Supplementary Material for
Learning to See Through a Baby's Eyes:
Early Visual Diets Enable Robust Visual Intelligence in Humans and Machines

S1. Implementation Details

Training Details.

CATDiet integrates the staged curricula of CDiet and ADiet by interleaving their respective schedules, yielding an eight-stage training curriculum with stage lengths of [10, 6, 1, 5, 1, 2, 3, 2] epochs. Across these stages, we jointly manipulate the Gaussian blur standard deviation σ and color-saturation ratio s. Specifically, we set $\sigma = [4, 3, 2, 2, 1, 1, 0, 0]$ (in pixels) and assign the saturation ranges for stages one through eight as (0.20, 0.36), (0.36, 0.52), (0.36, 0.52), (0.52, 0.68), (0.52, 0.68), (0.68, 0.84), (0.68, 0.84), and (0.84, 1.0), respectively.

We accelerate data loading using FFCV [65]. Training hyperparameters are selected via grid search for optimal performance. For SimCLR, we employ a staged temperature schedule; the temperature values τ for CDiet, ADiet, TDiet, CATDiet, and CombDiet are listed in **Tab. S1**.

	τ value across stages	stage durations
Cdiet	[0.5, 0.4, 0.3, 0.2, 0.1]	[10, 7, 6, 5, 2]
Adiet	[0.5, 0.4, 0.3, 0.2, 0.1]	[10, 6, 6, 3, 5]
Tdiet	[0.1]	[30]
CATDiet	[0.5, 0.45, 0.4, 0.35, 0.3, 0.2, 0.15, 0.1]	[10, 6, 1, 5, 1, 2, 3, 2]
CombDiet	[0.5, 0.45, 0.4, 0.35, 0.3, 0.2, 0.15, 0.1, 0.1]	[10, 6, 1, 5, 1, 2, 3, 2, 70]

Table S1. **Temperature (τ) schedules used for SimCLR under different visual diets.** Each row corresponds to one visual diet. The second column specifies the τ value used in each stage, and the third column provides the duration of each stage (in epochs).

For DINO, temperature is fixed throughout training; we use a student temperature of $\tau_s = 0.1$ and a teacher temperature of $\tau_t = 0.04$ for all visual diets. For momentum, we use the update schedule from [16] for CATDiet and each individual diet (CDiet, ADiet and TDiet); in the two-phase CombDiet setting, the same update schedule is applied in Phase 1 and the momentum is reinitialized at the beginning of Phase 2 to adapt to SDiet.

Calculation of mCE. We compute the mean Corruption Error (mCE) following the protocol of [47]. ImageNet-C comprises 15 corruption types, each evaluated at 5 severity levels. For a classifier f, let $E^f_{s,c}$ denote the top-1 error under corruption type c at severity level s. The aggregated error for corruption type c is $E^f_c = \sum_{s=1}^{5} E^f_{s,c}$.

Because corruption types differ in difficulty, we normalize these errors by those of a baseline model, E^{baseline}_c. Following [47], we use AlexNet [62] as the reference, yielding the normalized corruption error:

$$CE^f_c = \frac{E^f_c}{E^{\text{AlexNet}}_c}$$

The mCE is then obtained by averaging CE^f_c over all 15 corruption types, providing an overall measure of corruption robustness for model f.

Mapping Classes between IN and SAY. The 15 IN [99] classes used in this study and their corresponding SAY [117] labels are listed below, where each IN label is followed by its corresponding SAY label in parentheses: n02124075 (cat), n09421951 (sand), n04399382 (plushanimal), n04590129 (window), n04239074 (door), n03125729 (crib), n02802426 (ball), n03201208 (table), n04344873 (couch), n03180011 (computer), n04099969 (chair), n03598930 (puzzle), n04285008 (car), n04204238 (basket), and n04462240 (toy).

S2. Ablation

Schedule Ablation. For each diet, we vary stage durations to be either uniform or oppositely ordered from first to last stages, keeping the total number of epochs fixed. In **Tab. S2, left**, our default achieves the best Acc across diets. In **Tab. S2, right**, we also titrate the warm-up ratio in CombDiet (30%, 50%, 70%) and find Acc is highest at 30% (our default).

Schedule	CDiet	ADiet	CATiet
Ours	**62.8**	**54.4**	**64.4**
Opposite	59.8	52.1	63.1
Uniform	62.6	54.0	62.4

Ratio	CombDiet
30%	**74.4**
50%	73.8
70%	73.1

Table S2. Schedule ablation in Acc for SimCLR-ViT on CO3D-10.

Temporal Stride Ablation. We ablate the temporal sampling stride by halving or doubling the default setting used in CombDiet. As shown in **Tab.** S3, the performance remains stable across different strides, indicating that the method is robust to the choice of temporal sampling.

Temporal Stride	Acc	mCE
Ours	83.0	58.4
0.5 x stride	83.0	58.7
2 x stride	83.3	57.7

Table S3. Temporal stride ablation (SimCLR-ResNet, CO3D-10).

Pairing Strategy	Acc	mCE
Ours	83.0	58.4
Adjacent	82.4	60.4
Throughout	70.6	73.0

Table S4. Pairing strategy ablation (SimCLR-ResNet, CO3D-10).

Pairing strategy ablation. We ablate whether artificial warping can replace temporally adjacent video frames for forming positive pairs. Instead of pairing adjacent frames from the video, we consider two alternative strategies: (1) **Adjacent**, where every other frame is replaced by a geometrically warped view and paired with its adjacent real frames to simulate local temporal continuity; and (2) **Throughout**, where a frame is paired only with warped views generated from different azimuth angles, without using any other real frames from the video. As shown in **Tab.** S4, Adjacent performs close to Ours with a small but consistent drop, whereas Throughout performs substantially worse. This suggests that artificial warping can partially substitute adjacent frames and offers a practical trade-off between performance and storage efficiency. However, single-view warping fails to capture long-range temporal continuity, indicating that temporal continuity remains essential.

S3. Object recognition performance of CATDiet on other models

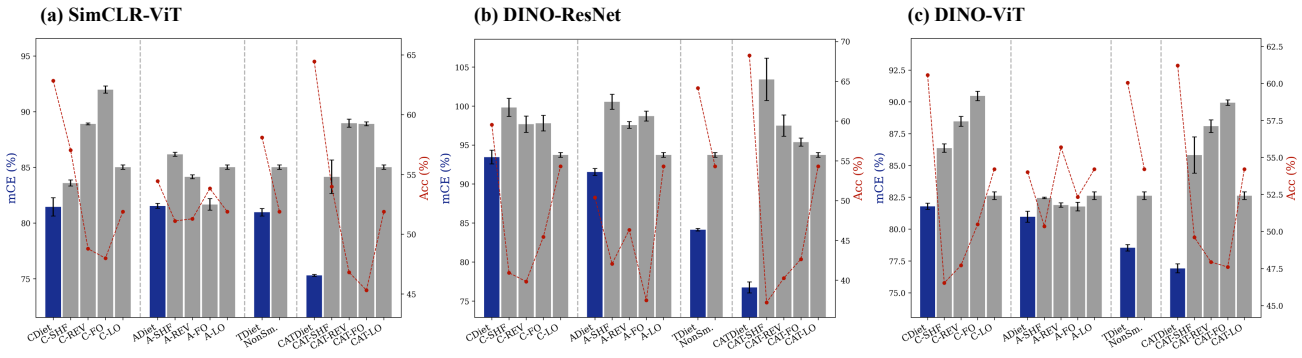

Figure S1. **Object recognition performance on CO3D (clean) and CO3D-C (corrupted) datasets for three SSL models pretrained on CATDiet and its individual diets.** Panels (a–c) show results for SimCLR-ViT, DINO-ResNet, and DINO-ViT, respectively. Within each panel, bars show mCE (↓, left axis) for our diets (blue) and their baselines (gray); the dashed red line indicates Acc (↑, right axis). Error bars reflect the standard error of the mean (SEM) of mCE over three runs. Within each panel, the four groups correspond to different attributes of the proposed visual diets: Color, Acuity, Temporality, and their combination.

As shown in **Fig.** S1, the conclusions in **Sec.** 5.1 consistently generalize across all SSL methods and backbones: Across all models (i.e., across all panels in **Fig.** S1), each individual diet (CDiet, ADiet, TDiet) consistently achieves lower mCE than its corresponding baselines, demonstrating that developmentally-inspired visual diets promote robust representation learning across diverse SSL configurations. Moreover, the integrated CATDiet not only attains the lowest mCE among all diets but also exhibits larger performance gains in mCE and Acc over its baselines than any individual diet across all models,

confirming that the synergistic benefit of integrating all developmentally-inspired diets generalize across other SSL methods and backbones.

S4. Object Recognition and Depth Order Classification Results of CombDiet pretrained on CO3D-10 and CO3D-50

Sec. 5.3 demonstrates that CombDiet improves model robustness and achieves superior performance on both object recognition and depth-order classification tasks. To assess whether these gains extend beyond SAY, we replicate the same experiments on CO3D. We begin with the 10-class CO3D subset used in the main text, and then evaluate scalability by repeating the experiments on an expanded 50-class version, which we denote CO3D-50.

Datasets.

CO3D-50 [97] and **CO3D-50-C**. We construct the **CO3D-50** benchmark using all 50 object categories in the CO3D dataset, sampling approximately 26,000 object instances for our experiments. Training and test splits are defined at the object-instance level within each category, and details of the frame-sampling procedure for both splits are provided in **Sec. 4.1**. To evaluate robustness, we additionally construct a corrupted version of the CO3D-50 test set following the ImageNet-C protocol [47], resulting in **CO3D-50-C**. For clarity, the 10-category subset used in the main text and its corrupted counterpart are referred to throughout as **CO3D-10** and **CO3D-10-C**, respectively.

IN-10, **IN-40**, **IN-10-C**, and **IN-40-C** [99]. For out-of-domain object recognition, we construct two ImageNet-21K subsets based on category overlap with CO3D-10 and CO3D-50. For CO3D-10, the 10 overlapping ImageNet-21K categories form **IN-10**. For CO3D-50, 40 categories align with ImageNet-21K, forming **IN-40**. To evaluate robustness, we generate corrupted versions of the IN-10 and IN-40 test sets using the ImageNet-C protocol [47], producing **IN-10-C** and **IN-40-C**. Complete class mappings between CO3D-10 and IN-10, and between CO3D-50 and IN-40, are listed below, where each ImageNet label is followed by its corresponding CO3D class in parentheses:

The overlapping classes between IN-10 and CO3D-10 are: n07930864 (cup), n04146614 (toybus), n03791053 (motorcycle), n04399382 (teddybear), n03809312 (toyplane), n04026813 (bicycle), n04099969 (chair), n03417042 (toytruck), n02958343 (car), and n02971579 (toytrain).

The overlapping classes between IN-40 and CO3D-50 are: n07930864 (cup), n04146614 (toybus), n03791053 (motorcycle), n04399382 (teddybear), n04552348 (toyplane), n02835271 (bicycle), n04099969 (chair), n03417042 (toytruck), n04285008 (car), n04335435 (toytrain), n03483316 (hairdryer), n04522168 (vase), n04263257 (bowl), n04507155 (umbrella), n03891332 (parkingmeter), n04442312 (toaster), n04447861 (toilet), n02992529 (cellphone), n07248320 (book), n02802426 (ball), n04344873 (couch), n03983396 (bottle), n03991062 (plant), n03085013 (keyboard), n03793489 (mouse), n01608432 (kite), n07873807 (pizza), n02823750 (wineglass), n02769748 (backpack), n07747607 (orange), n03709823 (handbag), n07753592 (banana), n04404412 (tv), n04074963 (remote), n07742313 (apple), n07697537 (hotdog), n07714990 (broccoli), n03891251 (bench), n03642806 (laptop), and n03761084 (microwave).

Detailed Settings. We evaluate CombDiet models pretrained on CO3D-10 and CO3D-50 across object recognition and depth-order classification. We summarize the specific training, linear probing, and test procedures below.

(1) *Pretrained on CO3D-10.* We pretrain four SSL models with CombDiet on the CO3D-10 training set. For object recognition, we train 10-way linear probes on CO3D-10 and evaluate them on both clean CO3D-10 test images and the corrupted CO3D-10-C set. To assess out-of-domain generalization, we train 10-way probes on IN-10 and evaluate on IN-10 and IN-10-C. Depth perception is measured using a 2-way linear probe trained and tested on the 3D-PC Depth Order dataset.

(2) *Pretrained on CO3D-50.* We pretrain four SSL models with CombDiet on the CO3D-50 training set. For object recognition, we train 50-way probes and evaluate them on CO3D-50 and CO3D-50-C. For out-of-domain evaluation, we train 40-way probes on IN-40 and evaluate on IN-40 and IN-40-C. Depth perception is again evaluated using a 2-way linear probe trained and tested on the 3D-PC dataset.

Results. Following the presentation style in **Sec. 5.3**, we report results in **Tab. S5** for both CO3D-10 (Columns 3–7) and CO3D-50 (Columns 8–12). On CO3D-10, CombDiet models achieve competitive or superior performance in Acc and dAcc, and outperform baselines by a substantial margin in mCE in most cases. We observe similar trends on CO3D-50, which includes a much broader set of object categories. These results demonstrate that the progressive ordering of developmental diets is essential: it substantially improves model robustness, whereas disrupting this order reduces performance to that of standard SSL training.

		CO3D					CO3D-50				
		CO3D-10 [97] Acc	CO3D-10-C [47] mCE	IN-10 [99] Acc	IN-10-C [47] mCE	3D-PC [68] dAcc	CO3D-50 [97] Acc	CO3D-50-C [47] mCE	IN-40 [99] Acc	IN-40-C [47] mCE	3D-PC [68] dAcc
SimCLR-ViT [19] [32]	*CombDiet*	74.1 ± 0.2	68.0 ± 0.5	**80.1** ± 0.7	**57.4** ± 0.4	**76.8** ± 1.2	**75.7** ± 0.2	**76.8** ± 0.2	**65.0** ± 0.3	**76.1** ± 0.3	**74.1** ± 1.5
	SHF	74.4 ± 0.3	70.0 ± 0.3	78.5 ± 0.5	61.6 ± 0.4	73.8 ± 1.6	70.9 ± 0.1	83.8 ± 0.1	61.3 ± 0.2	82.2 ± 0.5	68.5 ± 1.4
	STD	**75.3** ± 0.3	**67.8** ± 0.2	78.1 ± 0.1	58.3 ± 0.1	69.3 ± 0.9	73.0 ± 0.7	82.2 ± 0.6	62.1 ± 0.3	79.6 ± 0.2	65.3 ± 1.3
SimCLR-ResNet [19] [43]	*CombDiet*	**83.0** ± 0.2	**58.5** ± 0.3	**88.1** ± 0.4	**45.8** ± 0.5	**71.5** ± 1.3	**82.4** ± 0.0	**70.3** ± 0.1	**75.6** ± 0.2	**69.3** ± 0.2	**75.9** ± 1.7
	SHF	81.9 ± 0.4	62.9 ± 1.0	87.3 ± 0.6	50.4 ± 0.4	66.2 ± 1.1	79.8 ± 0.1	79.8 ± 0.4	73.8 ± 0.6	76.0 ± 0.1	68.8 ± 1.7
	STD	81.0 ± 0.1	64.1 ± 0.7	87.4 ± 0.5	51.9 ± 0.1	65.0 ± 1.1	78.7 ± 0.0	84.0 ± 0.4	73.5 ± 0.3	76.9 ± 0.4	69.7 ± 0.5
DINO-ViT [16] [32]	*CombDiet*	**77.7** ± 0.7	**62.4** ± 0.2	83.7 ± 0.9	**53.8** ± 0.5	**81.0** ± 1.2	80.2 ± 0.1	**67.8** ± 0.1	71.9 ± 0.6	**66.6** ± 0.3	**80.3** ± 1.1
	SHF	73.7 ± 0.1	68.6 ± 0.1	80.1 ± 1.3	62.0 ± 0.2	78.0 ± 0.9	16.8 ± 2.3	115.9 ± 2.4	12.9 ± 1.8	116.8 ± 2.5	57.8 ± 2.7
	STD	76.8 ± 0.6	64.6 ± 0.5	**84.3** ± 0.8	56.2 ± 0.5	80.0 ± 1.1	**80.4** ± 0.2	72.2 ± 0.7	**73.7** ± 0.4	68.6 ± 0.4	80.2 ± 1.6
DINO-ResNet [16] [43]	*CombDiet*	84.3 ± 0.5	**57.0** ± 0.1	89.1 ± 0.3	**49.3** ± 0.5	**73.1** ± 1.2	80.1 ± 0.5	**76.2** ± 1.1	76.2 ± 0.6	**74.5** ± 0.9	69.0 ± 1.9
	SHF	76.8 ± 0.6	76.4 ± 1.0	83.1 ± 1.4	73.5 ± 2.4	64.9 ± 0.8	50.7 ± 4.3	106.2 ± 2.0	47.1 ± 5.8	104.7 ± 2.3	59.1 ± 0.5
	STD	**85.1** ± 0.2	63.6 ± 0.5	**89.6** ± 0.3	59.8 ± 0.5	70.0 ± 0.8	**85.0** ± 0.1	76.6 ± 0.6	**80.3** ± 0.4	74.9 ± 0.4	**76.3** ± 1.2

Table S5. **Object Recognition and depth-order classification results of SSL models pretrained with CombDiet on CO3D-10 and CO3D-50 respectively.** Columns 3-7 denote results of models pretrained on CO3D-10, columns 8-12 denote results of models pretrained on CO3D-50. Each row corresponds to a specific [SSL]-[backbone] configuration (four in total). Within each row, three models are compared: CombDiet, SHF, and STD. Shaded rows highlight CombDiet performance. Values are reported in mean ± standard error of the mean (SEM) across three runs; best results are shown in **bold**.

S5. Additional Method Analysis

Training Convergence. To examine whether the models are sufficiently trained, we report online classification test loss and Acc over epochs. Both CombDiet and STD **(Fig. S2)** reach stable convergence; but CombDiet consistently achieves lower loss and higher Acc after saturation, indicating that our improved final performance of CombDiet over STD arises from better object representations rather than from early stopping.

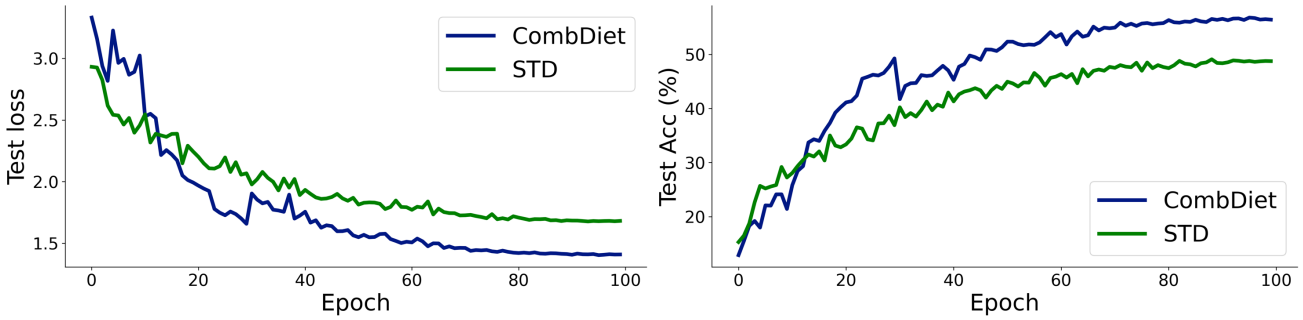

Figure S2. Online classification loss and Acc as a function of training epochs for models trained with STD (green) and CombDiet (blue) on SAY dataset.

Method Comparison. We compare our method with two prior works on curriculum learning, including PAPN [120] and MADAug [52]. PAPN targets fine-grained classification by aligning multiple augmented views of increasing strength across network depth. MADAug is developed for supervised classification, where augmentation policies are learned and adapt dynamically throughout training. Under the same training setup and evaluation protocols, CombDiet consistently outperforms both PAPN and MADAug, as shown in **Tab. S6**. These findings suggest that ecologically valid visual diets derived from infants are more effective than hand-crafted curricula for model training.

	Acc	mCE
CombDiet	**83.2**	**58.7**
PAPN [120]	72.0	67.4
MadAug [52]	51.7	101.6

Table S6. Method Comparison for SimCLR-ResNet on CO3D-10. Best results are in bold.

	Acc	mCE
CombDiet	**66.0**	**47.1**
SHF	65.7	61.5
STD	63.5	65.2

Table S7. Results for SimCLR-ResNet on TOYS. Best results are in bold.

Effectiveness on Synthetic Object-Centric Video Datasets. Toys-200 and Toys4K [115, 116] construct object-centric synthetic video datasets designed to approximate manipulable objects that infants encounter in early visual experience. We further evaluate our method on Toys4K, where we select 10 object classes and render multi-view images from diverse camera poses, denoted as **TOYS**. As shown in **Tab. S7**, CombDiet consistently achieves the best performance on TOYS, supporting the use of scalable synthetic object-centric environments for studying object learning under view continuity.

Statistical Analysis. While a few settings in **Tab. 1** and **Tab. S5** favor STD mainly on Acc, CombDiet still achieves lower or comparable mCE, suggesting reduced shortcut learning. We perform t-tests based on the results reported in both tables. The p-values indicate statistically significant improvements of CombDiet over STD on mCE (p-value < 0.05, **Tab. S8**).

p-value	SAY-C	IN-C	CO3D-10-C	IN-10-C	CO3D-50-C	IN-50-C
CombDiet vs STD	< 0.001	0.001	0.011	0.003	0.002	0.030

Table S8. p-values from t-test on mCE (CombDiet vs. STD).

S6. Limitations

There are several promising directions for future work. First, while we validate a sparse set of hyperparameter choices, exhaustive evaluation across all combinations remains infeasible due to the large search space. Future studies could efficiently expand the hyperparameter search to identify more optimal visual diets. Second, although results on SAY suggest encouraging generalization, CombDiet is primarily designed for temporally coherent video clips featuring a single large, centrally located object. Its effectiveness in fully unconstrained, cluttered egocentric streams remains an open question. Third, prior work [164] has explored active visual settings with human attention estimation for distant objects in cluttered scenes. Incorporating such attention mechanisms could further enhance model robustness and improve the ecological validity of infant visual diets. Finally, while our current focus is on object recognition and depth perception, stress-testing SSL models trained with CombDiet on a broader range of downstream tasks—such as object detection, segmentation, and higher-level cognitive tasks [41, 54, 107, 132, 138, 157–159, 161], remains an important avenue for future research.