambda} = D \cdot \text{eigmin}(Q) \\ &\text{subject to} && \sum_{i=1}^{D} u_i = \frac{1}{2}, \quad u_i \geq 0, \\ &&& \sum_{i=1}^{M} v_i = \frac{1}{2}, \quad v_i \geq 0. \end{aligned} \tag{29}$$

noting that Q is the matrix of the quadratic form of the problem and $q(\bar{\lambda}) = -1 \cdot \bar{\lambda} = n \cdot \text{eigmin}(Q)$ is the lower bound derived from Q. □

B.3 Online Setting

We start by defining a formulation of the separating function proposed in the last section in an online learning setting in the form $Out^t = Out^{t-1} - \eta^t L(Out^{t-1})$ where some outcome Out is iterated using step size η. We specify optimisation on real values and prove an algorithm with active regret normalisation on informations.

Definition 8. We underline the difference of full and partial information settings characterised by a specified number τ of rounds where the function has access to observed information denoted as α:

$$\alpha_t = \begin{cases} 1 & \text{if } t \leq \tau \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

Definition 9. A computational solution ω is updated by an online process where a point is mapped to the next nearest point in the convex set C of potential variable values

$$\omega^t = \begin{bmatrix} \vec{w}^t \\ bias^t \\ \gamma^t \end{bmatrix} \leftarrow proj_c \left(\begin{bmatrix} \vec{w}^{t-1} \\ bias^{t-1} \\ \gamma^{t-1} \end{bmatrix} - \eta_t \cdot \alpha_t \cdot \nabla_\omega L(\omega^{t-1}) \right), \text{ where } \nabla_\omega L(\beta^{t-1}) = \begin{bmatrix} \nabla_{\vec{w}L} \\ \nabla_{bias} L \\ \nabla_\gamma L \end{bmatrix} \tag{31}$$

and gradients $\{\nabla_{\vec{w}L, \nabla_{bias}L, \nabla_\gamma L}\}$ are given by $\sum_{i=1}^N u_i \nabla(\vec{w}^T x_i - bias - \gamma) + \sum_{i=1}^M v_i \nabla(\vec{w}^T y_i - bias + \gamma)$ with $\nabla_{\vec{w}L}$ including the additional regularisation step $\lambda \nabla_{\vec{w}(\|\vec{w}\|_2^2 - 1)}$.

Lemma 2. Let regret be the difference in Expectation $\mathbb{E}[\pi_t(out_t)]$ and $\mathbb{E}[\pi_t^*(out_t)]$, where π^* is the optimal policy in the set Π. Then for every t to τ,

$$\text{Regret}_t = \text{Regret}_{t-1} \cdot (1 - \eta_t \cdot \frac{\mathfrak{b}_{\mathfrak{u}t-1} - \mathfrak{b}_{\mathfrak{d}t-1}}{\mathfrak{B}_{t-1}}) + \text{constant}. \tag{32}$$

Remark 2. We have the equivalence between regret and the distance between $\mathfrak{b}_\mathfrak{u}$ and $\mathfrak{b}_\mathfrak{d}$ such that an indicator function $\mathbb{I}$

$$\text{Regret} \equiv \sum_{i=1}^{M} \left[\mathbb{I}\left(\vec{w}^\top \mathfrak{b}_{\mathfrak{u}i} + bias \le 0\right) \cdot \mathbb{I}\left(\mathfrak{b}_{\mathfrak{d}i} = 1\right) + \mathbb{I}\left(\vec{w}^\top \mathfrak{b}_{\mathfrak{u}i} + bias > 0\right) \cdot \mathbb{I}\left(\mathfrak{b}_{\mathfrak{d}i} = -1\right) \right]. \tag{33}$$

then regret in relation to the dual form in Equation 29 is

$$\begin{aligned} \text{Regret}^t \sum_{i=1}^{M} &= \left[\left(\max(0, \vec{w}^\top x_i + bias) \cdot (1 - y_i) + \max(0, -(\vec{w}^\top x_i + bias)) \cdot (1 + y_i) \right) \cdot \alpha_t \right] \\ &+ \alpha_t \frac{1}{2} \left(\sum_{i=1}^{M} \left| \vec{w}^\top x_i + bias \right| \cdot \alpha_t \cdot (\mathfrak{b}_{\mathfrak{u}i} - \mathfrak{b}_{\mathfrak{d}i}) \right). \end{aligned} \tag{34}$$

Proof of Lemma 2. The proof of Lemma 2 builds on the definition of redundancy in universal compression and a subsequent proof given in Cover and Joy (2006, Ch.13). We present a worked example of the above with a set of inputs $X = x_1, x_2, \ldots, x_n$ following an unknown distribution p_θ. For every p in P, there is an associated prior distribution $h \in \{h_1, h_2, \ldots, h_m\}$ over parameters Θ. The definition for the $Redundancy(p_\theta, q)$ of a code is estimated using the Kullback-Leibler distribution

$$Redundancy(p_\theta, q) = KL(p_\theta \| q) = \sum_x p_\theta(x) \log \left(\frac{p_\theta(x)}{q(x)} \right) \tag{35}$$

where q is the inferred distribution. The aim is to minimise the minimax formulation of $Redundancy$ for all p_θ:

$$Redundancy^* = min_q max_\theta KL(p_\theta \| q). \tag{36}$$

We assume that $X = \{1, 2, 3\}$ and $\theta = \{0, 1\}$ with a probability associated when observing $X = 2$ in both distributions p_1 and p_2 denoted as Ↄ. If $p_1 = (1 - \text{Ↄ}, \text{Ↄ}, 0)$, then the KL divergence between q and p_1 is

$$KL(p_1 \| q) = (1 - \text{Ↄ}) \log \left(\frac{1 - \text{Ↄ}}{q_1} \right) + \text{Ↄ} \log \left(\frac{\text{Ↄ}}{q_2} \right) \tag{37}$$

and for p_2 where $p_2 = (0, \text{Ↄ}, 1 - \text{Ↄ})$:

$$KL(p_2 \| q) = \text{Ↄ} \log \left(\frac{\text{Ↄ}}{q_2} \right) + (1 - \text{Ↄ})\text{Ↄ} \left(\frac{1 - \text{Ↄ}}{q_3} \right). \tag{38}$$

The optimal q for the problem is the candidate with minimum KL for both p_1 and p_2 is $q_1 = q_3 = \frac{1-\text{Ↄ}}{2}$ and $q_2 = \text{Ↄ}$

$$q = \left(\frac{1 - \text{Ↄ}}{2}, \text{Ↄ}, \frac{1 - \text{Ↄ}}{2} \right). \tag{39}$$

We now turn to the information capacity of a channel Cap. In the minimax case, we have

$$Cap = \max_{h(\theta)} \min_{q} \sum_\theta h(\theta) KL(p_\theta \| q) \tag{40}$$

recalling that q is inferred. Then we derive $q_h(x)$ over all Θ as follows:

$$q_h(x) = \sum_{\theta=1}^{m} h(\theta) p_\theta(x). \tag{41}$$

Let the mutual information $\mathbb{M}$ representing the change in uncertainty on observing the input X be

$$\mathbb{M}_h(\theta; X) = \sum_{\theta} h(\theta) \sum_{x} p_\theta(x) \log\left(\frac{p_\theta(x)}{q_h(x)}\right). \tag{42}$$

Given the inferred distribution

$$q_h(x) = \left(\frac{1-\Game}{2}, \Game, \frac{1-\Game}{2}\right) \tag{43}$$

the $\mathbb{M}$ for $\theta = 1$ and $\theta = 2$ is

$$\mathbb{M}_h(\theta = 1; X) = h_1 \left[(1-\Game)\log\left(\frac{1-\Game}{\frac{1-\Game}{2}}\right) + \Game \log\left(\frac{\Game}{\Game}\right)\right] = 0.5\left[(1-\Game)\log 2\right] \tag{44}$$

and

$$\mathbb{M}_h(\theta = 2; X) = h_2 \left[\Game \log\left(\frac{\Game}{\Game}\right) + (1-\Game)\log\left(\frac{1-\Game}{\frac{1-\Game}{2}}\right)\right] = 0.5\left[(1-\Game)\log 2\right]. \tag{45}$$

Then finding the optimal $q^*(\theta)$ yields the maximum Cap. Since $h(\theta) = \{0.5, 0.5\}$ maximises the mutual information, the optimal distribution also is $h^*(\theta) = \{0.5, 0.5\}$:

$$q_{h^*}(x) = \left(\frac{1-\Game}{2}, \Game, \frac{1-\Game}{2}\right) \tag{46}$$

□

B.4 Proof of Theorem 1

Remark 3. We follow Krichevsky (1968) in drawing an equivalence between redundancy and regret in the context of information theory. In the online setting, the dual form in Equation 33 is extended to express the relation between minimising regret and increasing the separation between $\mathfrak{b}_{\mathfrak{u}}$ and $\mathfrak{b}_{\mathfrak{d}}$ in the presence of observed information α. To find the supremum of the infimum in the minimax form, we have

$$\begin{aligned}\text{Regret}_t^* = \sup_{\mathfrak{b}_{\mathfrak{d}}} \in \mathfrak{B} \inf_{Q_{X_n}} &\left[\left(\max(0, \vec{w}^\top x_i + bias)\cdot(1-y_i) + \max(0, -(\vec{w}^\top x_i + bias))\cdot(1+y_i)\right)\cdot \alpha_t\right] \\ &+ \frac{1}{2}\sum_{i=1}^{M} \left|\vec{w}^\top x_i + bias\right| \cdot \alpha_t \cdot (\mathfrak{b}_{\mathfrak{u}i} - \mathfrak{b}_{\mathfrak{d}i})\end{aligned} \tag{47}$$

where the relation of α to $\vec{w}$ is $|\langle \alpha, \vec{w}\rangle| \leq \|\alpha\| \cdot \|\vec{w}\|$.

Proof of Theorem 1. We start with the framework of empirical risk minimisation Rk in online learning listed in several forms by Cesa-Bianchi et al. (2004)

$$Rk(\omega) = \frac{1}{M}\sum_{i=1}^{M}\left[\max(0, \vec{w}^\top x_i + bias)\cdot(1-y_i) + \max(0, -(\vec{w}^\top x_i + bias))\cdot(1+y_i)\right]\cdot\alpha_t \tag{48}$$

where the parameters are scaled by α_τ such that

$$\min_{\forall t,\ \alpha_t = 1}\left(\frac{1}{M}\sum_{i=1}^{M} f(x_i, \vec{w}, bias, y_i)\right) \leq Rk(\omega). \tag{49}$$

Then in the setting where α is not present, the information capacity of the channel is dependent on the function f_{mon} over the processes of minimising $\min Rk(\omega)$ and maximising distance $\max d$ being monotonic:

$$\max_{\pi(\theta)} Cap_\pi(\theta; X) \propto f_{mon}\left(\min_{\forall t,\ \alpha_t = 1}\left(\frac{1}{M}\sum_{i=1}^{M} f(x_i, \vec{w}, bias, y_i)\right),\ \min Rk(\omega),\ \max d_{Up,Down}\right). \tag{50}$$

Reordering the above in terms of the solution in Equation 31, f_{mon} holds *iff* a policy π^* that penalises the squared deviation such that the approximation of $\vec{w} + bias$ after the period τ is within Δ of the estimated μ in relation to the empirical mean μ_{emp} when α is present:

$$\begin{aligned} f_{mon}\left(\min_{\forall t,\ \alpha_t=1}\left(\frac{1}{M}\sum_{i=1}^{M} f(x_i, \vec{w}, bias, y_i)\right),\ \min Rk(\omega),\ \max d_{Up,Down}\right) \iff \\ \min_{\forall pred,\ \alpha_t=0}\left(\frac{1}{M}\sum_{i=1}^{M} \pi_i^* \cdot \left|\omega(x_i) - \frac{1}{2}(\vec{w} + bias)\right|^2\right) \le \Delta \end{aligned} \tag{51}$$

We note that in the above $pred$ is the period following τ. The monotonicity observed in numerical results in Table 1 is supported by defining a solution ω_{zo} where the objective is to predict a policy that results in $\left(\mu' - \mu_{emp}\right)^2$. Building on the dual form of regret in Equation 47, we have the correlation between α and $\vec{w}$ such that

$$\alpha \cdot \vec{w} = \sum_{i=1}^{M} \alpha_i \cdot \vec{w}_i. \tag{52}$$

Then by Cauchy-Schwarz, γ is supported when constraining the inner product between α_i and $\vec{w}$

$$\gamma \le \lambda \cdot \sum_{i=1}^{M} \alpha_i \cdot \vec{w}_i \tag{53}$$

such that the result of the factor λ is equivalent to the result of π^*. In terms of the process for the solution ω_{zo}, we have

$$\gamma \le \left(\frac{\lambda}{\min_{\omega_{zo}}\left(\frac{1}{M}\sum_{i=1}^{M}\left|\mu' - \mu_{emp}\right|_{\text{pred}}\right)}\right) \cdot \sum_{i=1}^{M} \alpha_i \cdot \vec{w}_i \tag{54}$$

and extend the dual form in Equation 29 for an upper bound on a full run with ω_{zo} that includes both periods τ and $pred$:

$$\begin{aligned} \gamma \le \left(\frac{\lambda'}{\min_{\omega'}\left(\frac{1}{M}\sum_{i=1}^{M}\left|\mu' - \mu_{emp}\right|_{pred}\right)}\right) \cdot \sum_{i=1}^{M}\Bigg(\int_{t=1}^{pred} \vec{w}_{i,t}\, dt \\ + \int_{t=pred}^{\tau} \alpha_{i,t} \cdot \vec{w}_{i,t}\, dt\Bigg). \end{aligned} \tag{55}$$

□

C DATA

C.1 Benchmarks

Scenes in EffEd are composed from a set of models curated from ShapeNetCore topLevelSynsetId 04460130 (*category: Tower*) (Chang et al., 2015)) and a floor mesh in RGBA $[0.9, 0.9, 0.7, 1.0]$. The dataset comprises 60 glb files with language descriptions. Human-generated descriptions of objects are formed into a sentence for each viewpoint with a template. The true description is one of 5 samples presented as a list. To ensure matches from the true sample, matching modifiers in negative samples are replaced. Each text file is composed of descriptions from 6 viewpoints and a single summary string for the 60 scenes in the dataset. A feature is identified in the description for one of the viewpoints where visibility is confirmed.

PartialViewEf scenes are generated from ShapeNetCore topLevelSynsetId $[3001627, 4379243]$ (*category: Chair, Table*). Objects are separated by a partition mesh intersecting a floor mesh and limiting visibility to a single item from four of six cardinal viewpoints. A three-stage process to synthesize descriptions for scenes starts with human-generated natural language descriptions of constituent objects from Han et al. (2020). Modifiers are verified by hand and corrected to match the objects. Sentences are generated for viewpoints by adding an indefinite article and period. Summary descriptions of the scene consist of object-level texts conjoined into a single sentence. The PartialViewEf benchmark consists of a total of 60 glb files and paired description files. Language descriptions refer to the 3D scene from one of 6 viewpoints.

C.2 Diagnostic

Multi-information estimation methods where $n > 2$ variables represent cross-modal inputs are evaluated with the UCMScm dataset. Data design is balanced with 24 uniform scenes and an equal number of complex scenes. Each scene contains two polygon objects initialised in one of 6 position groups defined in relation to the centroid of a floor mesh. Abstract polygons are selected to limit the impact of class bias in the training data of VLM systems. Language descriptions and scenes are split between concentration on colour or geometry. Complexity relates to matches or differences between objects on these criteria. VLM systems provide predictions on the correctness of descriptions in relation to 6 viewpoints for each 3D scene.

D ADDITIONAL DETAILS ON THE METHOD

D.1 Controller

We provide the full specification for our controller. A set of functions models inputs from the VLM and MI-ZO to predict a sequence of actions that lead to viewpoints defined by $(X, Y)-$ and $z-$axes. Functions are described by subprocess and in a figure with detailed specifications on individual operations at each stage.

Interval Estimation To enable working directly with continuous MI scores, a function converts values to intervals i with selection of interval size based on entropy maximisation

$$X_{MI} \leftarrow \text{interval}_i = \arg\max_{\Delta_i} H(x_k, \Delta_i) \tag{56}$$

on proposals for cut points generated with Halton sequences (Halton, 2005). We prefer this estimation method to a random number generator to reduce variation. Variance is further reduced by computing a mean over cut point proposals returning maximum entropy. The resulting process is automatic and requires no specification of interval widths in experiments.

Component Models Two component models filter data in the responses with the proxy labels $\hat{Y}$ generated during the first step in Algorithm 2. During the correction round, proxy labels replace the boolean values on the match of a viewpoint with its description that are provided as actual labels during the measurement round. To limit processing time, modeling is performed in both models with iterative least squares.

Component Model 1 increments the probability of prediction error $\mathbb{P}^{[x \neq x']}$ by the system for a viewpoint Vp where an error was marked in the prior round. Coefficients are measured for the set of decisions Dec with corresponding viewpoint labels and the demonstration data are updated. Test error rates and score-based measures on scene attributes are processed by Component Model 2 that ranks $z-$axis levels Dim^z for each viewpoint. Traces of the covariance matrices tr in each component model are retained as indicators of model fit $\lim_{n \to \infty}$.

Central Unit Outputs are passed to the Central Unit of the controller. Acceptance of VLM feedback and decisions on the $z-$axis level by viewpoint are modeled using the covariance traces normalised and converted to an inverse factor. View-level data are passed to update element values over a low dimensional representation of the scene in the form of an interaction matrix.

Interaction Matrix An interaction matrix $\mathbb{A}$ is a graph XYZ and a superset over the scene composed of an element drawn from set XY (element xy) with set Z (element z). Elements in each case are unique instances by position. Set XY is a cyclic graph consisting of adjacent nodes with a direct edge to any element in the fully connected graph of set Z. Pendant nodes in the factor graph XY define the adjacency matrix of graph XYZ. A graph product will result in the targeted structure for the graph and a strong product $\boxtimes$ provides a specific advantage in preserving the connectivity of the factors in the edges of any vertex set.

Algorithm 2 Controller Module

```
Input: Visual scene $\varsigma$, Description $l$, Correctness $y$
Generate proxy labels on few samples with derivative-free estimation
for views in scene $\varsigma$ do
    $MI \leftarrow g_{MI-ZO}(H(Dim_n), Y)$ ▹ Compute MI scores with MI-ZO
    $X_{MI} \leftarrow \text{interval}_i = \arg\max_{\Delta_i} H(x_k, \Delta_i)$
    Return proxy labels by viewpoint
    return $\hat{Y}$
end for
for $\varsigma$ in $\mathcal{D}$ do
    $g_{CM_1} \leftarrow 0$
end for

Function Component Model 1
Input: Decision $Dec$, proxy labels $\hat{Y}$
for views in scene $\varsigma$ do
    $\{\mathbb{P}_i\}_{i=1}^{m} \leftarrow \min_{Dec} \sum_{k=1}^{n} (Dec(Vp_k) - \hat{y}_k)^2$ ▹ Model prediction errors
    $tr\left(\{R_k\}_{k=1}^{n}\right) \leftarrow tr\left(\left\{(Dec(Vp_k) - \hat{y}_k)^2\right\}_{k=1}^{n}\right)$ ▹ Compute trace
    return Set of error probabilities per viewpoint $\mathbb{P}^{[x \neq x']}$, Metric on model fit $tr\left(\{R_j\}_{j=1}^{n}\right) \mathbb{P}^{[x \neq x']}$
end for

Function Component Model 2
Input: $z-$axis levels $Vp_{Dim^z}$, proxy labels $\hat{Y}$
for views in scene $\varsigma$ do
    $\{CS_j\}_{j=1}^{n} \leftarrow \min_{Vp_{Dim^z}} \sum_{k=1}^{m} (Vp_{Dim^z k} - \hat{y}_k)^2$ ▹ Model $z-$axis rankings
    $tr\left(\{R_j\}_{j=1}^{n}\right) \leftarrow tr\left(\left\{\left(Vp_{Dim^z j} - \hat{y}_j\right)^2\right\}_{j=1}^{n}\right)$ ▹ Compute trace
    return Confidence in $z-$axis level predictions $CS_j$, Metric on model fit $tr\left(\{R_j\}_{j=1}^{n}\right) CS_j$
end for

Function Central Unit
Input: Error probability per viewpoint $\mathbb{P}^{[x \neq x']}$, Confidence in $z-$axis level predictions $CS_j$, Metrics on model fit $(\mathbb{P}^{[x \neq x']}, tr\left(\{R_j\}_{j=1}^{n}\right) CS_j)$
for views in scene $\varsigma$ do
    $\sum_{i=1}^{p} \mathbb{P}_i \cdot \left(\sum tr\left(R_j^2\right)\right)^{-1} \leq \tau$ ▹ Viewpoint operations
    $\sum_{j=1}^{n} CS_j \cdot \left(\sum tr\left(R_j^2\right)\right)^{-1} \leq \tau$ ▹ Confidence score operations
    $\{\mathfrak{a}_i\}_{i=1}^{p} \leftarrow \boxtimes \left(\mathbb{A}(XY, Z), (Out_x, Out_y)\right)$ ▹ Update interaction matrix
end for
return Set of camera actions $\{\mathfrak{a}_1, \mathfrak{a}_2, \ldots, \mathfrak{a}_n\}$ =0
```

E ADDITIONAL SETTINGS FOR EXPERIMENTS

In this section, we provide core specifications and computing requirements to perform the experiments in the main paper. A run consists of $R = 2$ rounds for the reported results with a sequence of actions estimated during the correction round following a measurement round and n demonstrations. Control operations performed by all evaluated methods are updating demonstration data with system decisions in the measurement round for (X, Y) viewpoints and z-axis levels, updating coefficients measured on the set of decisions with corresponding viewpoint labels, and predicting the set of camera actions for the correction round. Methods receive prediction errors for viewpoints where an error was marked in demonstrations and during the measurement round.

To assess the results in Tables 2 and 3 in the main paper, we present additional details on task settings and technical specifications. **Number of Demonstrations** Demonstrations are set at 5% of scenes for each benchmark: $n = 3$ for object occlusion (PartialViewEf) and feature identification (EffEd). **Additional VLM system specifications** Weights for VLM systems are llava-v1.5-13b for Video-LLaMA-13B and vicuna-13b-v1.5 for Chat-UniVi-13B (Liu et al., 2023; Zheng et al., 2023). **Additional camera and video settings** A main scene camera is initialised. Video is recorded at $60 fps$. **Software** In-scene camera operations and actions are defined with open source versions of Harfang 3.2.4 and HarfangHighLevel libraries (HARFANG, 2024). Video is created with OpenCV 4.9.0.80. **Infrastructure** A single NVIDIA A100 80GB GPU is used for all runs reported in the experiments section to support VLM systems at inference time. Our controller adds no GPU processing or VRAM memory requirements additional to the hardware allocation for running the VLM. In-scene camera operations are performed with a single NVIDIA GeForce RTX 2080 Super 8GB GPU.

F CODE AND DATA SAMPLES

We provide an outline of the code and data samples for review as a supplementary upload in the following archive: `SupMat_11847.zip`. Please refer to the README in the code directory for details on the archive contents.

G ADDITIONAL RESULTS

G.1 Variation in Results for Experiments

We report standard deviations σ for the two benchmark experiments by system and method in Tables 5 and 6.

Distributions of scores over runs are presented in Figures 3 and 4. Means of each set of runs are displayed as black points and bars indicate the credible interval.

G.2 Permutation Tests on Experiment Results

In Tables 7 and 8, an analysis based on permutation testing is performed to assess the distributions of scores from the experiments on EffEd and PartialViewEf benchmarks. Tests are

	σ@8
Video-LLaMA-13B	
PID	0.52
Kalman-extended	0.50
Linear+SGD	1.16
Poly+ZO+MI (ours)	0.38
+GO-LED-OL	1.76
+GH-LED	1.32
+GO-LED-OL$_{ar}$	2.12
+GH-LED$_{ar}$	2.07
Chat-UniVi-13B	
PID	0.74
Kalman-extended	0.40
Linear+SGD	1.58
Poly+ZO+MI (ours)	0.53
+GO-LED-OL	1.18
+GH-LED	1.65
+GO-LED-OL$_{ar}$	1.12
+GH-LED$_{ar}$	1.45

Table 5: Standard deviation over 10 runs on the object occlusion benchmark.

	σ@5	σ@8
Video-LLaMA-13B		
PID	0.50	0.88
Kalman-extended	0.57	0.78
Linear+SGD	1.08	1.83
Poly+ZO+MI (ours)	0.89	1.10
+GO-LED-OL	1.26	1.94
+GH-LED	1.67	2.47
+GO-LED-OL$_{ar}$	1.98	2.65
+GH-LED$_{ar}$	1.96	2.16
Chat-UniVi-13B		
PID	0.43	0.64
Kalman-extended	0.47	0.51
Linear+SGD	1.11	1.79
Poly+ZO+MI (ours)	0.63	1.06
+GO-LED-OL	1.11	1.84
+GH-LED	1.56	1.61
+GO-LED-OL$_{ar}$	1.20	2.84
+GH-LED$_{ar}$	1.48	2.88

Table 6: Standard deviation over 10 runs on the feature identification benchmark.

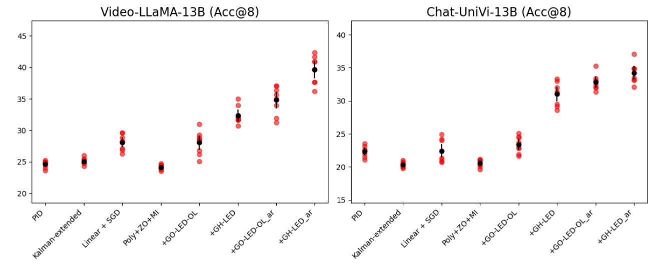


Figure 3: Variance over runs by method for feature identification on a budget of camera actions.

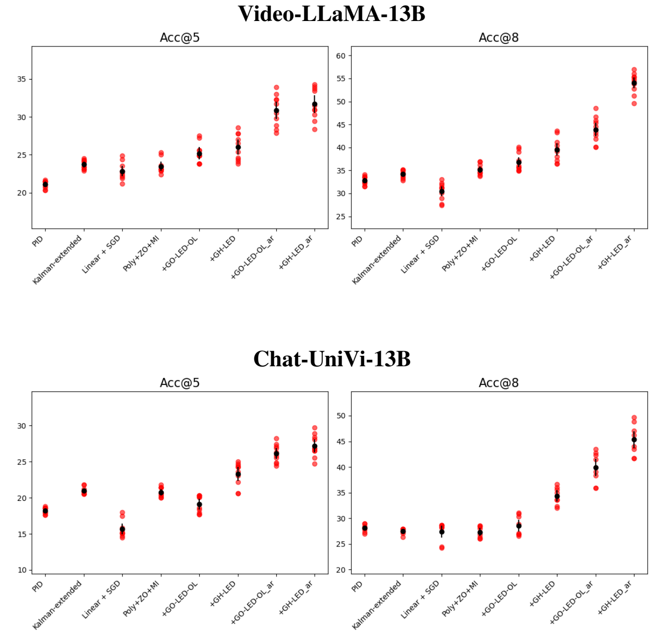


Figure 4: Variance over runs by method for feature identification on a budget of camera actions.

pairwise comparisons between X' the two multi-information metrics estimated with active regret normalisation - $GH\text{-}LED_{ar}$ and $GO\text{-}LED\text{-}OL_{ar}$) - in relation to Y' other methods. A difference Δ is computed on random permutations $\Delta_{rnd_{perm}} = \bar{X}' - \bar{Y}'$ where $N = 10,000$. Output values of the multiple assessments are adjusted with a Bonferroni correction. Our motivation in selecting this method for measuring significance is to minimise assumptions on underlying distributions (Hastie et al., 2017).

G.3 Score Distributions and Analysis by Median for MI Metrics in the Numerical Analysis

Multi-information with active regret minimisation concentrates scores around the median in comparison to multivariate and single visual input measures. The distribution of values for GH, GL, OL, and LED by system response are detailed in relation to the equivalent scores for *ar* variants (see Figures 1 and 6).

	Video-LLaMA-13B				Chat-UniVi-13B			
	+GO-LED-OL ar		+GH-LED ar		+GO-LED-OL ar		+GH-LED ar	
	μ	adj-p	μ	adj-p	μ	adj-p	μ	adj-p
Acc@8								
PID	10.19	-	15.04	-	10.46	-	11.87	0.002
Kalman-extended	9.76	-	14.61	-	12.46	0.001	13.87	0.003
Linear + SGD	6.71	-	11.56	0.003	10.38	0.001	11.79	-
Poly + ZO + MI (ours)	10.67	0.002	15.52	-	12.24	-	13.65	0.002
GO-LED-OL	6.77	0.001	11.63	0.002	9.41	0.001	10.82	-
GH-LED	2.45	0.151	7.30	0.001	1.80	0.181	3.21	0.006

Table 7: Analysis of the distributions of scores over 10 runs for the experiments on object occlusions. Permutation tests are performed with $N = 10,000$) to analyse the observed differences in means for the two methods with our multi-information estimation with active regret minimisation in relation to other methods.

	Video-LLaMA-13B				Chat-UniVi-13B			
	+GO-LED-OL ar		+GH-LED ar		+GO-LED-OL ar		+GH-LED ar	
	μ	adj-p	μ	adj-p	μ	adj-p	μ	adj-p
Acc@5								
PID	9.76	-	10.58	-	7.90	-	8.97	-
Kalman-extended	7.14	-	7.96	-	5.15	-	6.22	-
Linear + SGD	8.01	-	8.83	-	10.46	-	11.53	-
Poly + ZO + MI (ours)	7.38	0.001	8.20	-	5.34	-	6.41	-
GO-LED-OL	5.66	-	6.48	-	6.98	-	8.05	-
GH-LED	4.86	-	5.68	0.001	2.83	0.002	3.90	-
Acc@8								
PID	11.10	-	21.27	-	11.84	0.001	17.28	0.001
Kalman-extended	9.72	-	19.89	-	12.46	-	17.90	-
Linear + SGD	13.39	-	23.56	-	12.51	-	17.95	-
Poly + ZO + MI (ours)	8.71	-	18.88	-	12.59	-	18.03	-
GO-LED-OL	7.07	-	17.24	-	11.38	0.002	16.81	-
GH-LED	4.36	0.016	14.53	-	5.56	0.002	11.00	-

Table 8: Permutation tests analyse the differences in scores on feature identification between $GH\text{-}LED_{ar}$ or $GO\text{-}LED\text{-}OL_{ar}$) and other methods.

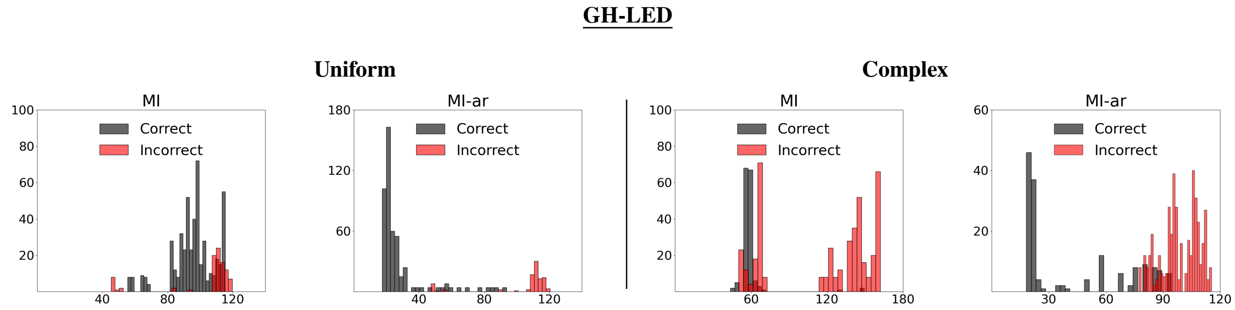
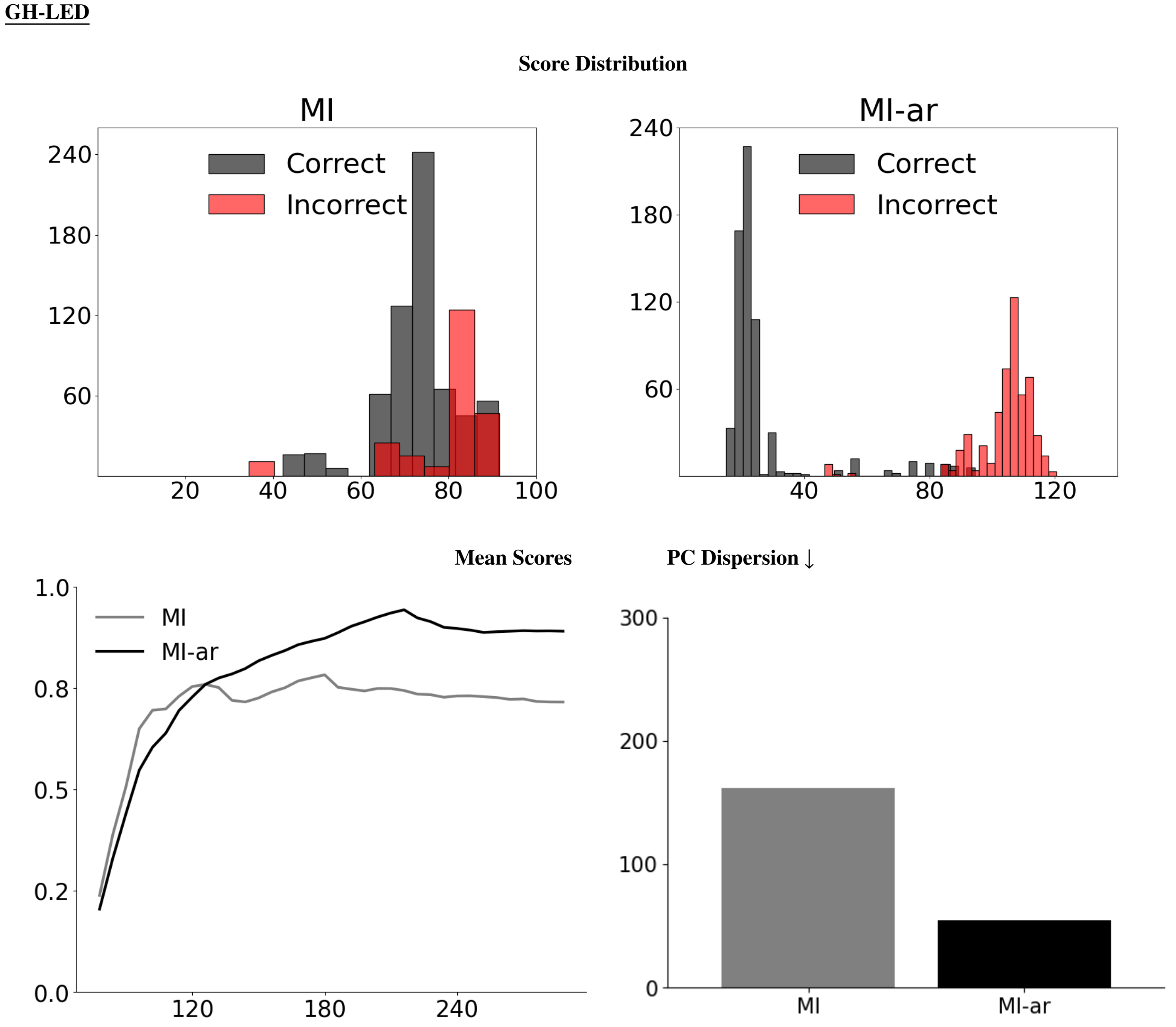


Figure 5: Active minimisation of regret distributes correct and incorrect scores in identifiable concentrations for a second multi-information metric.



GH-LED

Score Distribution

Mean Scores

PC Dispersion ↓

Figure 6: Score distribution and posterior concentration for correct and incorrect responses in relation to information setting. MI-ar is run with full-information and groups the respective centre of mass for correct and incorrect decisions over separate ranges of the score distribution.

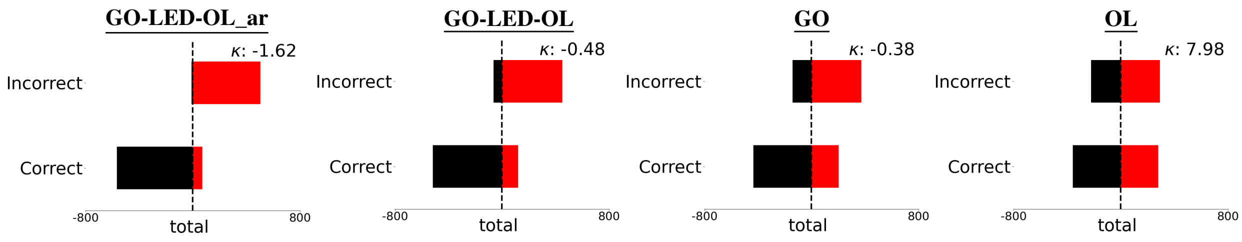


Figure 7: GO-LED-OL scores for correct and incorrect responses distributed around the median. LED score is presented in the figure below (see Figure 8). The median value is preferred to the mean as a measure with lower skew in tests with small sample sizes. Kurtosis κ is provided as a check on the fourth moment and indicates moderate to low concentration in tails for multi-information with active regret minimisation.

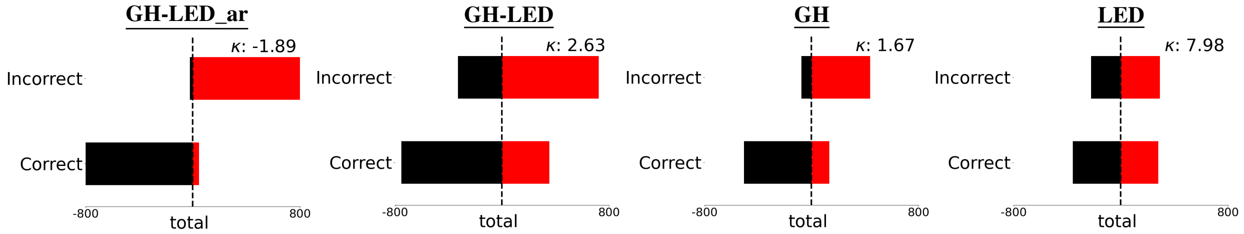


Figure 8: GH-LED scores for correct and incorrect responses distributed around the median.

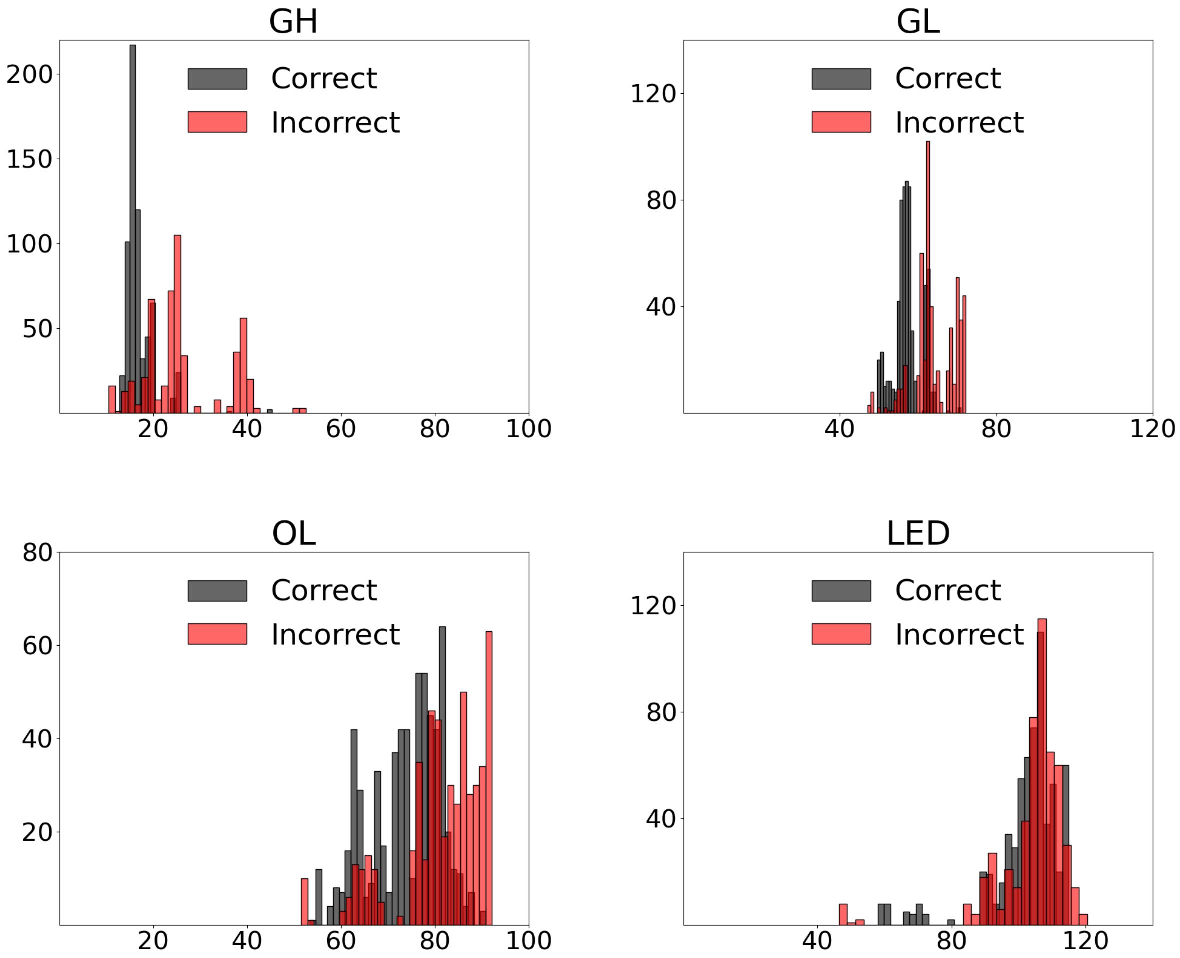


Figure 9: Score distributions for single visual inputs in relation to correct and incorrect responses.